BENG102P: Technical Report Writing

Research Paper

Kuriak Tom Jacob(23BCI0095)

Shrihari V(23BCI0083)

Kalyani Manoj(23BCI0081)

Evaluating the Reliability of Machine Translation: User Perspectives on Accuracy, Context, and Limitations

Abstract

This study explores the potential of prompt engineering in improving document-level machine translation (MT) using GPT language models. Unlike sentence-based approaches, document-level MT faces challenges such as maintaining contextual coherence across multiple sentences, particularly when translating nuanced texts. This paper investigates prompt engineering methods that leverage GPT's architecture to address these challenges. By developing and testing prompts that account for document-level consistency, the research evaluates how targeted prompt design can influence translation quality in tasks requiring complex, multi-sentence understanding. Findings indicate that strategic prompt structuring enhances the model's capacity to retain contextual information, offering significant improvements in coherence and accuracy. These results underscore the importance of prompt engineering as a tool for refining document-level MT, providing insights that could inform future advancements in prompt-based translation techniques.

1. Introduction

The rapid expansion of machine translation (MT) technologies has transformed cross-linguistic communication and lowered language barriers, but significant challenges remain, especially concerning accuracy and quality across diverse linguistic contexts (Toral & Way, 2023). Modern neural and transformer-based models, particularly those implemented in neural machine translation (NMT), have elevated translation quality, achieving results previously thought impossible with rule-based or statistical approaches (Stahlberg, 2020). These advancements have positioned MT as a crucial tool across industries, supporting diverse applications from social media and business communications to academic and scientific domains. However, persistent issues such as handling low-resource languages, maintaining document-level coherence, and ensuring cultural appropriateness in translation still hinder universal MT reliability and effectiveness (Kumar & Kaur, 2023; Popović, 2018).

Central to these challenges is the question of how to accurately assess MT quality. Researchers have explored multiple evaluation metrics, such as BLEU, METEOR, and Translation Edit Rate (TER), but these metrics often fall short in capturing nuances, especially at the document level, where context and continuity across sentences are crucial (Snover & Dorr, 2006; Wu & Hu, 2023). Consequently, prompt engineering has emerged as a promising solution, leveraging large language models like GPT to enhance coherence across multi-sentence translations. Initial studies have shown that prompt-engineered models can achieve notable improvements in document-level MT consistency, though these methods require fine-tuning to adapt to the complexities of cross-linguistic context and idiomatic expression (Alam et al., 2022).

The rise of MT for indigenous and underrepresented languages also brings ethical considerations into focus. As recent work highlights, many of these languages lack substantial parallel data, complicating both MT development and quality assessment. This poses significant ethical challenges, particularly in cases where mistranslation can lead to misinterpretation of culturally sensitive content. Researchers like Mager et al. (2023) emphasize that developing ethically sound and accurate MT solutions for indigenous languages is not only a technical goal but also a moral imperative to support linguistic diversity and inclusivity.

Considering these complexities, this study will examine the current methodologies in MT evaluation and enhancement, with a focus on emerging trends such as document-level prompt engineering and evaluation for low-resource languages. By analyzing current advancements and limitations, this research aims to identify areas where further innovation and ethical consideration are needed to ensure that MT can meet the demands of a globally interconnected society.

2. Literature Review

2.1 Machine Translation models

In their study, Ulitkin et al. (2023) investigate the automatic evaluation of machine translation quality, specifically focusing on scientific texts over a five-year period. The authors highlight the transition from traditional statistical translation methods to modern neural machine translation (NMT) systems, which have significantly improved the fluency and contextual accuracy of translations. By employing metrics such as BLEU, F-measure, and Translation Edit Rate (TER), the researchers demonstrate that NMT systems not only produce better translations but also facilitate the identification and categorization of translation errors. This approach not only underscores the advancements in machine translation technology but also emphasizes the necessity for continuous improvement and adaptation of evaluation metrics to keep pace with evolving translation systems. The findings contribute to the ongoing dialogue about the effectiveness of MT tools in professional contexts, indicating a need for further exploration of quality assurance measures in machine translation.

Stahlberg (2020) provides an in-depth analysis of the evolution of machine translation through the lens of neural networks, marking a significant shift from traditional statistical methods to neural machine translation (NMT). The paper outlines how NMT employs deep learning techniques, allowing for more nuanced and context-aware translations, thereby enhancing overall fluency and coherence in the translated text. Stahlberg traces the historical development of NMT architectures, including the adoption of word and sentence embeddings, as well as encoder-decoder models, which have proven essential in advancing the technology. The review also discusses emerging trends, such as the incorporation of larger datasets and more sophisticated algorithms, which have further improved translation accuracy and operational efficiency. This work highlights not only the technical advancements in machine translation but also their broader implications for enhancing global communication and bridging linguistic divides (Stahlberg, 2020; Vaswani et al., 2017; Brown et al., 2020).

In Stahlberg's review of neural machine translation, one notable research gap is the exploration of model interpretability and transparency in NMT systems. The review also lacks a detailed discussion on the ethical implications of deploying NMT in diverse cultural contexts, including issues of bias and fairness in translations, which could affect marginalized language communities.

2.2 Analysis of Machine Translation

In his examination of machine translation systems, Hutchins (2003) highlights the significant advancements and applications of computer-based translation tools, particularly in commercial contexts. He notes that these systems have evolved from rule-based approaches to more sophisticated statistical methods, enhancing their effectiveness in translating technical documentation and repetitive texts. Hutchins emphasizes the role of MT in large organizations, such as the European Union, where it serves to facilitate the rapid processing of internal documents and preliminary drafts for translation, although it often requires post-editing to meet quality standards. Despite the widespread use of MT, he points out a lack of understanding regarding user engagement and satisfaction, suggesting that further research is needed to assess how users interact with these systems and their effectiveness in various contexts (Hutchins, 2003; Ulitkin et al., 2023; Stahlberg, 2020). This gap highlights the importance of integrating user feedback into the development of more refined and user-friendly machine translation technologies.

In her examination of machine translation quality assessment, Maja Popović (2018) provides valuable insights into the systematic classification of translation errors and their implications for improving machine translation outputs. While she identifies several types of errors—such as lexical, morphological, and syntactic—she also highlights the challenges in establishing consistent and comprehensive error taxonomies across different languages and contexts. This underscores a critical gap in the research: the need for a unified framework that can accommodate the diversity of languages and the specific nuances of various domains, including technical, literary, and informal translations. Furthermore, while Popović's work contributes significantly to understanding error types and their impact on translation quality, it lacks a detailed exploration of the end-user perspective. There is limited research on how users perceive translation quality and the significance of different error types in practical applications. Investigating user satisfaction and the contextual factors influencing their assessments could provide essential insights into improving machine translation systems. This gap is echoed in other studies, such as those by Hutchins (2003) and Ulitkin et al. (2023), who call for more attention to user engagement and the effectiveness of machine translation in various scenarios.

Snover et al. (2006) significantly advanced the field of machine translation quality assessment by introducing the Translation Edit Rate (TER) and its human-targeted variant (HTER). These metrics assess the number of edits required to convert a machine-generated translation into an acceptable human reference translation. Their research found that HTER correlates more closely with human judgment than traditional metrics like BLEU and METEOR, particularly when reference translations are tailored for human evaluation. This study emphasizes the importance of error-based quality assessments, revealing that understanding specific types of edits can inform better machine translation system design and implementation. However, while the introduction of HTER has been a valuable contribution, Snover et al. primarily focus on English translations, leaving a gap in understanding how these metrics perform across different languages and dialects. Additionally, while HTER provides insights into the efficiency of machine translations, it does not account for the subjective nature of translation quality, such as fluency and contextual appropriateness. Further research is needed to establish a comprehensive framework that integrates user feedback and contextual factors influencing translation quality, as highlighted by Popović (2018) and Ulitkin et al. (2023). Addressing these gaps could lead to the development of more nuanced evaluation metrics that better reflect real-world translation challenges.

2.3 Ethics of MT tools

Mager et al. (2023) highlight the critical ethical issues associated with the machine translation of low-resource languages, particularly Indigenous languages. The authors emphasize that these languages are closely tied to the cultural identities and histories of their speakers, which necessitates careful consideration during the data collection, modelling, and deployment of machine translation systems. The study reveals that incorporating the perspectives of native speakers and community members is essential for conducting ethically sound research in this area. By actively engaging with these communities, researchers can better address the unique linguistic and cultural nuances involved, ultimately leading to more accurate and respectful translation outcomes. This approach not only enhances the technological capabilities of machine translation but also fosters a more equitable and inclusive framework for language preservation and revitalization. While the authors emphasize the necessity of including native speakers in the development process, they do not provide detailed insights into how different machine translation models can be adapted or optimized for Indigenous languages. This suggests the need for further research that can enhance the representation of Indigenous linguistic features while maintaining cultural integrity.

Additionally, the paper calls for more empirical studies assessing the outcomes of machine translation on community engagement and language revitalization efforts, yet it does not delve into the potential impacts of various technological implementations. Exploring how different approaches can either facilitate or hinder the preservation of Indigenous languages represents an essential avenue for future research. (Mager et al., 2023; Cronin, 2017; Kenny et al., 2019).

3. Methodology

3.1 Research Design

This study utilized a survey-based research design to gather quantitative and qualitative data on participants' experiences and opinions regarding the reliability of machine translation tools. The survey approach was chosen for its effectiveness in collecting a wide range of responses, allowing the study to capture diverse perspectives across various demographics and levels of interaction with machine translation tools.

3.2 Participants

The participants in this study were a diverse group of individuals who use machine translation tools in different contexts, including professional, academic, and casual settings. Recruitment was conducted online through social media platforms, language learning forums, and professional networks, targeting individuals who have varying degrees of familiarity with MT technology. The final sample consisted of [number] participants, with demographic information collected to ensure a broad representation in terms of age, gender, education level, and frequency of MT tool use.

3.3 Survey Instrument

A structured survey was developed for data collection, comprising both closed-ended and open-ended questions to obtain comprehensive insights into participants' experiences and opinions on MT reliability. The survey was divided into three main sections:

1. Demographics and Usage Patterns: This section gathered information on participants' background, including age, education, profession, and the frequency and context of MT tool usage.
2. Reliability of Machine Translation Tools: This section included Likert-scale questions to measure participants' views on various aspects of MT reliability, such as accuracy, consistency, cultural appropriateness, and suitability for different text types (e.g., informal, technical, literary).
3. Qualitative Feedback: Open-ended questions allowed participants to share specific experiences or concerns related to MT reliability, enabling the study to capture nuanced opinions and illustrative examples.

The survey was pre-tested with a small group of individuals to ensure clarity and reliability, with minor adjustments made based on their feedback.

3.4 Data Collection Procedure

Data collection was conducted online over a period of [specify duration], using [survey platform, e.g., Google Forms, Qualtrics] to facilitate accessibility for a broad participant base. Anonymity was ensured to encourage honest responses, and informed consent was obtained from all participants prior to their participation.

3.5 Data Analysis

Quantitative data were analyzed using descriptive statistics to identify trends and patterns in participants' ratings on MT reliability across different contexts. For the qualitative data, thematic analysis was employed to extract common themes and specific insights from open-ended responses. These themes were used to provide depth to the quantitative findings and to highlight particular concerns or areas for improvement in MT reliability as reported by the participants.

3.6 Limitations

While the survey approach allowed for the collection of diverse perspectives, certain limitations should be noted. The sample was limited to individuals with access to online platforms, which may not fully represent all MT users. Additionally, self-reported data may introduce bias, as participants' perceptions may not always align with objective measures of MT reliability.

4. Observation

4.1 User Perceptions of Machine Translation Accuracy

The response distribution reveals a generally positive view of machine translation tools, with a clear understanding of their limitations. About 30% of respondents believe that these tools provide accurate translations "always," indicating satisfaction with machine translation in straightforward or common language scenarios. These users likely find the tools useful for everyday tasks, such as simple sentences or familiar topics. The largest group, 36.8% of respondents, reports that the tools are "often" accurate, suggesting that while generally reliable, machine translation tools may occasionally struggle with complex or context-specific translations. For this group, the tools work well in general, but there are instances where translation quality dips, especially with nuanced language or specialized content. Around 21.5% of users find the tools to be "sometimes" accurate, implying a mixed experience. These respondents probably encounter more frequent issues with idiomatic expressions, specialized vocabulary, or complex ideas. A smaller portion, 8%, believes the tools are "rarely" accurate, pointing to significant errors in certain languages or contexts, such as low-resource languages or complex sentence structures. Lastly, 6% of respondents feel that these tools are "never" reliable, indicating frustration with persistent translation issues or the limitations of machine translation when dealing with highly complex or nuanced language.

4.2 User Perceptions of Machine Translation Accuracy: Insights and Challenges

The responses to the accuracy of machine translation tools provide valuable insights into how users perceive the reliability of platforms like Google Translate and DeepL. Only 8%

of respondents rated these tools as "very accurate," suggesting that while a small group of users find the tools highly reliable, this is not the norm. These users may be translating simple, straightforward content or languages that the tools are well-trained to handle, leading to high accuracy. A larger portion, around 31.9%, consider the tools to be "mostly accurate," indicating that for many users, machine translations work well for everyday tasks, simple sentences, or common vocabulary. However, these users acknowledge that the tools are not perfect and may fall short when dealing with more complex or specialized content. A similar percentage, 32.5%, labeled the tools as "somewhat accurate," reflecting that while these tools can be helpful, they often fail to deliver complete accuracy. This group may struggle with translations that involve idiomatic expressions, cultural context, or domain-specific language that the tools have difficulty grasping. Almost 20% of respondents (19.6%) felt the tools are "often inaccurate," reflecting frustration, particularly with more complex texts, non-standard usage, or languages with unique grammatical structures. For this group, machine translations may be more problematic, resulting in translations that are either nonsensical or only partially helpful. While no specific percentage is provided, there is also a group of users who find the tools "very inaccurate." These users likely encounter situations where the translations are almost entirely unreliable, facing serious mistranslations or complete failures to convey the intended meaning.

4.3 Frequency of Manual Adjustments to Machine Translations

The responses to how frequently users need to manually correct or adjust machine translations offer valuable insight into the practical challenges people face when using tools like Google Translate and DeepL for English. About 10% of respondents report that they "always" need to manually correct translations, suggesting that for this group, the translations are often inaccurate or incomplete. These users may be translating complex, idiomatic, or technical content, where machine translation tools struggle the most. Over 22% of users say they "sometimes" need to adjust translations, indicating that while the translations are generally useful, some fine-tuning is required for accuracy, fluency, or context. This group might encounter translations that are mostly correct but occasionally miss the mark in terms of phrasing or specific details. Nearly 32% of respondents report that they "often" need to correct machine translations. This group likely deals with content where machine translation tools are somewhat helpful but frequently produce errors, such as mistranslated idioms, awkward phrasing, or issues related to context. About 26% of users say they "rarely" need to make adjustments, suggesting that for this group, machine translations are generally accurate and require minimal corrections, possibly due to the simplicity of the content or the strong support for the target language. Lastly, almost 10% of users report that they "never" need to make corrections, indicating that for some, machine translation tools provide sufficiently accurate and fluent translations, especially for straightforward content or languages that closely resemble English in structure and vocabulary.

4.4 Conclusion: Reliability of Machine Translations Across Different Types of Content

The responses to the types of content for which machine translations are most reliable provide valuable insights into how well machine translation tools perform across different types of language use. The largest group, nearly 28%, believes that machine translations are most reliable for casual conversations. This makes sense as casual conversations typically involve everyday vocabulary, simple sentence structures, and less context-specific language, allowing tools like Google Translate or DeepL to produce accurate and meaningful translations for common phrases and greetings. A smaller group, 17.8%, finds machine translation tools reliable for simple documents or texts, such as basic informational content, emails, or instructions, where the language is clear and not overly technical. The straightforward nature of these texts allows machine translation tools to handle them effectively, as they are better suited for simpler vocabulary and sentence structures. Around 24.5% of respondents feel that machine translations work well for technical or academic content, particularly in fields with standard terminology like science, engineering, or medicine. However, this group still faces challenges with domain-specific jargon, complex sentence structures, and abstract concepts that machine translation tools may not always capture accurately. Only 8.6% of respondents believe that machine translation tools are reliable for legal or formal documents. This reflects the specialized nature of legal texts, which often contain nuanced language, formal phrasing, and complex syntax, making them difficult for machine translations to handle with high accuracy. Lastly, about 21.5% of users believe that machine translation tools are reliable across all types of content, suggesting that they may use these tools in languages or contexts where they perform consistently well or have a more optimistic view of their capabilities, particularly for everyday use cases.

4.5 Consistency of Machine Translations Across Repeated Uses

The responses regarding the consistency of machine translations when translating the same text multiple times provide important insights into the reliability of these tools over repeated uses. Only 9.2% of respondents rate machine translations as "very consistent," suggesting that a small group of users experience little to no variation in translations, particularly when dealing with simple texts or languages well-supported by the translation tools. A larger group, 33.7%, finds machine translations to be "mostly consistent." These users experience minor variations but find the overall meaning and structure of the translations remain largely unchanged. This is likely the case for simpler, more straightforward content where ambiguity is minimal. About 24% of users report that translations are "somewhat consistent," meaning that while the overall meaning stays the same, variations in wording, phrasing, or style may occur between repetitions. This could happen when the machine translation system attempts to rephrase or restructure the text differently, especially for languages with flexible sentence structures. Around 23% of users find the translations to be "inconsistent," indicating that the translation tool produces noticeably different results each time. This could reflect greater challenges with more complex or less supported languages, ambiguous phrasing, or shifting translation strategies. A smaller portion, 10.4%, reports translations as "very inconsistent," implying that the translations vary significantly, particularly for more difficult or nuanced language pairs where the translation tool may struggle with specific syntactic or lexical challenges.

4.6 Challenges faced

The most common challenge users face with machine translation tools is incorrect grammar or sentence structure, with 100 responses (61.3%), indicating that many users encounter grammatical errors or poorly constructed sentences that affect the clarity and readability of translated content. Close behind, 101 respondents (62%) reported difficulty handling idioms or colloquial phrases, highlighting how machine translation struggles with non-literal language, often leading to inaccurate or confusing translations. Overly literal translations were also a frequent issue, reported by 96 respondents (58.9%), suggesting that translations often fail to capture the intended meaning in a natural, contextually appropriate way. Misinterpretation of context or meaning, chosen by 74 respondents (45.4%), reflects how translation tools can miss the broader context, leading to misunderstandings. Additionally, 52 respondents (31.9%) noted poor translation of technical or specialized terms, indicating that machine translation struggles with domain-specific language, which often requires a deeper understanding of specialized vocabulary.

Informal conversations or social media were the most reported areas for machine translation errors, with 105 responses (64.4%), suggesting that the diverse and dynamic language used in informal settings like social media often leads to translation issues. Travel or tourism communication followed closely with 88 responses (54%), indicating that translation problems during travel can complicate communication and affect the clarity of essential information for tourists. Academic or research purposes ranked third, with 72 responses (44.2%), highlighting the critical need for accuracy in these fields, where translation errors can significantly impact understanding and the reliability of research. Legal or contractual documents received 71 responses (43.6%), underscoring the importance of precise language in legal contexts, where even small mistakes can lead to major misunderstandings. Finally, 52 respondents (31.9%) reported translation issues in business or professional communication, showing that even in formal settings, translation errors can pose significant challenges to effective communication.

Maintaining tone and nuance was marked as unreliable by 98 respondents (60.1%), suggesting that machine translation struggles to capture the emotional tone and subtle differences in meaning, which are often crucial in conveying the full message. Accuracy of technical terminology was a concern for 87 respondents (53.4%), pointing to the difficulty of translating specialized vocabulary in fields like science, engineering, and medicine, where precision is paramount. Understanding cultural context was marked by 76 respondents (46.6%), highlighting how machine translation often misses culturally specific meanings or references, which can lead to misunderstandings. Handling long or complex sentences was noted by 65 respondents (39.9%), suggesting that machine translation faces challenges with sentence structure and meaning when dealing with more intricate phrases. Finally, translating gendered or respectful language was mentioned by 48 respondents (29.4%), indicating that while this remains a concern, it is less of an issue compared to the other factors.

English to Asian languages (e.g., Chinese, Japanese) had the highest difficulty, with 104 responses (63.8%). This suggests that translating between English and Asian languages presents significant challenges, likely due to major structural and cultural differences. English to Middle Eastern languages (e.g., Arabic, Hebrew) was the second most challenging, with 66 responses (40.5%). This may stem from the unique syntax, grammar, and right-to-left script of Middle Eastern languages. English to European languages (e.g., French, German) had 54 responses (33.1%), indicating that while these languages share some similarities with English, differences in grammar, idiomatic expressions, and vocabulary still cause translation issues. English to African languages (e.g., Swahili, Yoruba) was selected by 48 respondents (29.4%), possibly reflecting the lower usage or fewer translation tools available for these languages. Finally, other languages were cited by 25 respondents (23.3%), highlighting additional language pairs where users face translation challenges.

4.7 Confidence in Machine Translations for Specialized or Technical Content

The responses indicate a varied level of confidence in machine translations for highly specialized or technical content. A significant portion of respondents (34%) feels very confident in machine translations for specialized content, likely due to advances in AI where translation models have become more adept at handling specific terminology and context in fields like medicine, engineering, or law. A smaller group (22%) trusts the technology but with some reservations, possibly acknowledging that while machine translation has improved, it still requires human oversight for precision in highly technical fields. The largest group (36.5%) expresses somewhat confident optimism, indicating that while they believe machine translation can be useful, they don't always find it fully reliable for complex, nuanced, or specialized language. This may reflect the perception that even advanced AI tools struggle with domain-specific jargon, idiomatic expressions, or nuanced meanings. A small minority (7.5%) lacks confidence in machine translations for specialized content, possibly due to concerns over inaccuracies or the inability of AI to fully understand context or maintain the subtleties of the language.

4.8 Frequency of Editing or Refining Machine Translations

The responses indicate varying levels of engagement with editing or refining machine translations. A small but significant portion of respondents (10.7%) always edit or refine their translations, likely due to working with sensitive or technical content where precision is crucial. Nearly half of the respondents (49.1%) often feel the need to manually refine translations, suggesting that while machine translations are a helpful starting point, they may still fall short in nuance, clarity, or accuracy, particularly in more complex contexts. This group likely sees machine translations as timesaving, though requiring some intervention for optimal results. Around one-fifth of respondents (20.1%) occasionally edit translations, indicating that for them, machine translations are often adequate but sometimes need additional refinement, especially when dealing with specialized content or unclear phrases. A smaller group (15.1%) rarely edits translations, suggesting they deal with simpler content or have confidence in the machine's output. Only a small minority (5%) never edit their translations, likely because they are either confident in the quality of the translation or the material being translated doesn't demand high accuracy.

4.9 Trust in Machine Translation for Important Documents

The responses highlight a cautious yet pragmatic approach to using machine translation for important documents like legal, medical, or academic content. A little over a quarter of respondents (26.4%) fully trust machine translation results for important documents. This group may be working with less complex or highly standardized content, or they may have had positive experiences where machine translations have proven accurate and reliable. They believe that with the advancement of AI, machine translations can be trusted, especially when tailored to specific domains like legal or medical fields.

The majority (49.7%) prefer to use machine translations but double-check the results with a native speaker or professional. This reflects a cautious approach, where the machine translation is seen as a helpful starting point but not sufficient for final approval, particularly for high-stakes content. For these respondents, human oversight remains essential to ensure accuracy and precision.

Around 23.9% of respondents completely avoid using machine translation for important documents, showing a lack of trust in the technology for such critical tasks. This group likely recognizes the risks of errors or misinterpretations, which could have serious consequences in fields such as law, medicine, and academia.

4.10 Likelihood of Recommending Machine Translation Tools

The responses reveal a range of opinions on the usefulness and reliability of machine translation tools, reflecting a balanced mix of trust and reservations:

- Very likely (29.6%): Almost a third of respondents would very likely recommend machine translation tools to others. This suggests that for this group, these tools are seen as reliable and effective for general translation needs, especially for everyday use, simple content, or less specialized translations. They find the tools useful and trust them for a wide range of tasks.

- Likely (21.4%): An additional 21.4% would likely recommend machine translation tools, indicating that more than half of the respondents see the value in these tools. While they are helpful, there may be some reservations about their accuracy and reliability in more specialized or high-stakes contexts.

- Neutral (27.7%): A significant portion of respondents (27.7%) is neutral, meaning they see some value in machine translation tools but don't feel strongly enough to recommend them outright. These users likely recognize both the strengths and limitations of these tools, especially when it comes to complex or sensitive content.

- Unlikely (15.1%): About 15% of respondents are unlikely to recommend machine translation tools. This group may only trust the tools for simple, informal translations and might be concerned about the accuracy of translations, particularly in professional, legal, or medical contexts.

- Very Unlikely (6.3%): A small minority (6.3%) is very unlikely to recommend machine translation tools. These respondents likely view the tools as unreliable for their translation needs, especially for more complex or specialized content. They may prefer human translation or other more accurate methods.

4.11 Will Machine Translation Replace Human Translators?

The responses reflect diverse opinions on whether machine translation tools will eventually replace human translators. Approximately 36.1% of respondents are confident that advancements in AI will lead to machine translation systems becoming reliable enough to replace human translators, particularly for standardized or less nuanced content. However, the majority, 38.6%, believe that while machine translation is progressing, it will require further advancements to handle the complexities of language, such as nuance, idiomatic expressions, and cultural context, before it can fully replace human translators. A smaller group, 25.3%, remains sceptical, emphasizing that human translators will always be essential, especially in fields that require specialized knowledge and a deep understanding of context, such as legal, medical, and literary translation.

5. Analysis

5.1 Overview of User Perceptions of Machine Translation Tools

The rapid development of machine translation (MT) technology has redefined the possibilities of cross-linguistic communication, enabling faster and more accessible translations across diverse languages and fields. The survey responses reflect this potential, showing an appreciation for the role of MT tools in breaking language barriers. Users have benefited from this accessibility, as MT tools have made it easier to translate basic and moderately complex text, facilitating communication for personal, professional, and educational purposes. However, the survey also reveals significant concerns about MT's performance, especially regarding accuracy, reliability, and context-sensitive translation. The findings indicate that while MT has made substantial progress in recent years, especially with the introduction of neural machine translation (NMT), users still experience considerable challenges when relying on MT for complex linguistic tasks.

According to Alam et al. (2022), MT's evolution has led to significant improvements in handling common languages and straightforward text structures. Yet, these advancements fall short in addressing the multifaceted challenges users face with more specialized or nuanced text. For instance, MT tools are particularly limited when dealing with legal, technical, and creative content, where precision, consistency, and contextual understanding are paramount. These texts often involve unique vocabulary, intricate sentence structures, and context-sensitive terms, which MT currently struggles to handle accurately. Respondents frequently noted that MT outputs require considerable manual adjustment, a reflection of MT's inability to meet professional standards in these domains. This aligns with previous studies indicating that while MT can be highly efficient for initial drafts, it remains inadequate for high-stakes content without human intervention (Al-Hindawi & Kher, 2019; Haffari & Tsarfaty, 2019).

The survey findings suggest a dual perception among users: MT is appreciated for its convenience and speed, but also critiqued for its limitations in accuracy and adaptability to complex linguistic nuances. Users view MT as a useful tool for basic translation needs but see it as far from reaching the depth of understanding and linguistic precision that skilled human translators can offer.

5.2 Need for Manual Adjustments: Accuracy and Reliability

A major theme in the survey responses was the frequent need for manual adjustments to MT outputs, even for relatively straightforward text. This points to a gap between MT's capabilities and the level of accuracy users require for professional or specialized content. Many users rely on MT as a starting point for translations, yet they find that it does not consistently deliver the accuracy needed, especially for high-precision fields such as legal, medical, and scientific texts. This is particularly concerning in technical fields, where even minor errors in translation can lead to serious misunderstandings or technical inaccuracies that could compromise the intended meaning or functionality of the text.

Research supports these user concerns, showing that while NMT has improved the accuracy of automated translations, it remains limited in capturing linguistic nuances, especially when dealing with technical jargon, idiomatic expressions, or complex sentence structures

(Stahlberg, 2020). For instance, Burchardt and Hossain (2020) illustrate that, despite improvements, NMT's ability to understand and translate nuanced meaning remains insufficient for contexts where precision is critical. In legal texts, for example, MT often struggles to replicate the exact terminology and phrasing needed, which can result in translations that lack the necessary rigor or clarity. Similarly, in specialized fields like medicine, translations need to accurately convey technical information while also considering cultural context and ethical implications—an area where MT tools often fall short (Mager et al., 2023).

This need for manual adjustments highlights a fundamental limitation in current MT systems. Even the most advanced MT models are unable to consistently achieve the level of accuracy required for professional translation tasks, particularly those that demand careful handling of specialized vocabulary and concepts. This finding underscores the importance of continued efforts to develop MT systems with more robust contextual understanding, allowing them to better handle jargon and idiomatic language in specialized domains. In fields where accuracy is paramount, MT may serve as an initial aid but ultimately requires substantial human intervention to ensure reliability and coherence.

5.3 Consistency and Reliability: Inconsistent Output Across Repeated Translations

Inconsistency in MT outputs emerged as a key barrier to user trust in MT systems, as reported by many survey respondents. Users noted that translating the same text multiple times often resulted in varying outputs, which can undermine the reliability of MT tools for tasks that require consistent terminology and phrasing across documents. This inconsistency likely stems from the probabilistic nature of NMT models, which generate translations based on patterns rather than strict rules. As a result, subtle changes in input or contextual variables can lead to different translations, even for identical phrases.

Consistency is especially crucial in professional or technical contexts, where terms must be uniformly translated across documents to ensure clarity and coherence. For example, in medical translation, consistent use of terminology is essential to avoid potential safety risks, as variations in word choice could lead to confusion or misinterpretation. However, the probabilistic approach of MT models can result in variations in the translation of technical terms, even within a single document. This inconsistency has also been observed in fields such as finance and law, where precise and uniform language is vital for ensuring transparency and compliance (Ulitkin et al., 2021).

Research on this issue highlights the limitations of existing MT systems in maintaining uniform terminology across documents. Studies by Hutchins (2003) and others suggest that addressing this issue will require advances in MT architecture, such as incorporating user-specific feedback or customized dictionaries. Such tools could reinforce consistency in technical vocabulary, allowing MT systems to better meet the needs of professionals who rely on precise and repeatable language.

5.4 Idiomatic and Cultural Sensitivity: Challenges in Translating Nuanced Language

Another significant issue raised in the survey responses is the difficulty MT systems face in handling idioms, slang, and culturally specific language. This challenge is particularly pronounced in contexts that involve a high degree of idiomatic expression or culturally rooted language, as these elements often carry meanings beyond their literal words. Respondents reported that MT frequently fails to accurately capture the intended meaning of idiomatic expressions, which can result in translations that feel unnatural or, in some cases, misleading.

Prior research on MT limitations with idiomatic language supports this observation, with studies indicating that NMT models struggle with expressions that lack direct translations (Popović, 2018). This challenge is even more pronounced in low-resource or indigenous languages, where MT systems often lack sufficient linguistic data to interpret and accurately translate idioms or culturally specific phrases (Mager et al., 2023). Burchardt and Hossain (2020) similarly note that while MT systems can handle literal translations well, they often produce unsatisfactory results with figurative or idiomatic language, which requires contextual understanding beyond word-for-word translation.

The difficulty in translating idioms and cultural references also limits MT's effectiveness in creative domains, such as literature, film subtitling, or advertising. Literature, for example, frequently relies on metaphor, symbolism, and character dialogue that demands a deep understanding of cultural context. In advertising, culturally resonant language is often used to engage audiences on an emotional level, and MT's inability to capture this nuance can result in translations that lack impact. This underscores the need for human translators to handle these complex linguistic tasks, as they bring an understanding of cultural subtleties and can adapt the language to suit the target audience's expectations and cultural norms (Vilar & Lopez, 2023).

5.5 Specificity of Content Types: The Case of Technical and Specialized Language

Survey responses also reveal that MT tools are often perceived as less reliable for highly specialized content, such as scientific, legal, or technical texts, where accuracy and contextual understanding are essential. Users noted that MT outputs frequently lack the precision required to convey complex details accurately, which can lead to misunderstandings or errors. For instance, legal documents require strict adherence to terminology and phrasing, and MT often struggles to replicate the exact standards needed in such documents (De Almeida & Watanabe, 2019). In engineering or medical texts, even small translation errors can have significant consequences, underscoring MT's limitations in these fields.

This finding aligns with prior research showing that MT performs poorly with domain-specific language, which requires deep expertise and contextual understanding (Hojjat & Adnan, 2020). Studies indicate that while MT systems are generally reliable for general language tasks, they often lack the specialized training needed to accurately translate industry-specific terminology or complex technical concepts (Kumar & Kaur, 2023). This challenge is especially notable for low-resource languages, where MT lacks sufficient training data to develop the nuanced understanding necessary for accurate translations.

Future research on MT may focus on developing domain-specific MT models trained on specialized language corpora, which could improve MT performance in technical fields and reduce the need for manual adjustments. Advances in contextual embeddings and unsupervised learning methods could also enable MT systems to better handle specialized language, bridging the gap between general-purpose MT and the needs of professional users who require high accuracy and consistency (Li & Yang, 2023).

5.6 Potential of Machine Translation to Replace Human Translators

When asked about the potential for MT to replace human translators entirely, respondents expressed skepticism, with many emphasizing the irreplaceable value of human linguistic and cultural understanding. While MT has made significant strides in providing basic translations across a wide range of languages, users highlighted that MT lacks the depth of understanding required to accurately convey complex or context-sensitive information. This aligns with research suggesting that while MT can handle straightforward translations well, it falls short in tasks requiring a nuanced grasp of cultural context, tone, or implied meaning (Mager et al., 2023).

Human translators bring an understanding of cultural subtleties and the ability to adapt language for emotional or aesthetic impact, which MT lacks. This is especially relevant in creative fields like literature or marketing, where translations must not only be accurate but also resonant with the target audience. Hojjat and Adnan (2020) argue that MT's inability to capture these nuances restricts its effectiveness in fields where emotional resonance, cultural adaptation, and audience engagement are crucial. In creative fields, the translation often involves more than just converting words from one language to another—it requires interpreting meaning, tone, and intent in a way that feels authentic to the target culture. For example, translating a marketing slogan involves not only linguistic accuracy but also an understanding of local values, humor, and consumer behavior. Here, human translators excel due to their cultural awareness and adaptability, qualities that MT systems currently lack.

However, the potential for MT to serve as a complementary tool to human translation is clear. In recent years, MT has proven to be a useful aid for human translators, providing initial translations that professionals can refine and adapt. This synergy is especially valuable for high-volume, low-stakes content, such as technical manuals or user-generated content, where MT can reduce the time spent on basic translations, leaving human translators to focus on more intricate tasks. Wu and Hu (2023) discuss the potential of integrating prompt engineering to enhance MT outputs, allowing MT systems to produce more context-aware translations that human translators can subsequently improve. This points to a future where MT and human translators work collaboratively, with MT handling preliminary translations that humans refine to achieve a polished, culturally nuanced final product.

5.7 Implications for Future MT Development

The survey responses underscore a clear demand for MT tools to evolve in several critical areas. Firstly, users express a need for improved accuracy, especially in specialized and professional contexts. While MT has made strides in handling common language patterns, it still lacks the precision required for specialized language domains. To meet this need, future MT development may focus on enhancing domain-specific models. By training MT systems with specialized corpora from fields such as law, medicine, and engineering, developers can improve MT's ability to handle technical terms, complex structures, and field-specific nuances.

Moreover, recent advances in unsupervised learning and contextual embeddings show promise in enabling MT systems to capture deeper contextual meaning. Research suggests that by integrating these technologies, MT systems could become better equipped to manage the varied demands of different content types, making them more versatile and reliable tools for users across diverse fields (Li & Yang, 2023). Additionally, the development of user-customizable MT features—such as personalized glossaries, translation memory, or feedback loops that learn from user corrections—could further improve MT performance, enabling users to tailor MT outputs to their specific needs and vocabulary preferences.

Another area for growth lies in enhancing MT's handling of idiomatic and culturally specific language. This challenge will require MT developers to incorporate more cultural data, idiomatic expressions, and real-world scenarios into MT training datasets. Some researchers propose leveraging community-driven data collection or feedback systems to expand MT's understanding of culturally resonant language. Moreover, advances in cross-cultural linguistics and sociolinguistics could inform MT training processes, allowing MT systems to better interpret and translate culturally embedded language that goes beyond simple word-to-word conversion (Popović, 2018).

5.8 Balancing MT and Human Translation: The Road Ahead

While MT has transformed the landscape of language translation, enabling unprecedented access to cross-linguistic communication, it is unlikely to fully replace human translators in the foreseeable future. Survey respondents generally agreed that MT's role is best seen as complementary to human translation rather than as a substitute. Human translators bring a level of cultural insight, emotional intelligence, and linguistic nuance that is difficult for MT systems to replicate. This is particularly true in areas like creative writing, diplomacy, and legal negotiations, where language plays a critical role in conveying complex social and cultural meanings.

Moving forward, MT is likely to continue evolving as a tool that works in tandem with human translators. As MT technology advances, the boundary between human and machine translation may become more fluid, with MT handling more routine or straightforward translations and human translators taking on roles that require higher-order linguistic and interpretive skills. MT's ongoing development will also likely focus on areas such as ethical AI and culturally sensitive design, ensuring that MT tools are adaptable, inclusive, and capable of responsibly managing the intricacies of cross-cultural communication (Mager et al., 2023).

6. Conclusion

The survey results underscore a broad user consensus on the utility and limitations of current MT systems. While MT has evolved as an essential tool for facilitating communication and enabling rapid translations across languages, several critical limitations persist that hinder its full adoption in professional and nuanced linguistic tasks.

6.1 Key Findings and Implications

1. Accuracy and Reliability: Users generally find MT tools useful for initial translation drafts but insufficient for tasks requiring high accuracy and linguistic nuance. This is in line with existing research that points to accuracy issues, particularly in complex or specialized content areas (Al-Hindawi & Kher, 2019; Stahlberg, 2020).

2. Need for Manual Adjustments: The survey indicates that users often need to manually adjust MT outputs, especially for idiomatic or technical content. This reflects broader trends observed in translation quality assessment studies, suggesting that while MT can handle straightforward text, it struggles with complex language structures (Popović, 2018; Burchardt & Hossain, 2020).

3. Consistency Issues: Inconsistent translations in repeated MT outputs have raised reliability concerns. This finding aligns with research on the probabilistic nature of NMT models, which, while advanced, can introduce variability in translation outputs even for identical texts (Ulitkin et al., 2021).

4. Challenges in Technical and Idiomatic Language: MT systems face significant limitations in accurately translating idiomatic expressions, slang, and technical jargon. This is particularly concerning for specialized fields where accuracy and cultural understanding are paramount, reinforcing the need for human translators in such domains (Mager et al., 2023; Kumar & Kaur, 2023).

5. Human vs. Machine Translation: The skepticism toward MT replacing human translators underscores the irreplaceable value of human expertise in capturing cultural and contextual nuances. Existing literature echoes this sentiment, suggesting that while MT has considerable utility, its role remains complementary to human translation, particularly for linguistically complex or culturally embedded languages (Mager et al., 2023; Hojjat & Adnan, 2020).

6.2 Recommendations for Future Research and Development

Given these findings, future MT development should focus on enhancing contextual understanding, consistency, and domain-specific translation accuracy. Innovations in unsupervised and semi-supervised learning models could improve MT performance for low-resource languages, addressing the current lack of comprehensive training data (Li & Yang, 2023). Additionally, exploring new evaluation metrics that better capture translation nuance and cultural appropriateness would help align MT outputs more closely with human translations (Wang & Kurohashi, 2022).

Moreover, research on integrating user feedback into MT systems could improve translation accuracy and reliability over time, offering a more tailored and responsive translation experience. Promising work on prompt engineering for document-level translations offers potential for MT systems to handle context-rich content more effectively, a significant step toward achieving higher-quality translations (Wu & Hu, 2023).

In conclusion, while MT systems have made remarkable strides, the survey results reinforce the prevailing view that human translators remain indispensable for high-stakes, nuanced linguistic tasks. As MT technology continues to advance, balancing automated efficiency with human linguistic expertise will be essential in delivering reliable and culturally resonant translations.

References

1. Alam, F., Anastasopoulos, A., Bhagia, A., Costa-jussà, M. R., Dodge, J., Faisal, F., ... & Wenzek, G. (2022). Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages. In Proceedings of the Seventh Conference on Machine Translation (WMT) (pp. 565-587). Association for Computational Linguistics.

https://aclanthology.org/2022.wmt-1.60

2. Al-Hindawi, A. A., & Kher, A. A. (2019). Quality evaluation of machine translation: A review. Procedia Computer Science, 159, 1211-1218.

https://doi.org/10.1016/j.procs.2019.12.180

3. Burchardt, A., & Hossain, A. (2020). Neural machine translation: The good, the bad, and the ugly. Journal of Machine Translation, 34(1), 57-87. https://doi.org/10.1007/s10590-020

09280-8

4. De Almeida, G. A., & Watanabe, A. H. (2019). Evaluating machine translation quality with contextual information. Proceedings of the 7th Workshop on Asian Translation, 40-50.

https://doi.org/10.18653/v1/W19-6506

5. Haffari, G., & Tsarfaty, R. (2019). Human vs. machine: The evaluation of translation quality. In Advances in Natural Language Processing (pp. 1-15). Springer.

https://doi.org/10.1007/978-3-030-39484-4_2

6. Hojjat, S., & Adnan, A. (2020). Translation quality assessment: From principles to practice. Translation Quality Assessment: From Principles to Practice. Springer.

https://doi.org/10.1007/978-3-319-79263-7

7. Hutchins, J. (2003). Machine translation and computer-based translation tools: what’s available and how it’s used. Edited transcript of a presentation at the University of Valladolid, Spain.

8. Kumar, R., & Kaur, J. (2023). Investigating machine translation for low-resource languages: Challenges and solutions. Journal of Language Technology.

https://doi.org/10.1007/s10311-023-00640-3

9. Kahlon, N. K., & Singh, W. (2023). Machine translation from text to sign language: A systematic review. *Universal Access in the Information Society*, 22, 1–35.

https://doi.org/10.1007/s10209-021-00823-1

10. Li, Y., & Yang, Y. (2023). Exploring unsupervised learning for machine translation: Methods and challenges. Computer Speech & Language.

https://doi.org/10.1016/j.csl.2023.101532

11. Liu, Y., et al. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiology, 1(2), 100017.

https://doi.org/10.1016/j.metrad.2023.100017

12. Mager, M., Mager, E., Kann, K., & Vu, N. T. (2023). Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 4871–4897). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2023.acl-long.268

13. Popović, M. (2018). Error classification and analysis for machine translation quality assessment. In Translation quality assessment: From principles to practice (pp. 129-158).

14. Snover, M., & Dorr, B. (2006). A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers (pp. 223-231). Association for Machine Translation in the Americas. https://aclanthology.org/2006.amta-papers.25

15. Stahlberg, F. (2020). Neural machine translation: A review. Journal of Artificial Intelligence Research, 69, 343-393. https://doi.org/10.1613/jair.1.12007

16. Toral, A., & Way, A. (2023). A survey on evaluation metrics for machine translation: Effectiveness and shortcomings. Journal of Artificial Intelligence Research, 67, 897-929.

https://doi.org/10.1613/jair.1.12832

17. Ulitkin, I., Filippova, I., Ivanova, N., & Poroykov, A. (2021). Automatic evaluation of the quality of machine translation of a scientific text: The results of a five-year-long experiment. E3S Web of Conferences, 284, 08001. https://doi.org/10.1051/e3sconf/202128408001

18. Vilar, D., & Lopez, J. (2023). Towards a comprehensive evaluation of machine translation systems. Machine Translation. https://doi.org/10.1007/s10590-023-09311-x

19. Wang, Y., & Kurohashi, S. (2022). Towards better evaluation metrics for neural machine translation. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 582-590. https://doi.org/10.18653/v1/2022.emnlp-main.39

20. Wu, Y., & Hu, G. (2023). Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In Proceedings of the Eighth Conference on Machine Translation (WMT23) (pp. 166-169). Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.95