Jurnal Kata : Penelitian tentang Ilmu Bahasa dan Sastra
Volume 10, Nomor 1, Tahun 2026
E-ISSN: 2502-0706
Open Access: http://publikasi.lldikti10.id/index.php/kata

READABILITY ASSESSMENT OF AI-DRIVEN INDONESIAN JOURNALISTIC TRANSLATIONS: COMPARING CHATGPT AND DEEPL PERFORMANCE

Farizka Humolungo[1], Fanny Puji Rakhmi[2], Ina Sukaesih[3], Ince Dian Aprilyani Azir[4] Talenta Gloria Napitupulu[5]

Administrasi Bisnis, Politeknik Negeri Jakarta
[1]e-mail: *farizka.humolungo@bispro.pnj.ac.id*, [2]e-mail: *fanny.pujirakhmi@bisnis.pnj.ac.id*,
[3]e-mail: *ina.sukaesih@bisnis.pnj.ac.id*
[4]Institute of Education, University College London, e-mail: *ince.azir.23@ucl.ac.uk*
[5]Administrasi Bisnis, Politeknik Negeri Jakarta, e-mail: *talenta.gloria.napitupulu.an23@stu.pnj.ac.id*

Article history:

Received
23 January 2026

Received in revised form
27 April 2026

Accepted
30 May 2026

Available online
May 2026

Keywords:
ChatGPT; DeepL; Indonesian to English Translation; Journalistic Texts; Readability.

DOI
10.22216/kata.v10i1.3493

Abstract

This study investigates the readability of Indonesian to English translations of journalistic texts produced by ChatGPT and DeepL across the domains of law, politics, and environment using articles from Tempo and Kompas. A qualitative descriptive comparative design was employed. Six source texts were translated under controlled procedures, with standardized prompts for ChatGPT and standard input for DeepL. Two expert raters evaluated segment level readability using a three-point rubric adapted from prior work. Before scoring, the raters reviewed and aligned category descriptors to support shared interpretation. Readability scores were summarized by outlet and topic, and rater comments from a focused discussion were used to contextualize patterns. The findings show that ChatGPT was more often perceived as easier to read because it tended to smooth sentences and improve flow, whereas DeepL more often kept the original structure and details, which can help preserve precision but sometimes made the English feel stiff. The results point to domain sensitive strengths for each system and suggest targeted post editing as a practical lever to support clarity for public facing news.

INTRODUCTION

Recent developments in artificial intelligence (AI), especially large language models and neural machine translation, have made tools such as ChatGPT and DeepL widely used in both academic and professional contexts (Basha, 2024; Iqbal et al., 2024; Ou, 2024). Many people now rely on these tools to speed up routine tasks, shorten turnaround time, and generate structured content that previously required fully manual work (C. C. Lin et al., 2023; Rizvi, 2023). This shift is particularly visible in translation, where texts often need to be produced quickly, in large quantities, and in a way that still communicates meaning clearly across languages and cultures.

In education, AI is frequently described as a support tool that can help learners process information, make decisions, practice interactively, complete academic tasks, and solve problems (Misnawati, 2023; Fitriani et al., 2024; Suharmawan, 2024). In translation work, ChatGPT and DeepL are often viewed as useful because they can produce translations rapidly and, in many cases, with reasonably strong baseline quality (Handoyo et al., 2023). However, the value of AI translation cannot be judged by speed or surface-level correctness alone. Research in language education and translation studies emphasizes that effective language use depends on meaning-making in context supported by knowledge of discourse conventions, communicative purpose, and cultural references (Omaggio, 2001; Richards &

Corresponding author.
E-mail address: fanny.pujirakhmi@bisnis.pnj.ac.id

Rodgers, 2001). For that reason, AI translation quality should be assessed not only in terms of accuracy, but also in terms of whether target readers can understand the message easily and interpret it as intended in real communicative situations.

Existing studies also show that AI translation performance is not uniform. It varies by language pair, text type, and discourse complexity. Across the literature, a pattern shows that AI tends to perform better for high-resource language pairs than for lower-resource contexts. Another patterns suggest that idioms, pragmatic meanings, and culturally grounded references remain difficult and often lead to overly literal translations, and even when a translation reads fluently, it may contain omissions or subtle shifts that require human judgment to ensure appropriateness (Hendy et al., 2023; Gadd, 2024; Cetin & Duran, 2024). Research further suggests that genre matters. In some studies of literary and poetic translation, ChatGPT has been reported to produce more natural output and preserve meaning more effectively than certain other tools, although human translations still achieve higher overall quality (Gao et al., 2024; Sun, 2024). Taken together, these findings imply that "quality" in AI translation is multi-dimensional and depends heavily on what kind of text is being translated, so evaluations focused only on accuracy may overlook real-world readability and comprehension.

Readability refers to the degree to which a translated text can be easily understood by the target readers (Nababan et al., 2012). Readability ensures that the message from the source language is expressed naturally and clearly, so that it is easy for the audience to comprehend. A readable translation does not force readers to struggle with awkward wording, confusing sentence structure, or unfamiliar expressions. Instead, it presents the message in a clear, smooth, and accessible way. This means that even if the meaning of the source text has been transferred correctly, the translation may still be considered weak if readers find it difficult to follow. In other words, readability focuses on the reader's experience, whether the translation feels understandable and effortless to process.

Although the research base is growing, much of it still prioritizes accuracy-based evaluation and often uses academic, literary, or specialized texts. This leaves an important gap regarding journalistic translation, where readability is essential because news is intended for broad public audiences. Journalistic writing, especially in politics, law, and environmental reporting, often includes dense vocabulary, domain-specific terms, institutional phrasing, and culturally specific references. When such texts are translated, readability becomes a basic requirement for access. If translated news is hard to follow, readers may misunderstand key points or lose trust, which undermines the social role of journalism.

There is also limited comparative evidence on how ChatGPT and DeepL perform in Indonesian–English journalistic translation, particularly from the perspective of readability. While earlier studies have compared AI outputs with human translation and noted recurring problems with idioms and cultural nuance (Cetin & Duran, 2024; Gao et al., 2024), fewer studies have systematically examined journalistic genres, focused on the Indonesian–English direction, and used human raters to evaluate readability across multiple news domains. As a result, it remains unclear how and to what extent ChatGPT and DeepL differ in producing readable Indonesian–English journalistic translations, and whether the difficulty level changes across law, environment, and politics. This gap supports the novelty of the present study, which shifts attention from accuracy-dominant evaluation toward readability-centered assessment in a socially important genre while directly comparing two widely used AI tools in an Indonesian–English context.

Therefore, this study aims to evaluate and compare the readability of Indonesian–English journalistic translations generated by ChatGPT and DeepL in three domains: law, environment, and politics. To achieve this aim, the study (i) selected representative journalistic texts across the three domains, (ii) generated translations using ChatGPT and

DeepL, (iii) applied rater-based assessment to evaluate readability and clarity, and (iv) analyzed rater feedback to identify recurring strengths, weaknesses, and domain-related patterns that may reflect pragmatic, contextual, and cultural limitations in AI outputs (Hendy et al., 2023; Gao et al., 2024).

Accordingly, this study addresses the following research questions: (1) How does the readability of AI-generated translations differ between ChatGPT and DeepL across journalistic domains such as law, environment, and politics? (2) What insights can be inferred from rater evaluations regarding the overall performance and limitations of AI-based translation tools, particularly in relation to pragmatic, contextual, and cultural nuance as reflected in readability and clarity?

RESEARCH METHOD

This study employed a qualitative descriptive design with a comparative approach, and it was conducted to examine differences in readability between Indonesian–English journalistic translations produced by ChatGPT and DeepL (Creswell, 2018; Rusandi & Rusli, 2021). The research object comprised AI-generated translation outputs derived from Indonesian news texts representing three journalistic domains—law, environment, and politics—selected to reflect vocabulary density, domain-specific terminology, and culturally embedded expressions commonly found in public-information discourse. The study was carried out by compiling source texts, generating translations using both tools under standardized procedures, and evaluating the outputs through expert rater assessment in order to identify tool-based patterns of readability and domain-based variation. General procedures for qualitative descriptive inquiry and comparative analysis followed established guidance in qualitative research and translation studies (Creswell, 2018; Saldanha & O'Brien, 2013).

The scope of the data consisted of six Indonesian news articles drawn from two major Indonesian news portals, Tempo.co and Kompas.com. Each portal contributed three articles—one for each domain (law, environment, politics)—to ensure balanced thematic representation across sources. The unit of analysis was the full translated text for each article (Indonesian source text translated into English), resulting in 12 translation products in total (six translations produced by ChatGPT and six produced by DeepL). The articles were selected based on (i) thematic fit with the three domains, (ii) representativeness of journalistic language use (e.g., institutional terms, formal register, and topic-specific lexicon), and (iii) publication recency at the time the dataset was compiled.

Data collection was conducted in three stages. First, the researchers retrieved and archived the six selected Indonesian articles from Tempo.co and Kompas.com. Second, translations were generated using two AI tools under controlled and replicable conditions. For ChatGPT, the researchers applied the same standardized instruction across all texts to minimize prompt variability and ensure consistent task framing. For DeepL, translations were produced using the platform's standard text-input procedure, reflecting typical user practice. Third, the 12 translation outputs were prepared for assessment and anonymized by removing tool identifiers so that raters could evaluate readability without being influenced by the source of the translation.

Readability assessment was conducted by two expert raters who had served as university lecturers in translation-related fields and possessed professional and academic experience in evaluating translation quality. The raters were recruited purposively because their expertise aligned with the analytic needs of the study, particularly in judging naturalness, clarity, and comprehensibility of English target texts. Each rater evaluated all translation outputs independently using a readability rubric adapted from Nababan et al. (2012), which operationalizes readability into a three-point scale: (1) low readability, defined as unclear, unnatural, or difficult to understand; (2) moderate readability, defined as generally

understandable but containing awkward expressions or partial clarity; and (3) high readability, defined as natural, clear, and easy to understand. Routine procedures for rater-based scoring were implemented in line with standard rubric-based assessment practice; therefore, operational details that are methodologically common were treated as implied and anchored to the rubric reference (Nababan et al., 2012).

Data analysis was carried out by compiling the rater scores for each translation output and organizing them by tool (ChatGPT versus DeepL) and by domain (law, environment, politics). The researchers calculated average readability scores for each tool within each domain and used a comparative matrix to identify patterns of convergence and divergence across the two systems and three themes. The analysis also incorporated qualitative explanatory data from rater comments. After the independent scoring stage, the raters participated in a focused discussion to clarify their judgments, elaborate on recurring issues, and provide examples of features that reduced readability (e.g., unnatural collocations, ambiguous referents, overly literal phrasing, or culturally bound expressions rendered without pragmatic adaptation). These comments were then coded thematically to support interpretation of the numerical tendencies and to explain why particular domains or tools exhibited higher or lower readability. The integration of score comparison and rater commentary were used to generate inferences regarding the strengths and limitations of each AI tool in producing readable Indonesian–English journalistic translations, particularly when translating domain-specific discourse with pragmatic and cultural constraints.

In analyzing the data, the study applied a simple rubric at each step to ensure consistency and transparency. First, data organization was considered *adequate* when all 12 translations had complete scores from both raters and were correctly labeled by tool and domain; it was *less adequate* when minor score/label issues appeared but could be corrected. Second, readability scoring followed the three-level rubric adapted from Nababan et al. (2012): a score of 3 indicated the translation was clear and natural in English, 2 indicated it was generally understandable but contained awkward or partially unclear expressions, and 1 indicated it was difficult to understand. Third, comparison across tools and domains was treated as *strong* when average scores were clearly summarized in a tool × domain matrix and supported by the score patterns. Fourth, when raters gave different scores, the case was reviewed through discussion and the reason for the difference was noted to support accountability. Finally, rater comments were coded into recurring themes (e.g., unnatural wording, ambiguity, overly literal phrasing, and cultural expressions) and were considered *well-developed* when each theme was supported by clear examples from the translations, so that the qualitative notes could explain the numerical trends.

RESULT AND DISCUSSION

Based on the readability assessment and rater feedback, three findings were identified. First, legal texts were the most challenging domain for readable AI translation because they contained dense institutional terminology, Indonesian-specific acronyms, and complex modifier chains. Second, ChatGPT and DeepL showed broadly comparable readability overall, but they achieved it through different strategies: ChatGPT tended to improve flow through restructuring and rephrasing, while DeepL tended to preserve source sequencing and detail. Third, domain characteristics shaped readability outcomes more strongly than outlet differences, although outlet style still contributed to variation. These findings address the first research question by showing how readability differed between tools across law, environment, and politics, and they address the second research question by indicating which linguistic and discourse factors raters associated with higher or lower readability.

Thematic Challenges in Translating Journalistic Texts

Table 1 below illustrates the distribution of data based on three journalistic themes: politics, law, and environment. The highest of data entries is found in news articles related to law, totaling 81 entries or 56% of the overall dataset. This is followed by environmental news with 39 entries (26%), and political news with 26 entries (18%), making it the least represented in the dataset.

Table 1. Distribution of Data Based on News Themes

Theme	Amount of Data	Data Percentage (%)
Politics	26	18
Law	81	56
Environment	39	26
Total	146	100

Law-themed

The predominance of law-themed articles reflects the frequent coverage of legal issues in Indonesian media, which may be attributed to the country's dynamic legal landscape and public interest in high-profile cases, regulations, and law enforcement. The complexity and specificity of legal terminology also introduce significant challenges for machine translation systems. Legal-themed journalistic texts often contain jargon, formal structures, and culturally bound expressions that require precise and context-aware translation. These features demand not only linguistic accuracy but also a deep understanding of legal systems, making it harder for machine translation tools to produce natural and faithful equivalents in the target language.

One of the key challenges for machine translation systems when processing legal news is the frequent use of institutional and legal terminology, particularly those involving abbreviations and acronyms that are culturally specific to Indonesia. Examples include POLRI (Indonesian National Police), AKBP (Ajun Komisaris Besar Polisi or Police Superintendent), Komnas HAM (National Commission on Human Rights), Satgas (Task Force), TPNPB (West Papua National Liberation Army), KUHP (Criminal Code), and Bawaslu (Election Supervisory Board). These abbreviations are often utilized without elaboration on the source text and require context-specific knowledge to be translated accurately and meaningfully into English.

Machine translation tools like ChatGPT and DeepL may struggle with these terms, especially when direct equivalents do not exist in the target language. A literal or incorrect translation could lead to loss of meaning, ambiguity, or misinterpretation of the legal context. This makes the law-themed texts not only more abundant but also inherently more challenging to translate. The existence of such specialized vocabularies offer a critical benchmark to evaluate the translation tools' capabilities in handling nuanced, institutional, and domain-specific language in Indonesian journalistic discourse.

Environment-themed

Environmental news represents the second-largest portion of the dataset, with 39 entries or 26% of the total. In the Indonesian context, environmental issues are highly relevant, particularly in relation to plastic waste, marine pollution, and deforestation. Among these, plastic pullution has received growing media attention in recent years.

One significant feature of environmental journalism is that it often involves common terminology, much of which is borrowed directly from English. Terms such as *mikroplastik* (microplastic), *akuatik* (aquatic), biodegradable, and sustainable are often used in their

original or naturalized forms in Indonesian news articles. This linguistic overlap can make translation appear easier on the surface, especially when dealing with technical or scientific vocabulary that is already standardized across languages.

However, while the terminology may be familiar to both source and target languages, challenges remain. A key difficulty lies in ensuring contextual appropriateness and consistency. For instance, although a term like *mikroplastik* may have a direct equivalent in English (microplastic), the surrounding description, implications, or cultural framing may differ between Indonesian and English discourse. Machine translation systems may struggle with capturing nuanced descriptions or the scientific tone required in environmental reporting.

Another challenge involves the frequent mention of geographical names and local placed-based references, such as rivers, beaches, islands, or provinces where environmental issues occur. These proper nouns are typically not translated, yet they often appear alongside technical information that must be translated clearly. For instance, a sentence mentioning pollution in Sungai Citarum (Citarum River) must preserve the place name while accurately conveying the environmental condition described.

For high-resource languages like English, both ChatGPT and DeepL benefit from extensive training data, including scientific and environmental texts, which enables them to accurately translate many technical terms such as *mikroplastik* or *akuatik* easily. However, the true test lies in how well these systems handle the integration of local context, particularly when translating Indonesian place names, culturally specific references, or regionally framed environmental narratives.

Politics-themed

Political news in this dataset contributes the smallest portion with 26 entries or 18% of the total. This suggests that political news article, at least those sampled in this study, tend to be shorter compared to legal or environmental news. The briefness of political texts may result from the nature of the reporting, which often focuses on event-driven coverage such as press statements, campaign updates, or official announcements, rather that in-depth analysis.

In terms of language, political texts share several characteristics with legal texts, especially in their frequent references to governmental and institutional actors. Terms such as POLRI, Komnas HAM, and KPU (General Elections Commision) regularly appear, highlighting the interrelation between law enforcement and political processes. What distinguished political news, however, is the presence of election-related terminology, including culturally specific phrases like *pemilu* (general election), *caleg* (legislative candidate), and *kampanye* (political campaign). These expressions are tightly connected to Indonesia's political system and require context-sensitive translation to retain their full meaning in English.

Difficulties arise when machine translation tools like ChatGPT and DeepL encounter culturally bound political expressions that lack direct equivalents in English. For instance, translating *BAWASLU* as "Election Supervisory Board" may be correct linguistically, but it still requires contextual explanation to convey its unique role in Indonesia's electoral system. Similarly, acronyms and shorthand terms widely recognized in Indonesian media may be left untranslated or awkwardly rendered if the system lacks contextual understanding.

In this sense, while ChatGPT and DeepL can navigate the structural aspects of political texts with relative ease, they still face challenges in ensuring that translations preserve the institutional and cultural specificity of Indonesia's political environment. It is important to maintain precision in rendering election-related terms to avoid misinterpretation, especially when translating content for international readers unfamiliar with the Indonesian political landscape.

Readability Scores

Here are the readability scores of ChatGPT and DeepL's translations covering three main themes: law, politics, and environment. The data analyzed in this study were drawn from two leading Indonesian media platforms–Kompas and Tempo. To assess the quality of translation output, two independent raters were involved, each evaluating the readability of translations produced by both tools. Table 2 shows the varying perceptions between the two raters. In some cases, both raters provided similar evaluations, while in others, their judgments differed notably, indicating a degree of subjectivity in assessing textual readability. These variations highlight the nuanced nature of translation assessment, particularly when dealing with complex journalistic content.

Table 2 Readability Scores

Media & Theme	Data Total	ChatGPT		DeepL	
		RATER 1	RATER 2	RATER 1	RATER 2
TEMPO 1 - Law	17	2.5	2.6	1.9	2.6
TEMPO 2 - Politics	16	2.8	2.8	2.3	2.8
TEMPO 3 - Environment	7	2.1	3	2.9	3
KOMPAS 1 - Law	64	2.9	2.6	2.4	2.7
KOMPAS 2 - Politics	10	2.4	3	2.5	2.9
KOMPAS 3 - Environment	32	2.7	2.9	2.4	2.9
TOTAL & AVERAGE	**146**	**2.6**	**2.8**	**2.4**	**2.8**

There are prominent differences in translation style between ChatGPT and DeepL. ChatGPT tends to summarize or compress article content, and in some cases, even introduces subheadings that are not present in the original source. While this approach can improve flow and readability for new readers, it also carries the risk of omitting key details or changing the original intent of the text. In contrast, DeepL tends to adhere more closely to the source material, maintaining the structure, sequence, and details of the original article. Although this can sometimes result in a translation that reads more rigidly, it offers greater fidelity to the source, which is valuable in contexts where precision and completeness are essential.

Readability of ChatGPT's translation

The readability scores of ChatGPT's translations show generally consistent performance across media sources. For Tempo articles, ChatGPT received an average score of 2.5 and 2.8 from the two raters. Meanwhile, Kompas articles scored slightly better at 2.6 and 2.8. These results suggest that ChatGPT's translations are perceived as slightly more readable when translating Kompas articles compared to Tempo. The consistent score of 2.8 from one of the raters across both media also indicates a stable perception of ChatGPT's readability.

However, these scores vary depending on the theme of the text. For legal-themed texts (*Law*), ChatGPT received average readability scores of 2.7 and 2.6, indicating relatively high readability. Political-themed texts (*Politics*) were rated slightly lower by one rater (2.6) but higher by another (2.9), suggesting some variation in perception but overall consistent readability. Environmental-themed texts (*Environment*) received the lowest score from one rater (2.4) but the highest from the other (3.0), indicating that while the readability may be more variable in this category, it can still be rated highly. Overall, ChatGPT maintains a

generally strong level of readability across different themes, with minor variations likely influenced by the nature of the content.

Readability of DeepL's translation

On the other hand, the readability scores of DeepL's translations show a slightly lower but still consistent performance. For Tempo articles, DeepL received scores of 2.3 and 2.8, while for Kompas articles, the scores were 2.4 and 2.8. These results indicate that although one rater consistently found DeepL's translations to be equally readable across both media, the other rater rated them slightly lower, especially for Tempo articles. This suggests that DeepL's commitment to source fidelity may come at the expense of flow and naturalness, leading to a translation that is more accurate but potentially less fluid for the reader.

The scores also vary depending on the text theme. For law-themed texts (*Law*), DeepL received scores of 2.1 and 2.6, indicating lower readability compared to other themes, especially from one rater. Political-themed texts (*Politics*) were rated more consistently at 2.4 and 2.9, suggesting a moderate to high level of readability. For environment-themed texts (*Environment*), DeepL received higher scores of 2.6 and 2.9, showing that its translations in this category were perceived as more readable. Overall, DeepL's readability performance improves with political and environmental content but tends to be less strong with legal texts.

Evaluative Insights from Translation Outputs

Readability patterns across the corpus crystallize around a small set of stable contrasts that interact with outlet, topic, and rater preference. Across the full set of materials, the translations produced by ChatGPT tend to be read as slightly more fluent by one rater, while the second rater judges the two systems as essentially equivalent overall. This recurring split is not random variation but a predictable effect of how each system balances fluency and literalness on Indonesian journalistic prose. This split can be read through the lens of *interference from the source text* versus *standardization* in translation behavior: outputs that preserve source sequencing tend to show stronger source-text interference, while outputs that smooth and normalize structures can appear more "standardized" and thus more readable in the target language (Toury, 1995). ChatGPT typically smooths clausal stacking, evens out parataxis, and supplies connective tissue that helps readers track topic–comment relations; the result is a more flowing English paragraph where sentence openings are varied and clause-final modifiers are reattached in ways that align with English information structure. DeepL's outputs, by contrast, more visibly preserve Indonesian syntactic sequencing and explicitness; this helps retain detail and institutional terminology with minimal risk of omission, but it can also reproduce calques or dense nominal strings that read stiffly in English even when they are perfectly faithful. A directly relevant comparative study between ChatGPT and DeepL reports systematic differences in how each system handles meaning reconstruction and risk of misunderstanding, supporting the idea that "more readable" can be tied to how much restructuring a system performs (Sun, 2024). In other words, the systems employ two complementary strategies that are both defensible on quality, and the raters' different emphases, one slightly favoring cohesion and idiomaticity, the other rewarding careful retention of source structure and terms, reliably propagate into their readability scores. This dependence on evaluator preference is also consistent with user-perspective evaluation comparing LLMs and NMT, showing that perceived translation performance varies with what users (or raters) prioritize (e.g., readability/naturalness vs. faithful transfer) (Son & Kim, 2023).

These global tendencies become more legible when examined through the lens of media outlet. Texts from Kompas, as represented in the dataset, invite higher readability on

average for both systems and both raters than texts from Tempo. The simplest explanation, consistent with the materials, is that the Kompas selections display more regular paragraphing, more explicit local topic sentences, and fewer stacked institutional qualifiers, all of which make both neural approaches look better downstream. This outlet sensitivity is compatible with news-translation scholarship that frames news translation as shaped by institutional routines and genre norms (e.g., how information is packaged for audiences), meaning that clearer source-text scaffolding can yield more readable target texts even before "tool differences" are considered (Troqe & Marchan, 2017). Tempo's selections, especially in its law subset, often concentrate dense references to Indonesian bodies, ranks, statutes, and acronyms that have no routine stock translation in English. DeepL's choice to keep sequence and form intact protects against accidental semantic drift but can lead to long stretches where clause-final legal qualifiers are chained together in English without the rhetorical cushioning that a native editor would supply. ChatGPT's choice to compress and rephrase these chains is often an advantage for readability, especially when long sentences are split and support details are distributed more evenly across a short paragraph. Yet the very same mechanism can attenuate or bury a modifier that is legally consequential if the model's salience heuristics underweight it. The divergence we see in Tempo's law items, where one rater markedly prefers ChatGPT while the other sees the two systems as similar, maps neatly onto this stylistic trade-off. This is a classic justification for post-editing: using an MT/AI draft as a base, then revising selectively to improve readability while controlling for omissions or terminological drift (O'Brien & Simard, 2014; Vieira, 2019).

Topic interacts with outlet to produce further texture. In the law subset, the items that contain dense legal formulations and highly culture-specific institutional references are the most challenging to render as readable English without editorial assistance. This is not merely a matter of vocabulary; it concerns genre-specific conventions. Indonesian legal journalism frequently foregrounds institutional actors and legal bases with sequences of acronyms, rank markers, and statutory references that are entirely natural for local readers but can overload English sentences if translated literally clause by clause. DeepL's fidelity preserves the semantic spine, which is critical for downstream fact-checking, yet the resulting English sometimes clusters too many modifiers before the main predicate or tucks them into long postmodifying phrases that slow processing. ChatGPT's smoothing is often exactly what a human editor would attempt such as split the sentence, bring the actor to the front, keep the legal basis explicit, and redistribute the modifiers to maintain emphasis, so the readability gains are real. However, in the absence of a strict prompt to preserve every qualifier verbatim, there are cases where a low-salience detail is compressed, which a rater attentive to legal nuance will notice. The data support this duality: a consistent, albeit modest, preference for ChatGPT's flow on law items by one rater coexists with near-parity judgments by the other rater, who appears to value explicitness highly and penalizes even small risks of compression-related ambiguity (Nababan et al., 2012). This finding also aligns with context-level MT evaluation research showing that systems differ in how they handle discourse phenomena such as reference, ellipsis, and terminology across larger spans—factors that can affect readability and perceived clarity (Castilho et al., 2023).

The politics subset sits closer to the center of the readability distribution. Much of the terminology related to election such as *pemilu, caleg*, *kampanye* maps to well-established English equivalents, and the sentence architecture of the source is comparatively less culture-bound than in legal discourse. Both systems therefore start from a friendlier baseline. The raters' scores converge, suggesting that the differences in stylistic preference have less to latch onto when the source texts are less dense with institutional qualifiers. ChatGPT's cohesion advantages are still present, manifesting as fewer ungainly strings of premodifiers and more varied sentence openings, while DeepL's sequence retention continues to guarantee

that list structures and parallelisms survive the transfer. But because the core lexicon is familiar and the factual burden is distributed across shorter propositions, the gap between an idiomatic paraphrase and a faithful calque narrows, and so do readability judgments. This convergence is consistent with user-perspective comparisons indicating that when expressions have stable equivalents, and the discourse load is lower, perceived differences between LLM and NMT outputs often become less pronounced (Son & Kim, 2023).

Environmental reporting shows yet another pattern. Here, the outputs of both systems are typically judged highly readable by both raters. The reason is straightforward: many key terms in environmental discourse are internationalized and English-aligned such as *microplastics, biodegradable materials, aquatic ecosystems, circular economy*, so neither system needs to manufacture an idiomatic analogue in the way legal discourse often does. DeepL's literal carryover is especially well tuned to this domain; when a source sentence consists of a short backbone plus a technical term that already exists in English, there is little to be gained from paraphrase. ChatGPT, for its part, maintains high readability as well because the smoothing it performs is mostly at the level of sentence rhythm and information packaging rather than at the level of replacing any technical term. Where variations appear in the data, most noticeably for Tempo's environmental news, one can usually locate a local feature that explains the difference: a cluster of local place names and programmatic references, a stretch of enumerated policy levers, or a sentence whose final position in Indonesian carries a weight that English renders more naturally at the start. In those cases, DeepL's sequence preservation protects the relationship between clause parts, whereas ChatGPT's reordering for flow can create small shifts of emphasis. The overall effect on readability is mild, but visible to raters who read for the alignment of emphasis.

A separate but equally instructive perspective concerns inter-rater patterns and the nature of the readability rubric. The study uses the three-point readability scale introduced by Nababan et al. (2012), which, precisely because of its simplicity, yields high near-agreement and moderate exact agreement. On short segments such as those in the present dataset, most items collect near the middle of the scale, and the differences that do exist are explainable by what each rater rewards. One rater places a premium on discourse-level cohesion and idiomatic rhythm, thereby consistently assigning higher readability to outputs that minimize calques and reduce paratactic accumulation; the other rater appears to give more credit to items that carry over every overt qualifier and structural cue from the source, even at the cost of stiffness. These dispositions are not contradictory. They represent legitimate emphases within a coarse rubric. What matters for interpretation is that the variation is principled, not noise, and that it maps to system tendencies in ways that practitioners can act on. This is consistent with translation quality assessment theory that highlights how different evaluative criteria, and textual functions can produce legitimate differences in judgement, even when raters apply a rubric consistently (House, 2015).

The dataset also lends itself to a practical reading of where each system is best suited with minimal post-editing. For law articles, especially those with many acronyms for Indonesian agencies, units, and ranks, or with layered legal bases, readability can be optimized by combining the systems' virtues. One tactically effective approach is to start from a DeepL draft to anchor terminology and sequence, then apply light editorial smoothing in the manner of ChatGPT or human post-editors to split long sentences, supply connective tissue, and front key actors. Another approach is to prompt ChatGPT with explicit constraints that discourage compression of legal qualifiers and require first-mention glosses for institutions (for example, "*Bawaslu* (Indonesia's Election Supervisory Board)"). Both approaches are visible in the data as "what worked" when readability improved without loss of legal content. This recommendation directly reflects post-editing scholarship describing MTPE as targeted revision to increase comprehensibility while maintaining adequacy—

especially important in terminology-heavy texts (O'Brien & Simard, 2014; Vieira, 2019). In environmental and general policy coverage, where the lexicon is internationally shared and the propositional content is less culture-bound, either system can serve as a dependable first pass; in those settings, the main post-editing need is stylistic tidying rather than factual or terminological repair.

What this shows is not that one tool is always better than the other, but that each has its own strengths and limitations depending on the domain. ChatGPT shows a clear advantage in readability, especially in cases where smoother phrasing helps readers follow the meaning more easily and where clause reordering makes the emphasis clearer in English. DeepL, on the other hand, performs more consistently in situations where preserving detail is especially important or where technical terms already match English well, making a more literal translation effective. Since journalistic translation requires both clear expression and accurate detail, the most useful conclusion is that the choice of system and post editing strategy should be adjusted to the specific domain rather than based on the assumption that one tool will perform best in every genre (Nababan et al., 2012; Saldanha & O'Brien, 2013). This interpretation is compatible with the scoring patterns in the spreadsheet: small mean differences across systems, clustered by topic and outlet, explainable by predictable stylistic features and by raters' articulated preferences.

CONCLUSION

Overall, this study shows that ChatGPT and DeepL produced Indonesian–English news translations that were broadly comparable in readability, with only small but consistent differences. ChatGPT was more often perceived as easier to read because it tended to smooth sentences and improve flow, whereas DeepL more often kept the original structure and details, which can help preserve precision but sometimes made the English feel stiff. The results also made it clear that topic mattered more than outlet: environmental texts were generally the easiest for both tools, while legal texts were the hardest because they contain dense institutional terms, acronyms, and culture-specific references. In terms of contribution, this study adds a readability-focused, domain-based comparison of two widely used AI tools in Indonesian–English journalistic translation, highlighting that readability depends on the fit between the tool's translation style and the genre's demands rather than on one tool being universally "better." Practically, the findings suggest that better outcomes come from using domain-aware workflows: legal and political texts need consistent handling of acronyms and institution names (including short first-mention explanations) and light post-editing to reduce sentence density, while environmental news usually needs less intervention and mainly benefits from small edits to maintain cohesion and integrate local context smoothly.

ACKNOWLEDGMENT

This work was supported by Politeknik Negeri Jakarta under Grant number 721/PL3/PT.00.06/2025

BIBLIOGRAPHY

AlAfnan, M. A. (2024). *Large language models as computational linguistics tools: A comparative analysis of ChatGPT and Google machine translations*. Journal of Artificial Intelligence and Technology, Advance online publication. doi:10.37965/jait.2024.0549

Basha, J. Y. (2024). *The negative impacts of AI tools on students in academic and real-life performance*. International Journal of Social Sciences and Commerce, 1(3), 1–16. doi:10.51470/IJSSC.2024.01.03.01

Castilho, S., Quinn Mallon, C., Meister, R., & Yue, S. (2023). Do online machine translation systems care for context? What about a GPT model?. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 393–417. European Association for Machine Translation.

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, 9(15), 120–173.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Thousand Oaks, CA: SAGE.

Fitriani, E., Miru, A. S., Lamada, M. S., & Fitrani, E. (2024). *Analisis tentang pemahaman, persepsi dan aspek ChatGPT oleh mahasiswa* [Analysis of students' understanding, perceptions, and aspects of ChatGPT use]. Jurnal Ilmiah Multidisipliner (JIMU), 1(1), 1–12.

Gadd, S. (2024). *Data and AI governance*. Journal of Financial Transformation, 59, 8–19.

Gao, R., Lin, Y., Zhao, N., & Cai, Z. G. (2024). *Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL*. Humanities and Social Sciences Communications, 11, Article 835. doi:10.1057/s41599-024-03363-0

Handoyo, E. R., Sugiarto, J., Lolo, A., & Chai, K. (2023). *Identifikasi pengaruh penggunaan ChatGPT terhadap kemampuan berpikir mahasiswa* [Identifying the influence of ChatGPT use on students' critical thinking]. KONSTELASI: Konvergensi Teknologi dan Sistem Informasi, 7(2), 45–53. doi:10.24002/konstelasi.v3i2.7241

Hendy, A., Abdurrahman, A., McCarthy, A., Tamchyna, A., & Bojar, O. (2023). *How good are GPT models at machine translation? A comprehensive evaluation*. doi:10.48550/arXiv.2302.09210

House, J. (2015). *Translation quality assessment: Past and present* (2nd ed.). Routledge.

Iqbal, M., Khan, N. U., & Imran, M. (2024). The role of artificial intelligence (AI) in transforming educational practices: Opportunities, challenges, and implications. *Qlantic Journal of Social Sciences and Humanities*, 5(2), 348–359. doi:10.55737/qjss.349319430

Kazu, I. Y., & Kuvvetli, M. (2023). The influence of AI-based pronunciation training on vocabulary acquisition in English learning. *International Journal of Psychology and Educational Studies*, 10(2), 480–493. doi:10.52380/ijpes.2023.10.2.1044

Kumar, O. P. (2023). Investigating the impact of artificial intelligence and technology in English language learning. *Advances in Social Behavior Research*, 3(1), 27–36. doi:10.54254/2753-7102/3/2023026

Lin, C. C., Huang, A. Y. Q., & Lu, O. H. T. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: A systematic review. *Smart Learning Environments*, 10, Article 41. doi:10.1186/s40561-023-00260-y

Misnawati, M. (2023). *ChatGPT:* Keuntungan, risiko, dan penggunaan bijak dalam era kecerdasan buatan [ChatGPT: Benefits, risks, and wise usage in the AI era]. *Jurnal Teknologi Pembelajaran Indonesia*, 13(2), 1–10. doi:10.55606/mateandrau.v2i1.221

Nababan, M. R., Nuraeni, A., & Sumardiono, B. (2012). Pengembangan model penilaian kualitas terjemahan [Developing a model for translation quality assessment]. *Kajian Linguistik dan Sastra*, 24(1), 39–57. doi:10.23917/kls.v24i1.101

O'Brien, S., & Simard, M. (2014). Introduction to special issue on post-editing. *Machine Translation, 28*(3–4), 159–164. https://doi.org/10.1007/s10590-014-9166-8

Omaggio Hadley, A. (2001). *Teaching language in context* (3rd ed.). Boston, MA: Heinle & Heinle.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching* (2nd ed.). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511667305

Rizvi, M. (2023). *Investigating AI-powered tutoring systems that adapt to individual student needs, providing personalized guidance and assessments*. Eurasia Proceedings of Educational & Social Sciences, 31, 67–73. doi:10.55549/epess.1381518

Rusandi, E., & Rusli, T. (2021). *Designing Basic/Descriptive Qualitative Research and Case Study*. Al-Ubudiyah: Jurnal Pendidikan dan Studi Islam, doi:10.55623/au.v2i1.18

Saldanha, G., & O'Brien, S. (2013). *Research methodologies in translation studies*. Manchester, UK: St. Jerome. doi:10.1080/1750399X.2015.1016282

Son, J., & Kim, B. (2023). Translation performance from the user's perspective of large language models and neural machine translation systems. *Information, 14*(10), 574. https://doi.org/10.3390/info14100574

Suharmawan, W. (2023). *Pemanfaatan ChatGPT dalam dunia pendidikan* [Utilization of ChatGPT in the world of education]. Education Journal: Journal Educational Research and Development, 7(2), 158–166. doi:10.31537/ej.v7i2.1248

Sun, R. (2024). Evaluating the translation accuracy of ChatGPT and DeepL through the lens of implied subjects. *Arab World English Journal for Translation & Literary Studies, 8*(4), 41–53. https://doi.org/10.24093/awejtls/vol8no4.5

Toury, G. (1995). *Descriptive translation studies and beyond*. John Benjamins.

Troqe, R., & Marchan, F. (2017). News Translation: Text analysis, fieldwork, survey. *In Empirical modelling of translation and interpreting* (pp. 277–310). Language Science Press. https://doi.org/10.5281/zenodo.1090974

Vieira, L. N. (2019). Post-editing of machine translation. *The Routledge handbook of translation and technology* (pp. 319–335). Routledge.