JOBIT

Journal of Digital Business and Information Technology

ChatGPT vs DeepL: Comparing the English Translation Quality of Digital Business and Information Technology Texts Using BLEU Metric

Haji Abdul Karim
Manajemen Pendidikan Islam, Pascasarjana, IAIN Palangka Raya
Jl. G. Obos Kompleks Islamic Centre, Palangka Raya, Kalimantan Tengah
Email: [1]hajiabdulkarim@iain-palangkaraya.ac.id

ABSTRACT

In the era of digital globalization, the demand for accurate and efficient translation tools has grown significantly. Advances in artificial intelligence (AI) have given rise to various automated translation platforms that facilitate cross-lingual communication. Among the most notable are ChatGPT and DeepL. This study focusses on translating digital business and information technology texts in English using both AI. All text is analyzed by determining their quality grammar and context-preserved, in addition, the accuracy assessment is being analyzed by using BLEU metric to ensure the translation quality threshold score. This study results that both have their own minuses and pluses. Both AI-translation tools have their strengths, with DeepL being slightly better for consistency in grammar rules and structure, while ChatGPT excels in clarity, accuracy, and adapting to the intended meaning. Meanwhile, the evaluation using the BLEU metric showed that ChatGPT scores ranged between 0.56-0.78, while DeepL scores ranged between 0.44-0.91.

Keywords : AI; ChatGPT; DeepL; Translation; BLEU Metric

ABSTRAK

Di era globalisasi digital, permintaan akan alat terjemahan yang akurat dan efisien telah berkembang pesat. Kemajuan dalam kecerdasan buatan (AI) telah melahirkan berbagai platform terjemahan otomatis yang memfasilitasi komunikasi lintas bahasa. Di antara yang paling terkenal adalah ChatGPT dan DeepL. Penelitian ini fokus pada penerjemahan teks bisnis digital dan teknologi informasi dalam bahasa Inggris menggunakan kedua AI tersebut. Semua teks dianalisis berdasarkan kualitas tata bahasanya dan sejauh mana konteksnya dipertahankan, serta penilaian akurasi dianalisis menggunakan metrik BLEU untuk mengetahui skor ambang batas kualitas terjemahan. Hasil penelitian ini menunjukkan bahwa keduanya memiliki kelebihan dan kekurangannya masing-masing. Kedua alat penerjemahan AI ini memiliki kekuatan tersendiri, dengan DeepL sedikit lebih unggul dalam konsistensi aturan tata bahasa dan struktur, sementara ChatGPT lebih unggul dalam kejelasan, akurasi, dan penyesuaian dengan makna yang dimaksud. Sedangkan, penilaian menggunakan metrik BLEU, hasil menunjukan ChatGPT berada pada ambang batas 0.56-0.78, sementara DeepL pada 0.44-0.91.

Kata kunci : AI; ChatGPT; DeepL; Penerjemahan; Metrik BLEU

Diterima Redaksi: 27-11-2024 | Selesai Revisi: 24-12-2024 | Diterbitkan Online: 06-01-2025

1. PENDAHULUAN

In the era of digital globalization, the demand for accurate and efficient translation tools has grown significantly. Global interaction is inherently tied to interlingual activities, making translation an essential aspect of globalization. However, many individuals are either unable or unwilling to overcome language barriers on their own and must rely on translations provided by others to access information beyond their linguistic capabilities. Traditionally, human translators and interpreters have fulfilled this role, serving as linguistic and cultural mediators to grant access the information. In addition, the perception of intercultural differences also requires understanding and studying the mentality of the people in the language of the original text. The ability to apply lexical, grammatical, and stylistic techniques on the spot in accordance with the norms of translation improves the quality of translation (Popel et al., 2020; Tursunovich, 2022).

Advances in artificial intelligence (AI) have given rise to various automated translation platforms that facilitate cross-lingual communication. Among the most notable are ChatGPT, an AI-powered model developed by OpenAI, known for its conversational and contextual understanding (Brown et al., 2020; Dalayli, 2023), and DeepL, a translation engine renowned for its linguistic precision, accuracy, fluency, and naturalness (Linlin, 2024).

Texts related to Digital Business and Information Technology pose unique challenges for translation. These challenges stem from the presence of technical terminology, complex sentence structures, and the need to preserve context and accuracy. For example, information technology texts often include jargon and abbreviations that require contextual understanding for accurate interpretation (Munasinghe et al., 2021). Errors in translating such texts can lead to misunderstandings, misguided decision-making, or even financial losses (He et al., 2021). Consequently, evaluating the capabilities of automated translation tools in handling these specialized texts is crucial.

While both ChatGPT and DeepL have been widely adopted, systematic comparisons of their translation quality,

DOI: 10.23971/jobit.v1i2.297

particularly within the domains of digital business and information technology, remain scarce. Studies such as Popović (2021) emphasize the importance of linguistic accuracy and domain-specific terminology in assessing translation quality, while more recent works e.g. (Soysal, 2023) highlight the potential of AI-driven translation tools to revolutionize industry practices. This study builds on these insights by analyzing the translation outputs of these two tools using linguistic and technical criteria.

Thus, this research seeks to answer the question: How does the translation quality of ChatGPT compare to DeepL when applied to texts in digital business and information technology? The findings of this study are expected to guide users in selecting the most suitable translation tool for their specific needs and contribute to the advancement of AI-based translation technology.

2. METODE

This study focuses on different texts in the Journal of Digital Business and Information Technology (JOBIT) Volume 1, Number 1, 2024. The aim of this study is to compare the quality of ChatGPT and DeepL in translating highly-related text with digital business and information technology. The comparison is presented in the results tables alongside the chosen texts. Although ChatGPT and DeepL offers a wide range of translation from other languages, the author focused on translating Indonesian to English due to significant number of research in those subjects. The method used in this study is by comparing both translation results from ChatGPT and DeepL whether the text can be understood well and comprehensible in linguistics realm.

Moreover, the writer will use BLEU metric to ensure the translation quality. BLEU (Bilingual Evaluation Understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "The closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU. BLEU was one of the first metrics to claim a high correlation with human judgements of quality, and remains one of the most popular evaluation metrics for machine translation. BLEU provides a

DOI: 10.23971/jobit.v1i2.297

quantitative measure of how similar the candidate translation is to the reference translation. BLEU's score is from number 0 to 1. This value indicates how similar the candidate text (AI translation) is to the reference texts (human translation), with values closer to 1 representing more similar accurate texts.

The BLEU score thresholds are explained in the following table:

Tabel 1. BLEU Score Thresholds

Thresholds	Indication
0.8 – 1.0	Excellent translations
0.6 – 0.8	Good translations
0.4 – 0.6	Adequate translations
Below 0.4	Poor Translations

(Sumber: (Papineni et al., 2002)

The writer will use BLEU metric that available online in https://huggingface.co/spaces/evaluate-metric/bleu to calculate the translation quality.

3. HASIL DAN PEMBAHASAN

The first text is related to the usage of pothole detection technology called YOLO v5.

Table 2. First Text

Original Text
Penelitian ini menggunakan metode YOLO v5.YOLO v5 dipilih karena algoritma ini telah terbukti memiliki kecepatan dan akurasi tinggi dalam mengenali objek pada gambar. Bahasa pemrograman yang digunakan adalah Python, Tensorflow digunakan sebagai library *untuk membangun model deteksi objek dan Google Colaboratory sebagai* notebook *untuk menjalankan program.*
Human Translation
This research uses the YOLO v5 method. YOLO v5 was chosen because this algorithm has proven to have high speed and accuracy in recognizing objects in images. The programming language used is Python, Tensorflow is used as a library to build an object detection model and Google Colaboratory is used as the notebook to run the program.
ChatGPT Translation
This research uses the YOLO v5 method. YOLO v5 was chosen because this algorithm has been proven to have high speed and accuracy in recognizing objects in images. The programming language used is Python, TensorFlow is utilized as the library to build the object detection model, and Google Colaboratory is used as the notebook to run the program.
DeepL Translation
This research uses the YOLO v5 method. YOLO v5 was chosen because this algorithm has proven to have high speed and accuracy in recognizing objects in images. The programming language used is Python, Tensorflow is used as a library to build an object detection model and Google Colaboratory as a notebook to run the program.

(Source: Surahmanto et al., 2024)

The excerpt above indicates a subtle difference. The DeepL translation uses "has proven", which is more concise and precise compared to "has been proven" in the ChatGPT translation. The expression "has proven" is more direct and common in technical language. The DeepL translation states "a library" and "an object detection model", which aligns with English grammar rules to indicate first-time usage. The DeepL translation structures "and Google Colaboratory as a

DOI: 10.23971/jobit.v1i2.297

notebook" more in line with the previous phrase structure, making it more cohesive. Furthermore, the BLEU score for ChatGPT is 0.77 and DeepL is 0.91. It means that both qualitative and quantitative assessment showed the DeepL translation is more accurate, it indicates as an excellent translation.

The second text is about the influence of the duration of gadget use and the duration of learning for informatics students on logical ability and mathematics learning outcomes.

Table 3. Second Text

Original Text
Uji heteroskedastisitas ini dilakukan untuk menguji apakah terdapat perbedaan varians pada model trimming *dari satu pengamat ke pengamat lainnya. Model yang variansnya homoskedastisitas merupakan model yang baik. Pengujian dengan Breusch Pagan Godfrey menggunakan Python (Gambar 12) menghasilkan nilai probabilitas (p-value) pada model* trimming *sebesar 0,127865*
Human Translation
This heteroscedasticity test is conducted to test whether there is a difference in variance in the trimming model from one observer to another. A model with homoscedastic variance is considered a good model. Testing with Breusch-Pagan-Godfrey using Python (Figure 12) results in a probability value (p-value) on the trimming model of 0.127865.
ChatGPT Translation
This heteroscedasticity test is conducted to determine whether there is a difference in variance in the trimming model from one observation to another. A model with homoscedastic variance is considered a good model. Testing with the Breusch-Pagan-Godfrey method using Python (Figure 12) resulted in a probability value (p-value) for the trimming model of 0.127865.
DeepL Translation
This heteroscedasticity test is conducted to test whether there are differences in variance in the trimming model from one observer to another. A model whose variance is homoscedasticity is a good model. Testing with Breusch Pagan Godfrey using Python (Figure 12) results in a probability value (p-value) on the trimming model of 0.127865.

(Source: Himmah & Kaestria, 2024)

The ChatGPT translation uses "difference in variance", which is more accurate than "differences in variance", as the context refers to the changing variance pattern from one observer to another. The use of the term "homoscedastic variance" in the first sentence is more technical and precise, while the phrase "whose variance is homoscedasticity" in DeepL translation is grammatically incorrect. ChatGPT translation uses "resulted", which is more consistent with the tense used, compared to "results" in DeepL. ChatGPT translation preserves the method's name as "Breusch-Pagan-Godfrey", adhering to the standard format commonly used in academic literature, whereas the DeepL omits the hyphen without justification. Then, one thing that very obvious, *pengamat* is translated to "observation" and "observer" by ChatGPT and DeepL respectively which the DeepL one is more appropriate. Meanwhile, the BLEU score for ChatGPT is 0.75 and DeepL is 0.77 which slightly higher. Even though,

DOI: 10.23971/jobit.v1i2.297

both AI-translation is in the threshold of good translation.

The third text is about decision support system using the Fuzzy TOPSIS Method.

Table 4. Third Text

Original Text
Berdasarkan penelitian ini, dapat diambil beberapa kesimpulan sebagai berikut: Penilaian menggunakan metode Fuzzy TOPSIS, di mana Fuzzy digunakan untuk pembobotan kriteria dan metode TOPSIS digunakan untuk perangkingan hasil penilaian.
Human Translation
Based on this research, conclusions can be drawn as follows: The assessment uses the Fuzzy TOPSIS method, where Fuzzy is used for weighting the criteria and the TOPSIS method is used for ranking the assessment results.
ChatGPT Translation
Based on this research, several conclusions can be drawn as follows: The assessment uses the Fuzzy TOPSIS method, where Fuzzy is applied for criteria weighting, and the TOPSIS method is used for ranking the assessment results.
DeepL Translation
Based on this research, the following conclusions can be drawn: The assessment uses the Fuzzy TOPSIS method, where Fuzzy is used for weighting the criteria and the TOPSIS method is used for ranking the assessment results.

(Source: Tamam & Rofiuddin, 2024)

The DeepL translation is better than ChatGPT because it is clearer and follows a more natural grammatical structure. Here are the reasons, the second sentence uses "the following conclusions can be drawn", which is more direct and formal compared to "several conclusions can be drawn as follows" in the first sentence. The phrase "Fuzzy is used for weighting the criteria" in DeepL is more straightforward and easier to understand than "Fuzzy is applied for criteria weighting" in ChatGPT. The DeepL translation maintains greater consistency in structure and the order of ideas, making it more organized. The BLEU score for ChatGPT is 0.78 and DeepL is 0.86. This means that, qualitatively and quantitatively, the DeepL is better.

The fourth text is related to the application of Matplotlib in data visualization for analysis of the relationship between gadget usage and learning outcomes.

Table 5. Fourth Text

Original Text
Penelitian ini menyimpulkan bahwa visualisasi data dapat dilakukan dengan menggunakan bahasa pemrograman Python kemudian menerapkan library *Matplotlib untuk mengubah data mentah menjadi informasi yang bisa dibaca melalui visualisasi data dalam berbagai bentuk grafik dan diagram. Harus dilakukan penelitian lanjutan untuk menemukan variabel lain yang juga mempengaruhi hasil belajar atau menggali lebih dalam terkait dengan durasi* gadget *digunakan untuk mengakses aplikasi apa saja sehingga dengan visualisasi yang tepat maka dapat lebih memahami dampak teknologi terhadap prestasi akademik mahasiswa.*
Human Translation
This study concludes that data visualization can be carried out using the Python programming language by applying the Matplotlib library to transform raw data into readable information through various forms

DOI: 10.23971/jobit.v1i2.297

of graphs and diagrams. Further research should be carried out to find other variables that also affect learning outcomes or to delve deeper into the duration of the gadget usage in accessing any application, so that with appropriate visualization, the impact of technology on students' academic performance can be better understood.
ChatGPT Translation
This study concludes that data visualization can be carried out using the Python programming language by applying the Matplotlib library to transform raw data into readable information through various forms of charts and graphs. Further research is needed to identify other variables that also affect learning outcomes or to delve deeper into the duration of gadget usage and the specific applications accessed. With appropriate visualization, the impact of technology on students' academic performance can be better understood.
DeepL Translation
This study concluded that data visualization can be done using the python programming language and then applying the matploplib library to convert raw data into information that can be read through data visualization in various forms of graphs and diagrams. Further research should be carried out to find other variables that also affect learning outcomes or explore more deeply related to the duration of the gadget used to access any application so that with the right visualization, it can better understand the impact of technology on student academic achievement.

(Source: Kaestria et al., 2024)

The ChatGPT translation uses "This study concludes", which is more appropriate for presenting research findings that are still valid or relevant. The DeepL translation uses "This study concluded", which tends to refer to finalized results, although this might be acceptable in certain contexts. The phrase "can be carried out" in the first sentence is more accurate than "can be done" in the second sentence, as it is more formal and technical.

The use of "Matplotlib" in the first sentence is consistent with the correct spelling, whereas "matploplib" in the second sentence is a typo. The first sentence has a better structure with more organized explanations, while the DeepL translation feels longer and somewhat convoluted (for instance, "so that with the right visualization, it can better understand" could be clearer if written more concisely). The BLEU score for ChatGPT is higher by 0.72 than the DeepL which is 0.44. It means that, in linguistics realm and BLEU metric, ChatGPT has similar assessment of better translation than the DeepL.

The last text is about the development of web-based expert system.

Table 6. Fifth Text

Original Text
Penelitian ini menggunakan metode forward chaining *dalam penyusunan aturan dan menambahkan metode kepastian faktor (CF) untuk menghitung nilai kepastian sehingga keputusan yang dihasilkan lebih akurat. Dari hasil perhitungan manual menggunakan sampel satu penyakit dengan kode P1, uji coba menunjukkan nilai akurasi sebesar 93,0736%.*
Human Translation
This research uses the forward chaining method for rule formulation and incorporates the certainty factor (CF) method to calculate the certainty values so that the resulting decision is more accurate. Based on manual

DOI: 10.23971/jobit.v1i2.297

calculations using a sample of one disease with the code P1, the trial showed an accuracy value of 93.0736%.
ChatGPT Translation
This study employs the forward chaining method for rule formulation and incorporates the certainty factor (CF) method to calculate certainty values, ensuring more accurate decisions. Based on manual calculations using a sample disease with the code P1, testing demonstrated an accuracy rate of 93.0736%.
DeepL Translation
This research uses the forward chaining method in the preparation of rules and adds the certainty factor (CF) method to calculate the certainty value so that the resulting decision is more accurate. From the results of manual calculations using a sample of one disease with code P1, the trial showed an accuracy value of 93.0736%.

(Source: Saad & Pratiwi, 2024)

The first sentence uses "This study employs", which is more formal and consistent with academic style, compared to "This research uses" in the second sentence, which sounds slightly more general. The first sentence uses "incorporates", which is more appropriate in a technical context, while "adds" in the second sentence feels less specific in describing the application of a method.

The ChatGPT translation uses "Based on manual calculations using a sample disease with the code P1", which is clearer and more structured than "From the results of manual calculations using a sample of one disease with code P1" in DeepL translation. The first phrase is more direct and easier to understand. "Accuracy rate" in ChatGPT translation is more accurate for describing the value derived from testing, while "accuracy value" in the DeepL translation is somewhat less precise in this context. Despite, the BLEU score for ChatGPT is 0.56 whereas DeepL is 0.70.

From those findings, in terms of linguistics realm, it can be discussed that DeepL is more consistent in verb usage, tense, article usage, and parallel structure. It excels in maintaining clarity, parallelism, and consistency in verb usage. Meanwhile, ChatGPT provides clearer, more accurate contextual adjustments. It has better verb agreement, more precise technical terminology, and ensures that the tone is appropriate for an academic style. Additionally, it maintains better coherence, clarity, and structure. However, in terms of quantitative assessment by BLEU metric, both ChatGPT and DeepL showed their own excellency. Thus, the evaluation using the BLEU metric showed that ChatGPT scores ranged between 0.56-0.78, while DeepL scores ranged between 0.44-0.91.

DOI: 10.23971/jobit.v1i2.297

4. KESIMPULAN

It can be concluded that there is no real winner in the debate over which is better and more high-quality in translating digital business and information technology texts to English between ChatGPT and DeepL. Both have their own minuses and pluses. Both AI-translation tools have their strengths, with DeepL being slightly better for consistency in grammar rules and structure, while ChatGPT excels in clarity, accuracy, and adapting to the intended meaning. Those AI-tools may not be suitable for translating highly academical, technical or specialised text especially journal article that requires domain-specific knowledge. In addition, BLEU metric is a useful and fast method to measure the quality of machine translations, but it should often be combined with other metrics or human evaluations for a more comprehensive assessment. Ultimately, the role of human translators cannot be overlooked. AI is merely a tool to facilitate the translation process, depending on how wisely we use it will determine its effectiveness.

5. UCAPAN TERIMA KASIH

A string of thanks is conveyed to Journal of Digital Business and Information Technology (JOBIT) for granting permission to the author to research their articles. It is hoped that the results of this study can serve as a consideration for JOBIT when translating their articles in the future, with a focus on internationalization and the goal of being reputable-indexed journals.

DAFTAR PUSTAKA

Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. *ArXivLabs*, *1*(1), 1–63. https://doi.org/10.48550/arXiv.2005.14165

Dalayli, F. (2023). Use of NLP Techniques n Translation by ChatGPT: Case Study. *Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC)*, 19–25. https://doi.org/10.26615/978-954-452-090-8_003

He, P., Meister, C., & Su, Z. (2021). Testing Machine Translation via Referential Transparency. *Proceeding of 43rd International Conference on Software Engineering (ICSE)*, 410–422.

DOI: 10.23971/jobit.v1i2.297

https://doi.org/10.1109/ICSE43902.2021.00047

Himmah, E. F., & Kaestria, R. (2024). Pengaruh Durasi Penggunaan Gadget dan Durasi Belajar Mahasiswa Informatika terhadap Kemampuan Logika dan Hasil Belajar Matematika. *Journal of Digital Business and Information Technology*, *1*(1), 9–21. https://doi.org/10.23971/jobit.v1i1.207

Kaestria, R., Himmah, E. F., & Irawan, R. (2024). Penerapan Matplotlib dalam Visualisasi Data untuk Analisis Hubungan Penggunaan Gadget dan Hasil Belajar. *Journal of Digital Business and Information Technology*, *1*(1), 29–39. https://doi.org/10.23971/jobit.v1i1.204

Linlin, L. (2024). Artificial Intelligence Translator DeepL Translation Quality Control. *Procedia Computer Science*, *247*(1), 710–717. https://doi.org/10.1016/j.procs.2024.10.086

Munasinghe, B., Bell, T., & Robins, A. (2021). Teachers' Understanding of Technical Terms in a Computational Thinking Curriculum. *Proceedings of the 23rd Australasian Computing Education Conference*, 106–114. https://doi.org/10.1145/3441636.3442311

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., & Žabokrtský, Z. (2020). Transforming Machine Translation: A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals. *Nature Communications*, *11*(1), 4381–4395. https://doi.org/10.1038/s41467-020-18073-9

Popović, M. (2021). Agree to Disagree: Analysis of Inter-Annotator Disagreements in Human Evaluation of Machine Translation Output. *25th Conference on Computational Natural Language Learning (CoNLL)*, 234–243. https://github.com/m-popovic/

Saad, M. I., & Pratiwi, H. (2024). Development of a Web-Based Expert System for Diagnosing Cocoa Diseases Using Forward Chaining and Certainty Factor Methods. *Journal of Digital Business and Information Technology*, *1*(1), 40–49. https://doi.org/10.23971/jobit.v1i1.223

Soysal, F. (2023). Enhancing Translation Studies with Artificial Intelligence (AI): Challenges, Opportunities, and Proposals. *International Journal of Philology and*

DOI: 10.23971/jobit.v1i2.297

Translation Studies, *5*(2), 177–191. https://doi.org/10.55036/ufced.1402649

Surahmanto, M., Aras, S., Rifki, I. A. M., & Ussalama, P. (2024). Deteksi Jalan Berlubang Menggunakan Algoritma Yolov5. *Journal of Digital Business and Information Technology*, *1*(1), 1–8. https://doi.org/10.23971/jobit.v1i1.198

Tamam, M. B., & Rofiuddin. (2024). Sistem Pendukung Keputusan Penentuan Lokasi Budidaya Ikan Tawar yang Cocok Dikabupaten Pamekasan dengan Menggunakan Metode Fuzzy Topsis. *Journal of Digital Business and Information Technology*, *1*(1), 22–28. https://doi.org/10.23971/jobit.v1i1.203

Tursunovich, R. I. (2022). Linguistic and Cultural Aspects of Literary Translation and Translation Skills. *British Journal of Global Ecology and Sustainable Development*, *10*(1), 168–173.

DOI: 10.23971/jobit.v1i2.297