Yagahara *et al.*
BMC Medical Informatics and Decision Making (2025) 25:66
https://doi.org/10.1186/s12911-025-02912-0

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Exploration of the optimal deep learning model for english-Japanese machine translation of medical device adverse event terminology

Ayako Yagahara[1*], Masahito Uesugi[2] and Hideto Yokoi[3]

Abstract

Background In Japan, reporting of medical device malfunctions and related health problems is mandatory, and efforts are being made to standardize terminology through the Adverse Event Terminology Collection of the Japan Federation of Medical Device Associations (JFMDA). Internationally, the Adverse Event Terminology of the International Medical Device Regulators Forum (IMDRF-AET) provides a standardized terminology collection in English. Mapping between the JFMDA terminology collection and the IMDRF-AET is critical to international harmonization. However, the process of translating the terminology collections from English to Japanese and reconciling them is done manually, resulting in high human workloads and potential inaccuracies.

Objective The purpose of this study is to investigate the optimal machine translation model for the IMDRF-AET into Japanese for the part of a function for the automatic terminology mapping system.

Methods English-Japanese parallel data for IMDRF-AET published by the Ministry of Health, Labor and Welfare in Japan was obtained from 50 sentences randomly extracted from the terms and their definitions. These English sentences were fed into the following machine translation models to produce Japanese translations: mBART50, m2m-100, Google Translation, Multilingual T5, GPT-3, ChatGPT, and GPT-4. The evaluations included the quantitative metrics of BiLingual Evaluation Understudy (BLEU), Character Error Rate (CER), Word Error Rate (WER), Metric for Evaluation of Translation with Explicit ORdering (METEOR), and Bidirectional Encoder Representations from Transformers (BERT) score, as well as qualitative evaluations by four experts.

Results GPT-4 outperformed other models in both the quantitative and qualitative evaluations, with ChatGPT showing the same capability, but with lower quantitative scores, in the qualitative evaluation. Scores of other models, including mBART50 and m2m-100, lagged behind, particularly in the CER and BERT scores.

Conclusion GPT-4's superior performance in translating medical terminology, indicates its potential utility in improving the efficiency of the terminology mapping system.

Keywords Machine translation, Deep learning, Medical device adverse event terminology

*Correspondence:
Ayako Yagahara
yagahara-a@hus.ac.jp
Full list of author information is available at the end of the article

Introduction

In Japan, it is required to report any malfunctions occurring during the operation of medical devices, the associated health problems, and information related to the investigation of causes, as Medical Device Malfunction Reports to the government. To promote the use of standard terminology in these reports, the Japan Federation of Medical Device Associations (JFMDA) has published an Adverse Event Terminology Collection (JFMDA Terminology Collection) [1]. At present, the sixth edition has been published, comprising a collection of 97 individual terminology sets and one common terminology set. The individual terminology sets are associated with groups of Japanese medical device nomenclatures, such as medical X-ray devices, catheters, and cardiac pacemakers. Each terminology set includes approximately 100 terms, with a total of around 9,000 terms. The common terminology set is an organized version of the Adverse Event Terminology released by the International Medical Device Regulators Forum (IMDRF-AET) in a format suitable for use in Japan, while the IMDRF-AET is utilized for the collection of adverse event information abroad [2]. For international harmonization of the JFMDA Terminology Collection, manual mapping work is conducted by the JFMDA Malfunction Terminology Working Group. The JFMDA terminology collection is organized in a two-level structure, with upper-level terms indicating general categories and lower-level terms representing the specific terms recorded in reports. On the other hand, the IMDRF-AET follows a three-level structure, with all terms being eligible for inclusion in reports. The mapping process involves a one-to-one mapping between the specific terms in items in the JFMDA terminology collection and the corresponding terms in the IMDRF-AET. This work involves manually translating the IMDRF-AET into Japanese and visually comparing the IMDRF-AET translation data with the JFMDA Terminology Collection. As both terminology collections are updated at least once a year, each update requires significant human resources and time due to the thousands of terms involved, leading to potential mapping errors and inconsistencies between the terminology collections. To improve on this, we have been working on the development of a computer-aided system for mapping between terminologies with a machine translation and sentence similarity evaluation tool [3–5]. In this paper, we focus on machine translation.

Research on machine translation targeting medicine, according to a review by Dew et al., is primarily aimed at Health Education and Clinical Communication, with no studies identified on the translation of documents related to medical devices [6]. Noll et al. reviewed research cases on machine translation targeting medical terms such as SNOMED, investigating the target glossaries, languages, machine translation tools, and evaluation metrics [7]. Research specifically focusing on Japanese is very limited, forming only 3% of the studies.

We have been working on the development of deep learning-based English-Japanese translation models specifically for the IMDRF-AET [4, 5]. These studies have shown that the translation accuracy of IMDRF-AET using Generative Pretrained Transformer 3 (GPT-3) was the best, but since its publication, deep learning technology in natural language processing has advanced. The evolution of ChatGPT and GPT-4, surpasses traditional methods in various tasks. For instance, in examinations for medical licenses and specialist examinations, GPT-4 has achieved or nearly achieved passing marks [8–12]. Regarding clinical applications, there are reports that explanations to patients by ChatGPT are more understandable than those from doctors [13], and it has been applied to various tasks such as generating and simplifying radiology reports [14, 15]. These outcomes suggest that GPT-4 could be useful for the translation task of IMDRF, and given the enhancements made to GPT-4 for languages other than English [16], it is also expected to perform well in translations into Japanese.

The purpose of this study is to identify an optimal machine translation model for IMDRF-AET translation, incorporating ChatGPT and GPT-4.

Methods

Data collection and trained translation model acquisition

The IMDRF-AET comprises a single set of terminologies applicable to all medical devices, with sections ranging from Annex A to G, covering terms related to medical device problems, cause investigation, health effects, and medical device components. The IMDRF-AET is structured into three levels, with each term assigned a definition. Reports may utilize terms from all levels.

For this research, bilingual data from the IMDRF-AET published by the Ministry of Health, Labour and Welfare [17] was acquired, and 50 sentences were selected by generating pseudorandom values from the terms and definition texts in annexes A, E, F, and G. The investigation terms and these definitions in annex B, C, and D were excluded as they utilize the translated version from the IMDRF-AET and do not require additional mapping.

For the acquisition of pretrained translation models, this study obtained the models used in the previous study: the google translation [18] and multilingual-T5 (mT5) [19] released by Google, multilingual bidirectional auto-regressive transformer (mBART) [20] and Many-to-Many multilingual translation model (m2m-100) [21], released by Facebook AI Research (now Meta AI Research), and GPT-3 [22] released by Open AI. For m2m-100, both the 418 million parameter model

(m2m-100–418 M) and the 1.2 billion parameter model (m2m-100-1.2B) were utilized. In addition to these other models, this study utilized ChatGPT, and GPT-4 [23], which are provided by OpenAI.

Translations using Google Translation and mT5 were conducted by having the publicly available models from an original Python program perform machine translation. For GPT-3, ChatGPT, and GPT-4, machine translation was executed by entering "Translate the following sentence into Japanese" into the prompt on the webpage provided by OpenAI. These tasks were carried out in July 2023. For machine translations using mBART and m2m-100, subword tokenization was performed as a preprocessing step for the English input through byte pair encoding (BPE). Subsequently, the subword-tokenized English texts were input into each model for translation into Japanese. The software utilized for translations with mBART and m2m-100 was fairseq [24].

Evaluation of machine translation

English sentences in test data were input into all the models to generate Japanese translated sentences. A random selection of 50 sentences was extracted from the test data for both quantitative and qualitative evaluations. For the quantitative evaluation, the Bilingual Evaluation Understudy (BLEU) [25], character error rate (CER), word error rate (WER), Metric for Evaluation of Translation with Explicit ORdering (METEOR) [26], and Bidirectional Encoder Representations from the Transformers (BERT) score [27] were used.

The BLEU metric [25] is widely utilized in assessing the accuracy of machine translations. It is an evaluation metric based on the n-gram match rate between the machine-generated text (generated sentence) and the baseline of the Japanese translation (reference sentence) found in the translated version of the IMDRF terminology. The BLEU score is calculated using the following formula:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log P(n)\right)$$

where, P(n) represents the n-gram (ranging from unigram to 4-gram) match rate between the generated sentences in the test data and the reference sentences. The brevity penalty (BP) is a factor that applies a penalty when the generated sentence is shorter than the reference sentence. The purpose of BP is to discourage overly short translations that could artificially inflate the match rate by being brief but not necessarily accurate or complete. The higher the BLEU score, the closer the generated sentence is to the reference sentence.

The CER indicates the percentage of characters that were incorrectly predicted. The lower the value, the better the performance of the translation system with a CER of 0 being a perfect score, CER can then be computed as:

$$CER = \frac{S_c + D_c + I_c}{N_c}$$

where S_c is the number of character substitutions, D_c is the number of character deletions, I_c is the number of character insertions, and N_c is the number of characters in the reference sentence.

Word error rate represents the percentage of words that were incorrectly predicted.

$$WER = \frac{S_w + D_w + I_w}{N_w}$$

where S_w is the number of word substitutions, D_w is the number of word deletions, I_w is the number of word insertions, and N_w is the number of words in the reference sentence. The lower the value of WER, the better the translation accuracy.

The METEOR [26] metric measures the quality of the generated text based on the alignment between the generated text and the reference text. The metric is based on the harmonic mean of the unigram precision and recall, with recall weighted higher than precision. The weighting is 1 for recall and 9 for precision, according to the literature, and the formula for the harmonic mean is as follows

$$\text{Harmonic mean} = \frac{10 \times \text{Precision} \times \text{Recall}}{\text{Recall} + 9 \times \text{Precision}}$$

A penalty factor is applied for lack of cohesion in the word order between the translation and the reference. The penalty is calculated based on the number and size of chunks (contiguous sequences of matched words) in the translation, with more and longer chunks indicating better order. The penalty formula is:

$$\text{Penalty} = 0.5 \times \left(\frac{\text{number of chunks}}{\text{number of unigrams matched}}\right)^3$$

The final METEOR score is computed by applying the penalty to the harmonic mean, as follows:

$$\text{METEOR} = \text{Harmonic mean} \times (1 - \text{Penalty})$$

The BERT score [27] is a metric designed to evaluate the quality of text generated by machine learning models. It utilizes the BERT model, a deep learning algorithm developed by Google. Unlike traditional evaluation metrics that frequently rely on surface-level text comparisons, the BERT score calculates the semantic similarity between the generated text and a reference text.

This process involves embedding both the generated and reference texts into high-dimensional vector spaces using BERT, followed by computing the cosine similarity between these vectors. This methodology allows for the assessment of textual similarity and quality based on contextual meanings, providing a more nuanced evaluation compared to conventional metrics.

For the qualitative evaluation, four evaluators conducted a visual assessment. One evaluator was a physician with 20 years of experience in medical device regulatory affairs, including approval reviews and safety corrective actions. Two other evaluators were natural language researchers, one with 5 years and the other with 15 years of experience as radiological technologists. The fourth evaluator had at least 10 years of experience in medical device regulatory affairs and an additional 10 years of experience serving on post-marketing safety committees of industry associations.

This assessment focused on the semantic consistency of the generated sentences. The evaluators determined whether the meanings of the translations were consistent with the original English texts, considering factors such as context, accuracy, and completeness. The percentage of generated sentences that received approval from at least three evaluators was deemed to have achieved semantic coherence. Inter-rater agreement was evaluated by calculating the κ values for each pair of evaluators, followed by computing the average κ value across all pairs, as described in [28]. κ value of 0.41–0.60 was considered to indicate moderate agreement, 0.61–0.80 was considered to indicate good agreement, and 0.81–1.00 was considered to indicate excellent agreement.

Results

For inter-rater agreement, the κ values for the six pairs were calculated, ranging from 0.44 to 0.59, with the asymptotic test yielding $p < 0.001$. The average κ value was 0.51, indicating a moderate agreement.

The scores of each model are presented in Table 1. The best performance in both the quantitative and qualitative evaluations was achieved by GPT-4. ChatGPT showed a capability comparable to GPT-4 in qualitative evaluation, but it did not reach the quantitative scores of GPT-4. For the other models, mBART50 achieved a CER second only to GPT-4, but its performance in the other quantitative evaluations and qualitative evaluation was poorer. The two models of m2m-100 did not achieve good values in either of the quantitative and qualitative evaluations, with these models ranking either last or second to last in the CER and BERT scores. Google Translation ranked second after GPT-4 in the BLEU and BERT scores, and its results in the qualitative evaluation followed GPT-4 and ChatGPT in the ranking. The mT5 had the third-best results in BLEU and WER but ranked lowest in the visual evaluation. The GPT-3 ranked fourth in the qualitative evaluation, but its BLEU score was also low.

Translation examples are shown in Tables 2, 3 and 4. In Table 2, the GPT-4 produced correct Japanese translations compared with the reference texts, but other models output incorrect words and transliterations not used in clinical practice. The example in Table 3 shows that the objective was to translate "regionally-limited" into the Japanese term which refers to treatment being confined to a lesion or its surrounding area, but the context was not adequately captured, leading to the translation as a geographical area. This phenomenon was observed across almost all models. In models excluding the three types of GPT included here, outputs included mistranslations, untranslated terms, and transliterations not used in clinical practice among the medical terms. In the case shown in Table 4, the use of prepositional phrases and the order of words were incorrect, leading to an erroneous output of causality, despite the words being almost equivalent to the reference text.

Table 1 Score of each machine translation models

Model	BLEU	CER	WER	METEOR	BERT score	Evaluators
mBART50	15.29	0.541	0.719	0.481	0.855	42%
m2m-100–418 M	17.29	0.639	0.731	0.416	0.825	34%
m2m-100-1.2B	21.75	0.679	0.673	0.460	0.831	42%
googletranslation	27.72	0.543	0.581	0.574	0.881	56%
mT5	25.91	0.573	0.601	0.440	0.837	34%
GPT-3	20.85	0.578	0.610	0.452	0.861	54%
ChatGPT	24.69	0.569	0.876	0.571	0.877	**72%**
GPT-4	**35.24**	**0.424**	**0.496**	**0.612**	**0.892**	**72%**

Table 2 The Japanese translation of "angioedema." Inside the parentheses is the English translation of the Japanese generated text

Model	Generated sentences	
mBART50	アンジオデマ	Transliteration not used in clinical practice
m2m-100-418M	アンジオデマ	Transliteration not used in clinical practice
m2m-100-1.2B	アンジオデマ	Transliteration not used in clinical practice
googletrans	アンギオエマ	Transliteration not used in clinical practice
mT5	血管腫	Mistranslation(hemangioma)
GPT-3	アンギオエデマ	Transliteration not used in clinical practice
ChatGPT	アンギオエデマ	Transliteration not used in clinical practice
GPT-4	血管性浮腫	Correct translation

Discussion

In both the quantitative and qualitative evaluations, GPT-4 achieved the highest scores, establishing itself as the optimal model for this task. The superiority of GPT-4 in the quantitative evaluations can be attributed to its precise translation of health problem terms and events derived from medical device malfunctions, such as "angioedema" and "erythema," into Japanese, as demonstrated in the examples from Tables 2 and 3. The tendency of GPT-4 to produce translations closely matching the terms used in the reference texts was a contributing factor. The quantitative evaluation metrics, BLEU, CER, WER, and METEOR, used in this study assess the string similarity to reference texts, which led to lower quantitative scores for ChatGPT when it produced phonetic transcriptions that differed from the reference terms, despite being equivalent in the qualitative evaluation to GPT-4. The reason why the ChatGPT BLEU score was approximately 10 points lower than that of GPT-4, but equivalent in the qualitative evaluation, is believed to be due to the acceptance of the ChatGPT phonetic transcriptions as valid translations.

While GPT-4 was identified as the best model, it exhibited two main issues. The first issue, as shown in Table 3, is the misinterpretation of "regionally-limited" to words with geographical meanings. This may be attributed to the predominance of "regional" being associated with geographical contexts in the corpus from which it learned. A potential countermeasure for GPT-4 involves prompt tuning. Here, by specifying in the prompt that "regionally-limited is not geographical," the correct output was generated. Therefore, if translation errors are known to occur with specific terms in advance, prompt tuning could be a useful approach.

The second issue concerns the degradation of translation accuracy due to the breakdown of causal relationships in longer sentences. In the example from Table 3, the reference text stated "extreme thirst accompanied by chronic excessive intake of water," but GPT-4 translated it as "extreme thirst in the throat due to chronic excessive water intake.," introducing an error in the causal relationship. Generally, it is said that language models learn only the patterns of word occurrences from their training data, which can lead to the generation of unfounded sentences or hallucinations. This case is considered to be a result of one such hallucination. One way around this would be the generation of multiple outputs. When prompted to regenerate the output, the correct translation "extreme thirst accompanied by chronic excessive water intake" was produced. Creating multiple generated texts and manually selecting the most accurate one could lead to the acquisition of higher-precision translations.

There are limitations to relying solely on GPT-4 for the translation of mission-critical medical documents. To mitigate such risks, it is recommended to incorporate human oversight and a proofreading process. While GPT-4 has demonstrated nearly 90% accuracy in qualitative evaluations and recorded the highest scores in quantitative assessments, its utility is undeniable. However, it is essential to use GPT-4 with an understanding of its output limitations.

Regarding the BERT score, it was observed that all models achieved favorable outcomes, with scores exceeding 0.8. This metric evaluates similarity by transforming sentences into vector representations, suggesting that the nuances of the sentences generated by all models were likely close to those of the reference texts, indicating a high probability that the translations did capture the intent of the original English sentences. However, it is important to note that this score is not fully semantic and may have difficulty accounting for changes in temporality

Table 3 The Japanese translation of "A regionally-limited response to an antigen, which may include inflammation, induration, erythema, pruritus or pain." Inside the parentheses is the English translation of the Japanese generated text

Model	Generated sentences	Explanation of underlined words
mBART50	炎症，耐久性，エリテマ，pruritus または痛みを含む抗原に対する地域的に制限された反応. (A locally restricted response to antigens including inflammation, durability, *erythema*, pruritus, or pain)	mistranslation, untranslated word and *transliterations* not used in clinical practice
m2m-100-418M	抗原に対する地域的に制限された反応で，炎症，刺激，エリテマ，プルチウス，または痛みが含まれる可能性があります. (A locally restricted response to the antigen, which may include inflammation, irritation, *erythema*, *pruritus*, or pain)	mistranslations, and *transliterations* not used in clinical practice
m2m-100-1.2B	抗原に対する地域的に限られた反応，炎症，硬化，赤毛，かゆみ，または痛みを含む可能性があります. (A locally limited response to an antigen may include inflammation, induration, red hair, itching, or pain.)	mistranslation
GPT-3	地域的な抗原への反応であり，炎症，硬化，紅斑，かゆみ，痛みを含む場合がある. (It is a localized response to an antigen, which may include inflammation, induration, erythema, itching, and pain.)	mistranslation
Googletrans	炎症，硬膜，紅斑，掻痒，疼痛を含むことができる抗原に対する根拠に限られた反応. (A response limited to a rationale for an antigen, which is able to include inflammation, dura mater, erythema, itching, and pain.)	mistranslation
mT5	抗原に対する地域限定的な反応 炎症 延滞 炎症 膿疱 痛みなどです (Such as a locally limited response to an antigen, inflammation, delay, inflammation, pustules, and pain.)	mistranslation
ChatGPT	抗原に対する地域的な反応で，炎症，硬結，紅斑，かゆみまたは痛みを含む場合があります. (A localized reaction to an antigen, which may include inflammation, induration, erythema, itching, or pain.)	mistranslation
GPT-4	抗原への地域的に限定された反応で，炎症，硬結，紅斑，かゆみ，または痛みを伴うことがある. (A locally restricted reaction to an antigen, which may include inflammation, induration, erythema, itching, or pain.)	mistranslation

or the direction of causality. Nonetheless, by utilizing qualitative evaluations, and considering that good results have been obtained in these evaluations, we believe that the aforementioned shortcomings can be mitigated.

The limitations of this study include the use of an outdated version of the terminology collection (the latest being the fifth edition), and the inherent imbalance in the distribution of terms across categories. Furthermore, the

Table 4 The Japanese translation of "Extreme thirst accompanied by chronic excessive intake of water." Inside the parentheses is the English translation of the Japanese generated text

Models	Generated sentences	Explanation of underlined words
mBART50	極端な渇きを伴う慢性的な過剰な水摂取. (Chronic excessive intake of water accompanied by extreme thirst)	Incorrect word order
m2m-100-418M	極度の渇きは，慢性過剰な水分摂取に伴います. (Extreme thirst accompanies chronic excessive water intake.)	Incorrect word order and *incorrect use of a prepositional phrases*
m2m-100-1.2B	慢性的な過剰な水分摂取に伴う極度の渇き. (Extreme thirst accompanying chronic excessive water intake.)	*Incorrect use of a prepositional phrases*
GPT-3	極端な口渇に伴う水の過剰摂取の症状. (Symptoms of excessive intake of water accompanied by extreme thirst)	Incorrect word order and mistranslation
googletrans	慢性的な過剰な水を伴う極端な渇き. (Extreme thirst accompanied by chronic excessive water.)	Mistranslation
mT5	激しい渇きと慢性的な水の過剰摂取が伴います (It is accompanied by severe thirst and chronic excessive water intake.)	Mistranslation
ChatGPT	慢性的な水の過剰摂取に伴う極端な口渇. (Extreme thirst associated with chronic excessive water intake.)	*Incorrect use of a prepositional phrases*
GPT-4	慢性的な過度な水摂取に伴う極端な喉の渇き. (Extreme thirst associated with chronic excessive water intake.)	*Incorrect use of a prepositional phrases*

accuracy of publicly available GPT versions may change with future updates.

Conclusion

The optimal machine translation model for translating IMDRF-AET was GPT-4, which achieved the highest scores in both quantitative evaluations and visual assessments. Moving forward, we plan to advance our examination of glossary mapping based on these translation results. This study has highlighted the current capabilities and limitations of machine translation using LLM, it also opens up new avenues for research that could significantly impact the future of translation technology in medical domain.

Abbreviations

BERT	Bidirectional Encoder Representations from the Transformers
BLEU	Bilingual Evaluation Understudy
BP	brevity penalty
CER	Character Error Rate
GPT	Generative Pretrained Transformer
IMDRF	the International Medical Device Regulators Forum
AET	the Adverse Event Terminology
JFMDA	the Japan Federation of Medical Device Associations
mBART	multilingual bidirectional auto-regressive transformer
METEOR	Metric for Evaluation of Translation with Explicit ORdering
mT5	multilingual-T5
m2m-100	Many-to-Many multilingual translation model
WER	Word Error Rate

Acknowledgements

Not Applicable.

Authors' contributions

AY: Conceptualization, Methodology, Software, Data curation, Investigation, Writing- Original draft preparation. MU: Methodology, Software. HY: Supervision, Writing- Reviewing and Editing.

Funding

This research is supported by the Research on Regulatory Science of Pharmaceuticals and Medical Devices from the Japan Agency for Medical Research and development, AMED (Grant Number 24mk0101234s0403). The funder had no role in the design of the study, analysis, and interpretation of the data, or the writing of the manuscript.

Data availability

The data that support the findings of this study are available from the Ministry of Health, Labour and Welfare (https://www.japal.org/wp-content/uploads/2022/12/T220908I0040.pdf) and the International Medical Device Regulators Forum (https://www.imdrf.org/working-groups/adverse-event-terminology).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not Applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Faculty of Health Sciences, Hokkaido University of Science, 7-Jo 15-4-1 Maeda, Teine, Sapporo, Hokkaido 006-8585, Japan. [2]Faculty of Medical Management and Informatics, Hokkaido Information University, Ebetsu, Japan. [3]Department of medical informatics, Kagawa University Hospital, Kita-gun, Japan.

Received: 14 April 2024 Accepted: 3 February 2025
Published online: 08 February 2025

References

1. Japan Federation of Medical Device Associations. Utilization of the Medical Device Malfunction Terminology Collection. In Japanese. Available: https://www.jfmda.gr.jp/activity/committee/fuguai/. Accessed 15 Aug 2023.
2. The International Medical Device Regulators Forum. Adverse Event Terminology. Available: https://www.imdrf.org/working-groups/adverse-event-terminology. Accessed 15 Aug 2023.
3. Yagahara A, Uesugi M, Yokoi H. Identification of synonyms using definition similarities in Japanese Medical device adverse event terminology. Appl Sci. 2021;11:3659. https://doi.org/10.3390/app11083659.
4. Yagahara A, Yokoi H, Uesugi M. Machine translation of English Medical device adverse event terminology using deep learning. Japan J Med Inform. 2022;42(5):211–5.
5. Yagahara A, Uesugi M, Yokoi H. Evaluation of Machine Translation Accuracy Focused on the Adverse Event Terminology for Medical Devices. Stud Health Technol Inform. 2024;310:1450–1451. https://doi.org/10.3233/SHTI231239. PMID: 38269691.
6. Dew KN, Turner AM, Choi YK, Bosold A, Kirchhoff K. Development of machine translation technology for assisting health communication: a systematic review. J Biomed Inf. 2018;85:56–67. https://doi.org/10.1016/j.jbi.2018.07.018. Epub 2018 Jul 19. PMID: 30031857.
7. Noll R, Frischen LS, Boeker M, Storf H, Schaaf J. Machine translation of standardised medical terminology using natural language processing: a scoping review. N Biotechnol. 2023;77:120–9. Epub 2023 Aug 29. PMID: 37652265.
8. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States Medical Licensing examination? The implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312. https://doi.org/10.2196/45312. PMID: 36753318; PMCID: PMC9947764.
9. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198. PMID: 36812645; PMCID: PMC9931230.
10. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the National Medical Licensing examination in Japan: evaluation study. JMIR Form Res. 2023;7:e48023. https://doi.org/10.2196/48023. PMID: 37831496; PMCID: PMC10612006.
11. Kunitsu Y. The potential of GPT-4 as a Support Tool for pharmacists: Analytical Study using the Japanese National Examination for pharmacists. JMIR Med Educ. 2023;9:e48452. https://doi.org/10.2196/48452. PMID: 37837968; PMCID: PMC10644185.
12. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style examination: insights into current strengths and limitations. Radiology. 2023;307(5):e230582. https://doi.org/10.1148/radiol.230582. Epub 2023 May 16. PMID: 37191485.
13. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing physician and Artificial Intelligence Chatbot responses to patient questions posted to a Public Social Media Forum. JAMA Intern Med. 2023;183(6):589–96. https://doi.org/10.1001/jamainternmed.2023.1838. PMID: 37115527; PMCID: PMC10148230.
14. Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. Radiology. 2024;310(1):e232756. https://doi.org/10.1148/radiol.232756. PMID: 38226883.
15. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. Radiology. 2023;309(2):e232561. https://doi.org/10.1148/radiol.232561. PMID: 37987662.
16. OpenAI. GPT-4. https://openai.com/research/gpt-4. Accessed 4 Apr 2024.
17. Ministry of Health, Labour and Welfare. Regarding the Revision of the Translated Version of the IMDRF adverse event terminology for Medical Devices (Part 2). https://www.japal.org/wp-content/uploads/2022/12/T220908l0040.pdf. Accessed 4 Apr 2024.
18. googletrans. Available: https://pypi.org/project/Googletrans/.. Accessed 15 Aug 2023.
19. Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer, arXiv, 2021.
20. Liu Y, Gu J, Goyal N, Li X, Edunov, Ghazvininejad M, Lewis M, Zettlemoyer L. Multilingual denoising pre-training for neural machine translation. Trans Association Comput Linguistics. 2020;8:726–42.
21. Fan A, Bhosale S, Schwenk H, et al. Beyond English-Centric Multilingual Machine Translation, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2020. p. 483–498.
22. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
23. OpenAI. ChatGPT. Available: https://openai.com/chatgpt. Accessed 15 Aug 2023.
24. Ott M, Edunov S, Baevski A et al. fairseq: A Fast, Extensible Toolkit for Sequence Modeling, arXiv, 2019. Available: https://github.com/pytorch/fairseq. Accessed 15 Aug 2023.
25. Papineni K, Roukos S, Ward T, Zhu WJ. (2002). BLEU: a method for automatic evaluation of machine translation. 40th Annual meeting of the Association for Computational Linguistics. 2002. p. 311–318.
26. Satanjeev Banerjee, Alon Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Proceedings of the Second Workshop on Statistical Machine Translation. 2007. p. 228–231.
27. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y, BERTScore. Evaluating Text Generation with BERT. International Conference on Learning Representations. 2, Proceeding of International Conference on Learning Representations (ICLR) 2020. 2019.
28. Kundel HL, Polansky M. Measurement of observer agreement. Radiology. 2003;228(2):303–8. https://doi.org/10.1148/radiol.2282011860. Epub 2003 Jun 20. PMID.

Publisher's note