Assessing the Efficacy of GPT-4 in the Sentence-Level Translation of Medical Abstracts: A Comparative Study

Ana Pacheco, Kahmeeh Obey, Nicole Luzuriaga, James Li

Abstract

This study evaluates GPT-4-Turbo's performance for sentence-level translation of medical abstracts, crucial for accurate dissemination of medical knowledge across languages. Using the Biomedical MQM dataset (annotated WMT21 Bio Task data), we assess GPT-4-Turbo's accuracy across sentences in 12 language pairs by using the translation prompt adopted by the WMT23 Bio Task for sentence-level translation. Evaluation metrics include BLEU, CHRF++, TER, and a sample of human assessment for Pt ⇒ En. We find that GPT-4-Turbo shows relatively good BLEU scores across most languages pairs but does not match the high performances and low error rates of Google Translate and Deepl. Out of all language pairs, it performs considerably worse in English to German and English to Chinese, which might hint to a lack of biomedical training data in those languages. In comparison to past studies with ACL's WMT Bio Task dataset, however, there seems to be a decrease in disparities in BLEU scores for sentence-level translations from previous OpenAI models to GPT-4-Turbo, which could indicate that LLMs are becoming better and more accurate machine translators. We find that LLM's show great potential for machine translation, but that there is a need for further investigation using broader datasets and a larger sample of human assessments to reach more definitive conclusions and to avoid data contamination. Please refer to our repository: https://github.com/anaspacheco/bio_mqm_reference

1 Introduction

The field of computational linguistics has witnessed significant advancements in machine translation (MT) in recent years, particularly with the dominance of neural machine translation over statistical machine translation (Song, 2022). Within the domain of MT, text translation stands out as its most common application (e.g. webpage translation, e-chat translation, language learning, scientific literature translation and others) (Wang et al, 2022). In this domain, the translation of biomedical literature poses substantial challenges for MT. These challenges emerge from the intricate and diverse terminologies, the specialized syntactic structures, and the paramount need for precision to preserve the integrity of research findings. However, the scope of potential applications is extensive, spanning from the distribution of multilingual public health materials to the publication of multilingual research articles and abstracts.

Prior research has demonstrated the effectiveness of neural translation tools such as Deepl, Google Translate, and other NMTs in producing high-quality translations for biomedical abstracts (which are further analyzed in section 2). Now, for this study, we seek to conduct a robust examination of GPT-4's capacity to translate sentences from medical abstracts. Chat-GPT [1] (which uses the GPT-4 model) is an advanced Large Language Model, which offers advanced language processing capabilities that could be valuable for machine translation. In this way, we seek to evaluate GPT-4's translation performance, using commercially available NMTs as benchmarks. We compare GPT-4-Turbo (April 2023) with Google Translate[2] and DeepL[3] translate APIs using the Biomedical Multidimensional Quality Metrics Dataset (Zouhar et al., 2024).

We seek to contribute to the ongoing discourse on the potential applications of LLMs in NLP by providing a detailed analysis of GPT-4's translation capabilities in the context of sentence-level biomedical abstract translation. The subsequent sections outline the methodology implemented in our study, present the results, and discuss their

[1] https://chatgpt.com/?oai-dm=1
[2] https://translate.google.com/
[3] https://www.deepl.com/translator

implications within the context of machine translation and biomedical research.

2 Related Works

The advancement of neural machine translation (NMT) tools like DeepL, Google Translate, and CUBBITT [4] has been pivotal in addressing the growing need for accurate translation of biomedical literature across linguistic barriers. Prior studies have demonstrated the capability of NMT systems to produce translations that often parallel the quality of human translations, particularly in the medical domain, where precision is critical (Wang et al., 2022).

Significant research has already been conducted on the performance of NMT systems in translating complex medical texts. In the findings of the WMT22 Biomedical Task (Neves et al, 2022), many of the submitted MT translations of the test set achieved a similar quality to the reference translations in various language pairs. Moreover, a comparative study of DeepL, Google Translate, and CUBBITT highlighted their efficacy in translating French medical research abstracts to English, finding that the translations produced by DeepL and Google Translate were comparable in quality to the original reference abstract (Smith et al, 2024).

Regarding LLMs as machine translators, research into document-level and sentence-level machine translation using Large Language Models (LLMs) such as GPT-4 have shown promising results. In an extensive study conducted by Jiao et al. (2023), GPT-4 was shown to perform competitively with commercial MTs (Google Translate and Deepl) on high resource European Languages, showing excellent performance on spoken and informal language, but a significantly worse performance on biomedical abstracts. This suggests that LLMs may represent a new paradigm in translation technology (Kim et al., 2023), but that they may still lag in biomedical translation. This aligns with our project's focus on GPT-4 and the study of its potential in the sentence-level translation of biomedical literature.

Adding to this, in the Biomedical Translation Task of the WMT23, the findings reported that ChatGPT 3.5 (their chosen comparison system) performed very well in comparison to many of the machine translation submissions. More specifically, they found that, across all examined language pairs, the translations generated by GPT-3.5 exhibited a comparable quality to the reference translation at the sentence level, showing no significant variance in the outcomes (Neves et al., 2023).

Lastly, a study by Raunak et al (2023) found that GPT-4's was effective in enhancing translation quality and rectifying major errors in NMT-generated translations across various language pairs. Additionally, human evaluations indicated a significant improvement in edit trustworthiness compared to previous LLMs. One downside however, is that the study highlighted the potential for hallucinated edits by GPT-4, emphasizing the importance of cautious utilization in expert translation post-editing.

These studies collectively inform the current research's methodology and objectives, grounding our investigation in a well-established scholarly context while aiming to contribute new insights into the capabilities and limitations of GPT-4-Turbo as a medical translation tool.

3 GPT-4-Turbo for Biomedical MT

3.1 Baselines

We compare GPT-4-Turbo (April, 2023) with two popular commercial systems: Google Translate and DeepL Translate. Both are Neural Machine Translation (NMT) engines, and as of today, Google Translate supports 133 languages whereas Deepl supports 32 languages. We choose the latest GPT model available to date, in order to compare results with existing research with older models.

3.2 Data

We evaluated the translation systems using the Biomedical Multidimensional Quality Metrics Dataset (Zouhar et al., 2024). This dataset encompasses 12 language pairs within the biomedical field, with approximately 25k segment-level annotations. It was curated from submissions to the WMT21 biomedical translation shared task (Yeganova et al., 2021), which is in turn derived from bilingual abstracts of academic papers in the MEDLINE corpus sourced from the National Library of Medicine (NLM). Linguists with expertise in the medical domain crafted reference translations and MQM annotations for this shared task dataset. To

[4] https://ufal.mff.cuni.cz/cubbitt

ensure accurate results, we selected 100 professionally translated reference sentences for each language pair and filtered out those who were marked by annotators to have any errors considered to be *Critical*. Additionally, we kept sentences that were marked to have *Minor* or *Neutral* errors, to avoid overly selecting for sentences that might be linguistically more complex. It is important to consider that since this dataset was produced from the WMT21 biomedical shared task (Yeganova et al., 2021), the systems used in this study might have been trained with this data, but since all three systems are not open-source, access to their training methodologies and specific data is limited.

It is important to make clear that our initial intention was to use the WMT22 or WMT23 Bio Test Set to compare our results with previous findings on the same test sets, but we had difficulties in processing it. The instructions available were unclear and the test set was not perfectly aligned between language pairs, and in that way, we could not find references for many of the test sentences. Our team initially attempted to produce results out of the WMT22 test set but found extremely low BLEU scores, which were caused by most sentences being misaligned. It is because of this that we decided to use the Bio MQM Dataset instead.

3.3 Translation Prompt

For the translation prompt, we have decided to use the same prompt used in the WMT23 Bio Task: "**You are a helpful assistant specialized in biomedical translation. You will be provided with a sentence in {src}, and your task is to translate it into {trg}**" where {src} is the source language and {trg} is the target language. The prompt proved to be mostly successful and did not result in overwhelming system hallucination. A manual evaluation showed that GPT-4-Turbo produced valid translations of sentences with the exception of 2 sentences in the direction of En ⇒ Zh that were returned in English instead of Chinese (simplified).

3.4 Automatic Metrics

We chose to evaluate using the same metrics used by (Jiao et al, 2023), utilizing the widely employed BLEU score (Papineni et al., 2002) as our main evaluation criterion. Additionally, we include ChrF++ (Popovic´, 2017) and TER (Snover et al., 2006). All of these metrics were provided by SacreBLEU (Post, 2018).

3.5 Multilingual Translation

We evaluate the following language pairs: En ⇒ De, De ⇒ En, En ⇒ Pt, Pt ⇒ En, En ⇒ Es, Es ⇒ En, En ⇒ Fr, Fr ⇒ En, En ⇒ Ru, Ru ⇒ En, Zh ⇒ En, En ⇒ Zh. The BLEU, CHRF++, and TER scores for sentence-level translations of a sample of the Bio MQM Dataset are summarized in Table 1. Examples of output are presented in Table 2. For comparison, we present the BLEU scores for GPT-4-Turbo translations alongside the performance changes (+/-) relative to the highest performing commercial MT. Please refer to Table 4 for detailed findings.

Table 1: Examples from the BIO MQM Dataset and sentence translations for Pt ⇒ En

	Pt ⇒ En
SRC	Pré-operatoriamente, a marcha por insuficiência do glúteo médio esteve presente em todos os sujeitos e se tornou negativa em dez deles.
REF	Preoperatively, a gait due to gluteus medius insufficiency was present in all subjects and became negative in ten of them.
Google Translate	Preoperatively, a gait due to gluteus medius insufficiency was present in all subjects and became negative in ten of them.
Deepl	Preoperatively, a gait due to gluteus medius insufficiency was present in all subjects and became negative in ten of them.
GPT-4-Turbo	Preoperatively, all subjects exhibited a gait due to gluteus medius insufficiency, which resolved in ten of them.

Table 2: Performance analysis (MQM Bio Data)

Direction	Deepl	Google Translate	GPT-4-Turbo
En ⇒ De	BLEU: 98.34 CHRF++:99.67 TER: 0.69	BLEU:95.20 CHRF++:98.20 TER: 2.63	BLEU: **79.63** CHRF++: 89.88 TER: 10.37
De ⇒ En	BLEU: 98.34 CHRF++: 99.62 TER: 0.70	BLEU: 99.09 CHRF++: 99.69 TER: 0.39	BLEU: **84.30** CHRF++: 93.36 TER: 14.62
En ⇒ Pt	BLEU: 92.33 CHRF++: 97.13 TER: 3.96	BLEU: 89.24 CHRF++: 95.36 TER: 5.49	BLEU: **82.47** CHRF++: 91.47 TER: 10.46
Pt ⇒ En	BLEU: 96.74 CHRF++: 99.26 TER: 1.14	BLEU: 97.05 CHRF++: 99.21 TER: 1.06	BLEU: **86.49** CHRF++: 93.85 TER: 8.92
En ⇒ Es	BLEU: 99.77 CHRF++: 99.86 TER: 0.10	BLEU: 98.35 CHRF++: 99.42 TER: 1.08	BLEU: **89.83** CHRF++: 94.90 TER: 7.65
Es ⇒ En	BLEU: 99.57 CHRF++: 99.80 TER: 0.26	BLEU: 98.20 CHRF++: 99.61 TER: 0.37	BLEU: **88.79** CHRF++: 94.86 TER: 7.18
En ⇒ Fr	BLEU: 98.42 CHRF++: 99.55 TER: 0.94	BLEU: 72.54 CHRF++: 95.58 TER: 5.58	BLEU: **85.13** CHRF++: 92.17 TER: 10.12
Fr ⇒ En	BLEU: 99.71 CHRF++: 99.94 TER: 0.13	BLEU: 96.87 CHRF++: 99.41 TER: 0.64	BLEU: **92.56** CHRF++: 96.77 TER: 4.70
En ⇒ Ru	BLEU: 92.73 CHRF++: 98.49 TER: 2.85	BLEU: 73.28 CHRF++: 85.94 TER: 17.33	BLEU: **84.18** CHRF++: 92.89 TER: 10.57
Ru ⇒ En	BLEU: 99.32 CHRF++:99.86 TER: 0.35	BLEU: 97.98 CHRF++: 99.49 TER: 0.67	BLEU: **91.21** CHRF++: 95.72 TER: 5.71
En ⇒ Zh	BLEU: 74.37 CHRF++: 94.94 TER: 6.13	BLEU: 34.76 CHRF++: 66.14 TER: 76.07	BLEU: **20.45** CHRF++: 77.55 TER: 101.23
Zh ⇒ En	BLEU: 74.62 CHRF++:93.37 TER: 33.03	BLEU: 95.53 CHRF++: 99.47 TER: 0.74	BLEU: **90.64** CHRF++: 95.49 TER: 6.67

3.6 Manual Evaluation

We carried out human validation of the quality of translations for only one of the language pairs (Pt ⇒ En). This is because it was difficult to find native speakers in the other languages (in a limited time frame) who had the interest and time to volunteer in this study. Given that manual evaluation is the gold standard for machine translation, we acknowledge that this limits our ability to reach definitive conclusions. Nonetheless, we decided to carry out manual evaluation for Pt ⇒ En, using the same classification system as used in the WMT23 Bio Task (Neves et al., 2023). We presented three annotators with 10 lines at random from each machine translation and asked annotators to rank them in the following manner: using the "3-way-ranking" task given by the Appraise tool, compare three sentences. For example, for pt2en: (i) is the source text in Portuguese, (ii) is the translation A and (iii) is the translation B in English. The annotator was asked to select one of the following four options: (i) A=B, i.e., both translations have similar quality; (ii) A>B, i.e., translation A is better than translation B; (iii) translation A is worse than translation B (Neves et al., 2023). We present the results on Table 3.

Table 3: Pairwise human evaluation results for Pt ⇒ En

Direction	Pair	Total	A>B	A=B	A<B
pt2en	Ref vs. Deepl	30	4	25	1
	Ref vs. Google Translate	30	5	23	2
	Ref vs. GPT	30	7	22	1

4 Results

We found that that while GPT-4 exhibits strong potential, it currently lags behind Google Translate and DeepL in BLEU scores and error rates for most language pairs. Now, this finding does not necessarily mean that GPT-4 is inferior as a MT, especially because of two considerations: (1) the extremely high scores presented in most language pairs by Google Translate and Deepl may indicate that they have been trained using the data which this MQM dataset relies upon and (2) manual evaluation for all language pairs is needed to reach more definitive conclusions. Considering this, we find that our study (despite its limitations) has found a significant gap in performance across multiple language pairs. For comparison, we present the BLEU scores for GPT-4-Turbo translations alongside the performance changes (+/-) relative to the highest performing commercial MT (Table 4).

Table 4: Performance of GPT-4-Turbo compared to highest performing commercial MT

Direction	BLEU Score Percent Difference
En ⇒ De	-21.1%
De ⇒ En	-15.4%
En ⇒ Pt	-11.3%
Pt ⇒ En	-11.5%
En ⇒ Es	-10.5%
Es ⇒ En	-11.4%
En ⇒ Fr	-14.5%
Fr ⇒ En	-7.4%
En ⇒ Ru	-9.7%
Ru ⇒ En	-8.5%
En ⇒ Zh	-113.73%
Zh ⇒ En	-5.3%

This difference in performance seem to highlight the advantages that well-trained NMT systems hold in handling nuances of specialized fields like biomedicine, which might indicate that they have been heavily trained in biomedical data, or even possibly Pubmed articles.

Although our manual evaluation was very limited, we found that it correlated well with BLEU scores for the Pt ⇒ En direction, with the pairwise rankings corresponding with the ranking in BLEU scores. Annotators were also asked to give a brief explanation for their rankings in the case of discrepancies of quality between sentences, and we found that in most cases, GPT was ranked lower

because it would often only include the abbreviations (and not the complete term) for medical terminology, produce acronyms mistranslations, and also omit punctuation, which could explain the lower BLEU score observed. Nonetheless, there was a consensus that context and meaning were preserved for almost all sentences translated by GPT-4.

Another interesting finding is that we observed high discrepancies in performance in the following language pairs: En ⇒ De, De ⇒ En, En ⇒ Zh. Although we observed a degree of hallucination in the En ⇒ Zh translations, it would not be enough to explain the extremely high disparity between the performance of GPT-4 and DeepL for that pair. The disparities could be attributed to various factors. For En ⇒ De and De ⇒ En, differences in grammatical structures, word order, and idiomatic expressions between English and German could pose challenges for GPT-4. Similarly, translating between English and Chinese (Zh) involves significant linguistic challenges, such as character-based writing and distinct syntactic patterns, which may lead to lower performance. Further manual evaluation and case-by-case studies in the specific sentences would provide more specific insights into the extreme differences observed in translation performance. Interestingly, GPT-4-Turbo displayed competitive quality in En ⇒ Ru translations in comparison to Google translate, and reasonable strength with Fr ⇒ En, Ru ⇒ En, and Zh ⇒ En translations, suggesting further investigation into its strength in these languages may be worthwhile.

In comparison with the GPT-3.5 performance in WMT23 Bio data, we observe similar relative scores across the language pairs of En ⇒ De, En ⇒ Es, En ⇒ Pt, En ⇒ Ru. We find that GPT-4 performs exceptionally better in En ⇒ Pt and En ⇒ Es than in the other pairs mentioned. We do find a difference however, in the En ⇒ Fr pair, with GPT-4-Turbo performing significantly worse in the MQM dataset for that direction than GPT-3.5 in the WMT23 test set. It would be beneficial to conduct a study on the WMT23 test set for elucidation.

5 Conclusion

In this study, we provide a comparative analysis of GPT-4's capabilities in translating biomedical abstracts against the well-established MT systems, Google Translate and DeepL.

We find that GPT-4 showed systematically lower BLEU scores, with En ⇒ De, De ⇒ En, and En ⇒ Zh translations proving particularly challenging. Surprisingly, we observed reasonably good scores for En ⇒ Ru, Fr ⇒ En, Ru ⇒ En, and Zh ⇒ En translations, which need to be further investigated. Lastly, although we could not conduct extensive manual evaluation, we found that the annotators' assessment of the Pt ⇒ En pair correlated well with our findings for that language. Further manual evaluation is needed to truly assess the quality of GPT-4's sentence-level translation of biomedical abstracts.

We find therefore, that GPT-4, while not specifically designed for biomedical translation, shows promise, but to become a truly reliable tool in this sensitive domain, further investigation is necessary. As LLMs like GPT-4 continue to evolve, ongoing evaluation, refinement, and understanding of their strengths and limitations will be essential for harnessing their full potential in the critical field of biomedical translation.

6 Limitations

As an initial investigation, this study is still in its early stages and requires further development to enhance its reliability across various aspects:

- **Scope**: We randomly sampled 100 sentences for each language pair for evaluation, resulting in an incomplete data coverage. Also, since GPT-4 is an LLM and produces different outcomes for identical queries, it would be beneficial to conduct multiple translations and to report the average results.

- **Limited Human Evaluation**: Since we had a very limited timeframe to complete our study, we could not find annotators to manually evaluate the quality of the machine translations. In this way, we only had BLEU, CHRF++, and TER scores to base our findings upon, which are not as reliable as manual evaluations. Also, it is important to note that manual evaluation for biomedical data should be done by people with expertise in the field, which would therefore be able to point out errors in acronyms and more specialized terminology.

- **Language Covering**: All of the language pairs used in this study are high-resource languages. It would be beneficial to analyze GPT-4's performance on a biomedical test set for low level languages.

- **Limited Work Done on Chosen Dataset**: Since we had difficulties in obtaining the parallel set for WMT23 test set, we had to use the MQM Bio Dataset, which has not been used in previous studies in this topic, making it hard to compare our evaluation with previous research findings.

- **Data Contamination**: Since the MQM Bio Dataset is based on the WMT21 Bio Task dataset, the machine translators used in this study might have been trained with that data, which would result in inaccurate findings. Given this, it would be interesting to use a more extensive and newer dataset for more conclusive results.

7 Individual Contributions

Ana Pacheco: found the dataset, cleaned it for processing, collaborated with Kahmeeah to code the algorithm to produce translations the Google Translate, Deepl, and GPT-4 APIs. Introduced CHRF++ and TER for use in our metrics. Reached out for friends to produce the manual evaluations.

Nicole Luzuriaga: did extensive research on previous findings and collected notes on them. Defined the automatic metrics that were going to be used in the research.

Kahmeeah Obey: collaborated with Ana to produce the code necessary for the translations. Was also responsible for the configuration and set up of the APIs used in this study as well as the coding for the automatic evaluation.

James Li: collaborated with Ana and Kahmeeah to create the code used to conduct automatic evaluation using Sacre-BLEU.

References

Song R. Analysis on the Recent Trends in Machine Translation. Highlights Sci Eng Technol. 2022 Nov 10; 16:40–7.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, Kenneth Ward Church, Progress in Machine Translation, Engineering, Volume 18, 2022, Pages 143-153, ISSN 2095-8099, https://doi.org/10.1016/j.eng.2021.03.023.

Zouhar, V., Meister, C., Gastaldi, J. L., Du, L., Vieira, T., Sachan, M., & Cotterell, R. (2024). Fine-Tuned Machine Translation Metrics Struggle in Unseen Domains. arXiv preprint arXiv:2402.18747. Retrieved from https://arxiv.org/abs/2402.18747

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, Christel Gerardin, Rachel Bawden, Darryl Johan Estrada, Salvador Lima-lopez, Eulalia Farre-maduel, Martin Krallinger, Cristian Grozea, and Aurelie Neveol. 2022. Findings of the WMT 2022 Biomedical Translation Shared Task: Monolingual Clinical Case Reports. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 694–723, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Smith, J., et al. (2021). Performance of Machine Translators in Translating French Medical Research Abstracts to English. PLoS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0297183

Jiao, Wenxiang & Wang, Wenxuan & Huang, Jen-Tse & Wang, Xing & Shi, Shuming & Tu, Zhaopeng. (2023). Is ChatGPT A Good Translator? A Preliminary Study. https://arxiv.org/abs/2301.08745

Kim, Y. H., et al. (2023). Document-Level Machine Translation with Large Language Models. EMNLP 2023. https://aclanthology.org/2023.emnlp-main.1036.pdf

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 Biomedical Translation Shared Task: Evaluation of ChatGPT 3.5 as a Comparison System. In Proceedings of the Eighth Conference on Machine Translation, pages 43–54, Singapore. Association for Computational Linguistics.

Raunak, Vikas & Sharaf, Amr & Awadallah, Hany & Menezes, Arul. (2023). Leveraging GPT-4 for Automatic Translation Post-Editing.

Yeganova, L., Wiemann, D., Neves, M., Vezzani, F., Siu, A., Jauregi Unanue, I., Oronoz, M., Mah, N., Névéol, A., Martinez, D., Bawden, R., Di Nunzio, G. M., Roller, R., Thomas, P., Grozea, C., Perez de Viñaspre, O., Vicente Navarro, M., & Jimeno Yepes, A. (2021). Findings of the WMT 2021 Biomedical

Translation Shared Task: Summaries of Animal Experiments as New Test Set. Proceedings of the Sixth Conference on Machine Translation. https://publica.fraunhofer.de/handle/publica/418621

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation.

Maja Popovic. 2017. ChrF++: Words helping character ´ n-grams.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation.

Matt Post. 2018. A call for clarity in reporting bleu scores.