Comparing ChatGPT's and Human Evaluation of Scientific Texts' Translations from English to Portuguese Using Popular Automated Translators

Notebook for the SimpleText Lab at CLEF 2023

Sílvia Araújo [1] and Micaela Aguiar [1]

[1] *University of Minho, Rua da Universidade, Braga, 4710 - 057, Portugal*

Abstract

This paper addresses the challenge researchers face when translating their work. Due to the high cost associated with human translation services, many researchers turn to automatic translation tools as a cost-effective alternative. Therefore, assessing the quality of these translations is crucial. This paper presents a comparative evaluation of translations using both human assessments and ChatGPT. Our study focuses on the translation of a scientific text excerpt from English to Portuguese. We analyze the performance of ChatGPT in two scenarios: comparative evaluations with all translations presented in a single prompt, applied five times to test for consistency, and 20 individual evaluations (five evaluations per translation) in separate chats. In both scenarios, ChatGPT's assessments exhibit higher consistency in terms of fluency, appropriateness, accuracy, and overall assessment compared to human evaluations. The results also reveal a consensus between human evaluations and ChatGPT assessments regarding the translation with the lowest score, while discrepancies arise in the evaluations of top-performing translations. Finally, the ability to engage in follow-up questions, receive suggestions for improvement, and compare translations using ChatGPT's own recommendations proves a valuable tool for researchers seeking to assess and improve the translation quality of their work.

Keywords

Automatic Translation, Scientific Texts, Machine Translation Evaluation template

1. Introduction

In an era of global scientific collaboration and knowledge sharing, the ability to promptly and effortlessly access scientific information is of vital importance [1]. As researchers and science communicators strive to disseminate their findings across linguistic boundaries, the need for reliable and accurate translation tools becomes increasingly apparent. The effective translation of scientific texts plays a pivotal role in bridging language barriers, promoting cross-cultural collaboration, and fostering the global exchange of knowledge. However, finding translators available at any time and for any purpose in the real world can be a daunting task, and the cost of relying on highly-trained bilingual individuals in terms of labor and time can be prohibitive [2].

Consequently, researchers frequently turn to automated translators, which offer the potential of efficient and cost-effective translations. However, the evaluation of these translations is essential to ensure that the conveyed scientific information remains accurate, coherent, and comprehensible in the target language. It is imperative to evaluate the quality of these automated translations to ensure that

[1] CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: saraujo@elach.uminho.pt (A. 1); maguiar60@mail.com (A. 2)
ORCID: 0000-0003-4321-4511 (A. 1); 0000-0002-5923-9257 (A. 2)

scientific communication remains accurate and easily understood. For individual researchers, manual translation evaluation presents similar drawbacks to human translation including time and cost constraints, limited adjustability, and a lack of reproducibility [3].

When it comes to automatic translation evaluation, multiple metrics have been developed, including BLEU [4], which measures n-gram overlap between machine translations and human references; ROUGE [5], which examines content overlap and coherence in model-generated and human reference summaries; or METEOR [6], which combines precision, recall, and flexible unigram matching, accommodating synonyms and morphological variants. Over the past few decades, language models have emerged as valuable resources for assessing and evaluating machine translations [7] [8] [9]. Recently, Kocmi and Federmann [10] employed pre-trained large language models (LLMs) to assess translation quality. The study evaluated seven distinct models of GPT (Generative Pretrained Transformer) [11] and found that GPT 3.5, as well as larger models such as Davinci-002, Davinci-003, and ChatGPT, yielded remarkably competitive outcomes.

Yet, these metrics and methods were primarily designed for research and technical purposes making their practical application extremely challenging for researchers who do not possess advanced technical skills. These evaluation metrics often require a deep understanding of computational linguistics, statistical analysis techniques, and programming languages. Researchers without a strong technical background would find it difficult to navigate and utilize these metrics effectively in real-world scenarios.

In light of this background, our primary goal in this paper is to evaluate the performance of ChatGPT, a user-friendly chatbot developed by Open AI that uses Large Language Models and is accessible to the general public, in the assessment of automatic translations of scientific texts. To accomplish this, we translated a segment of a scientific text using popular automated translators and subjected it to evaluation by both human assessors and ChatGPT. The objective was to determine the extent of agreement between human evaluation and ChatGPT's assessment. Three key research questions were developed to guide our investigation:

1. To what extent does ChatGPT's evaluation of automatic translations of scientific texts align with human assessment?
2. Can ChatGPT be a reliable tool for researchers and science communicators in evaluating automatic translations of scientific texts?
3. What are the strengths and/or limitations of ChatGPT in evaluating the quality of automatic translations of scientific texts?

Next, we will delve into the detailed description of the methodologies utilized in conducting this experiment. We will provide a comprehensive overview of the steps taken to carry out the evaluation, including the selection of scientific texts, the choice of popular automated translators, and the specific criteria used for the translation assessment.

2. Methods

ChatGPT, powered by Open AI's state-of-the-art Large Language Model GPT-3.5, is an advanced chatbot that was launched to the public in November 2022. It showcases an impressive ability to comprehend and respond to natural language, closely mimicking human communication patterns. The remarkable growth of ChatGPT's user community, attracting over one million subscribers in its initial week, has thrust it into the limelight, making ChatGPT a true "cultural sensation" [12]. This is one of the reasons why we selected ChatGPT as a potential tool for researchers and science communicators to assess their translations of scientific texts. These reasons can be succinctly summarized in three key considerations: (1) ChatGPT harnesses advanced artificial intelligence technology, including large language models, which have been recognized as cutting-edge evaluators of translation quality, as highlighted by Kocmi and Federmann [10], (2) ChatGPT has been specifically designed for use by the general public, offering a user-friendly interface and intuitive navigation, (3) ChatGPT has gained widespread popularity and is already being widely adopted by users.

In terms of the evaluated translations, we selected an excerpt from the article "General Gynecologic Evaluation" by David H. Barad [13], sourced from the MSD Manual Professional Version, as our original text in English. We assessed four different Portuguese translations for this study. Two of these translations were obtained from popular automated translation tools: Google Translate and DeepL.

Despite not being explicitly tailored for translation tasks, we included ChatGPT in our evaluation to gauge its performance. Several studies have examined ChatGPT's potential for Machine Translation [14] [15] and Hendly et al. [16] have demonstrated its ability to produce great quality translations in high-resource languages. Additionally, we included the official Portuguese translation of the corresponding article from the MSD Manuals. The original excerpt and its translations are listed in Table 1.

Table 1

Original Excerpt and Evaluated Translations

Source	Translation
Original Text	Most women, particularly those seeking general preventive care, require a complete history and physical examination as well as a gynecologic evaluation. Gynecologic evaluation may be necessary to assess a specific problem such as pelvic pain, vaginal bleeding, or vaginal discharge. Women also need routine gynecologic evaluations, which may be provided by a gynecologist, an internist, or a family practitioner; evaluations are recommended every year for all women who are sexually active or > 18 years. Obstetric evaluation focuses on issues related to pregnancy. A gynecologic or an obstetric evaluation may include a pelvic examination when indicated based on history or symptoms and agreed to by the patient.
Google Translate	A maioria das mulheres, particularmente aquelas que procuram cuidados preventivos gerais, requer uma história completa e exame físico, bem como uma avaliação ginecológica. A avaliação ginecológica pode ser necessária para avaliar um problema específico, como dor pélvica, sangramento vaginal ou corrimento vaginal. As mulheres também precisam de avaliações ginecológicas de rotina, que podem ser fornecidas por um ginecologista, internista ou médico de família; avaliações são recomendadas todos os anos para todas as mulheres sexualmente ativas ou > 18 anos. A avaliação obstétrica enfoca questões relacionadas à gravidez. Uma avaliação ginecológica ou obstétrica pode incluir um exame pélvico quando indicado com base na história ou sintomas e acordado pelo paciente.
ChatGPT	A maioria das mulheres, especialmente aquelas que procuram cuidados preventivos gerais, necessita de um histórico completo e de um exame físico, bem como de uma avaliação ginecológica. A avaliação ginecológica pode ser necessária para avaliar um problema específico, como dor pélvica, sangramento vaginal ou corrimento vaginal. As mulheres também precisam de avaliações ginecológicas de rotina, que podem ser realizadas por um ginecologista, um internista ou um médico de família; recomenda-se que essas avaliações sejam feitas todos os anos para todas as mulheres sexualmente ativas ou com mais de 18 anos. A avaliação obstétrica concentra-se em questões relacionadas à gravidez. Uma avaliação ginecológica ou obstétrica pode incluir um exame pélvico quando indicado com base no histórico ou nos sintomas e concordado pela paciente.

DeepL	A maioria das mulheres, especialmente as que procuram cuidados preventivos gerais, necessitam de uma história completa e de um exame físico, bem como de uma avaliação ginecológica. A avaliação ginecológica pode ser necessária para avaliar um problema específico, como dor pélvica, hemorragia vaginal ou corrimento vaginal. As mulheres também necessitam de avaliações ginecológicas de rotina, que podem ser efectuadas por um ginecologista, um internista ou um médico de família; recomenda-se a realização de avaliações anuais a todas as mulheres sexualmente activas ou com idade superior a 18 anos. A avaliação obstétrica centra-se em questões relacionadas com a gravidez. Uma avaliação ginecológica ou obstétrica pode incluir um exame pélvico quando indicado com base na história ou nos sintomas e com o acordo da paciente.
MSD Portuguese Translation	A maioria das mulheres, principalmente as que procuram o exame ginecológico para rastrede [sic] rotina, requer história e exame físico completos e avaliação ginecológica. A avaliação ginecológica pode ser necessária para diagnosticar um problema específico, como dor pélvica, sangramento ou corrimento vaginal. As mulheres também necessitam de avaliação ginecológica de rotina, que pode ser feita por ginecologista, médico generalista ou da família; este tipo de avaliação é recomendado anualmente para todas as mulheres sexualmente ativas ou com > 18 anos. As avaliações obstétricas focalizam questões relacionadas à gestação. A avaliação ginecológica ou obstétrica pode incluir exame pélvico quando indicado de acordo com a anamnese ou sintomas e com o acordado com a paciente.

Regarding the criteria, we drew upon the proposals and discussions presented by White et al. [17], Church & Hovy [18], Blanchon & Boitet [19], and defined three criteria: fluency (it is intuitively acceptable and can be reasonably interpreted by a native speaker), appropriateness (it is the degree to which the information present in the original text was conveyed in the translation), and accuracy (it pertains to the precise and accurate utilization of terminology). For each criterion, we employed a Likert scale ranging from 1 to 5, with 1 indicating poor, 2 denoting average, 3 representing fair, 4 signifying good, and 5 reflecting excellent.

The human evaluation took place in the context of the II Symposium on Post-editing and Multilingual Information Retrieval, organized by the University of Aveiro, held on May 17-18, 2023. Participants were asked to complete a form structured as follows. The header contained the task description, the criteria to be applied, and the original text. Following that, the four translations were presented, and for each translation, participants had to evaluate the three established criteria on a scale of 1 to 5 and provide an overall assessment.

We conducted two distinct scenarios to assess the reliability of the evaluations obtained from ChatGPT.

2.1.Scenario 1

In the first scenario, we provided the criteria, original text, and the four translations. The prompt used was as follows:

“Evaluate the translations based on the following criteria and rate each criterion on a Likert scale of 1-5. Use the following scale: 1 for poor, 2 for average, 3 for fair, 4 for good, and 5 for excellent.

Criteria: Fluency: it is intuitively acceptable and can be reasonably interpreted by a native speaker. Appropriateness: it is the degree to which the information present in the original text was conveyed in the translation. Accuracy: it pertains to the precise and accurate utilization of terminology.

Give an overall evaluation of 1 to 5 for each translation. Summarise the results in a table.

The original text is: '[text]'

Translation 1 : '[text]'

Translation 2 : '[text]'

Translation 3 : '[text]'

Translation 4 : '[text]'"

To verify the consistency of the results, we applied the same prompt in different conversations five times, by opening a new chat every time. This approach prevented ChatGPT from accessing the history of prior attempts.

2.2.Scenario 2

In the second scenario, we requested ChatGPT to assess a single translation in each chat, conducting a total of 20 evaluations (five evaluations per translation).The objective of this scenario is to compare the results obtained by each translation in the first scenario and determine if they remain consistent when evaluated individually. The prompt used was as follows:

"Evaluate the translation based on the following criteria and rate each criterion on a Likert scale of 1-5. Use the following scale: 1 for poor, 2 for average, 3 for fair, 4 for good, and 5 for excellent.

Criteria: Fluency: it is intuitively acceptable and can be reasonably interpreted by a native speaker. Appropriateness: it is the degree to which the information present in the original text was conveyed in the translation. Accuracy: it pertains to the precise and accurate utilization of terminology.

Give an overall evaluation of 1 to 5 for each translation. Summarise the results in a table.

The original text is: '[text]'

Translation: '[text]'"

In the following section, we will present the findings from both the human evaluation and the evaluation conducted with ChatGPT in the two scenarios.

3. Results

We will now proceed to present the outcomes of our experiment, starting with the human evaluation and then moving on to the evaluation conducted by ChatGPT in the two scenarios we designed. Finally, we will discuss the obtained results, drawing connections to the three initial questions we posed.

Table 2 displays the outcomes of the human evaluation. The table presents the averages of human evaluators' responses, as well as the standard deviation among their ratings.

Table 2

Comparative Evaluation of Translations by Human Evaluators

Criteria	Google Translate		ChatGPT		DeepL		MSD Portuguese Translation	
	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD
Fluency	2.5	1.130	4	1.224	3.7	0.833	2.3	0.866
Appropriateness	3	1	4.22	0.971	3.8	0.600	2.7	0.971
Accuracy	3.4	1.103	4	0.866	3.33	1	2.7	1.092
Overall Evaluation	3	0.866	4	0.866	3.5	0.527	2.5	0.881

It is evident that the translation generated by ChatGPT received the highest overall rating, scoring 4 out of 5. Close behind is DeepL with a score of 3.5, followed by Google Translate with a score of 3.

The official translation from the MSD manual obtained the lowest rating of 2.5. The ratings were assessed on a scale of 1 to 5, indicating the relative performance of each translation.

Regarding the evaluation of ChatGPT, we tested two scenarios. In the first scenario, translations were comparatively evaluated, all included within the same prompt. To test the consistency of the results, the identical prompt was utilized five times, initiating a fresh chat for each attempt. Table 3 presents the mean responses from ChatGPT, along with the corresponding standard deviation among the responses.

Table 3

Comparative Evaluation of Translations by ChatGPT (Scenario 1)

Criteria	Google Translate		ChatGPT		DeepL		MSD Portuguese Translation	
	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD
Fluency	4.6	0.55	4	0	4	0	3	0
Appropriateness	4.4	0.55	4.4	0.55	3.8	0.447	3.4	0.347
Accuracy	4.6	0.55	4.6	0.55	3.8	0.836	3.8	0.447
Overall Evaluation	4.54	0.508	4.32	0.16	3.87	0.375	3.4	0.2909

With respect to the overall evaluation, the translation from Google Translate attained the highest score, averaging 4.54. Following closely is ChatGPT's translation, with an average rating of 4.32. In third place is the translation from DeepL, averaging at 3.87, while the translation from the MSD manual ranks last with an average of 3.4.

In the second scenario, each translation was evaluated individually in separate chats, five different times. Table 4 presents the mean results of the individual assessments for the translations, along with the corresponding standard deviation among the responses.

Table 4

Comparative Evaluation of Translations by ChatGPT (Scenario 2)

Criteria	Google Translate		ChatGPT		DeepL		MSD Portuguese Translation	
	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD
Fluency	4.6	0.547	4.4	0.547	4	0	4.4	0.547
Appropriateness	4.6	0.547	5	0	4.4	0.547	4	0
Accuracy	4.4	0.547	4.8	0.447	4.6	0.547	4	0
Overall Evaluation	4.6	0.435	4.7	0.279	4.33	0.335	4	0

While the individual evaluation scores are generally higher, the overall ranking remains largely unchanged. In the global assessment, ChatGPT achieved the highest score of 4.7. Close behind is the translation from Google Translate with a score of 4.6. In third place is the translation from DeepL, averaging at 4.33, while the translation from the MSD manual ranks last with an average of 4.

4. Discussion

Let us now revisit the initial questions.

To what extent does ChatGPT's evaluation of automatic translations of scientific texts align with human assessment?

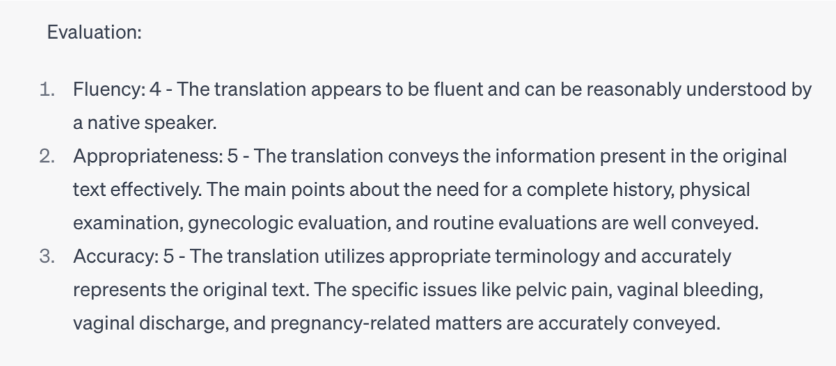
In terms of the global ranking, a consensus emerges from both the human evaluation and ChatGPT evaluation regarding the translation with the lowest score, which is the MSD manual translation. However, there is a notable discrepancy between the evaluations when it comes to the top-performing translations. In a collective evaluation (scenario 1), ChatGPT gives the highest rating to the translation by Google Translate. However, when evaluated individually (scenario 2), it assigns the highest rating to its own translation. Human evaluation aligns with ChatGPT's assessment in the second scenario, considering ChatGPT's translation as the best option. These findings suggest that ChatGPT’s translation is highly regarded, demonstrating its strong performance and quality according to both human evaluators and the self-evaluation conducted by ChatGPT.

Can ChatGPT be a reliable tool for researchers and science communicators in evaluating automatic translations of scientific texts?

When it comes to result consistency, the responses from ChatGPT demonstrate lower variation in both scenarios compared to the human evaluations. Specifically, focusing on the criterion of fluency, the average standard deviation was 0.137 for ChatGPT in the first scenario and 0.410 in the second scenario, and 1.013 for the human evaluation. Regarding the criterion of appropriateness, the average standard deviation was 0.523 for ChatGPT in the first scenario and of 0.273 in the second scenario, and 0.886 for the human evaluation. For the criterion of accuracy, the average standard deviation was 0.595 for ChatGPT in the first scenario and of 0,385 in the second scenario, and 0.993 for the human evaluation. Lastly, in terms of the overall evaluation, the average standard deviation was 0.333 for ChatGPT in the first scenario and 0.262 in the second scenario, and 0.785 for the human evaluation. These results highlight that ChatGPT's evaluations demonstrate greater consistency in terms of fluency, appropriateness, accuracy, and overall assessment when compared to the human evaluations.

What are the strengths and/or limitations of ChatGPT in evaluating the quality of automatic translations of scientific texts?

As observed from the reported results, the evaluations conducted by ChatGPT display some degree of variation and are not entirely aligned with the human evaluations. Hence, utilizing ChatGPT for formal translation assessments using this methodology may not be recommended. However, we posit that this approach can offer individual users, such as researchers and science communicators, a valuable tool to evaluate translations of their scientific texts. First and foremost, ChatGPT consistently provides explanations along with its evaluations (see Figure 1 for an example). These explanations can assist researchers and other science communicators in clarifying terminological or discursive choices, as well as simplifying scientific jargon [20] to make the texts more comprehensible for a wider audience.

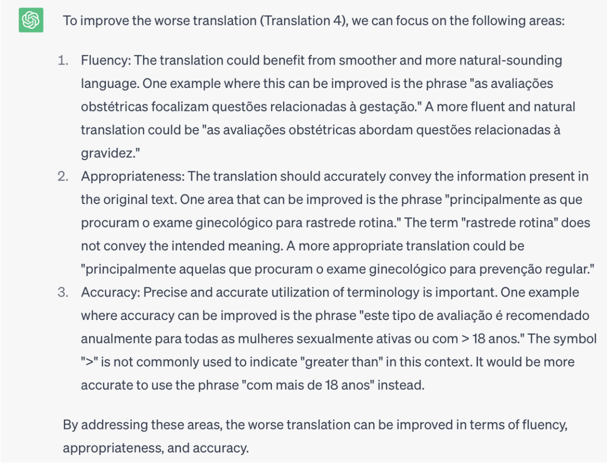


Evaluation:

1. Fluency: 4 - The translation appears to be fluent and can be reasonably understood by a native speaker.
2. Appropriateness: 5 - The translation conveys the information present in the original text effectively. The main points about the need for a complete history, physical examination, gynecologic evaluation, and routine evaluations are well conveyed.
3. Accuracy: 5 - The translation utilizes appropriate terminology and accurately represents the original text. The specific issues like pelvic pain, vaginal bleeding, vaginal discharge, and pregnancy-related matters are accurately conveyed.

Figure 1: ChatGPT'’s evaluation, accompanied by explanations for each score.

Another notable advantage of employing ChatGPT for assessing translations of scientific texts is the opportunity for follow-up inquiries. For instance, during one of the tests, we posed a question to ChatGPT regarding how to improve the lowest score translation. In response, ChatGPT provided concrete suggestions and examples of potential enhancements for each evaluation criterion. This feedback is exemplified in Figure 2.



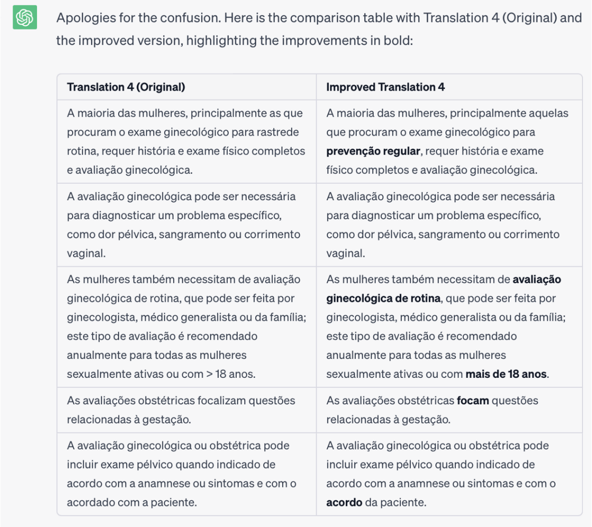
To improve the worse translation (Translation 4), we can focus on the following areas:

1. Fluency: The translation could benefit from smoother and more natural-sounding language. One example where this can be improved is the phrase "as avaliações obstétricas focalizam questões relacionadas à gestação." A more fluent and natural translation could be "as avaliações obstétricas abordam questões relacionadas à gravidez."
2. Appropriateness: The translation should accurately convey the information present in the original text. One area that can be improved is the phrase "principalmente as que procuram o exame ginecológico para rastrede rotina." The term "rastrede rotina" does not convey the intended meaning. A more appropriate translation could be "principalmente aquelas que procuram o exame ginecológico para prevenção regular."
3. Accuracy: Precise and accurate utilization of terminology is important. One example where accuracy can be improved is the phrase "este tipo de avaliação é recomendado anualmente para todas as mulheres sexualmente ativas ou com > 18 anos." The symbol ">" is not commonly used to indicate "greater than" in this context. It would be more accurate to use the phrase "com mais de 18 anos" instead.

By addressing these areas, the worse translation can be improved in terms of fluency, appropriateness, and accuracy.

Figure 2: ChatGPT''s response, providing concrete suggestions for improving the lowest score translation.

Following this response, we proceeded to ask ChatGPT to enhance the translation according to its own improvement suggestions. We requested ChatGPT to present both the original and improved translations in a table, with the improvements highlighted in bold. The response obtained from ChatGPT to this instruction is depicted in Figure 3.



Apologies for the confusion. Here is the comparison table with Translation 4 (Original) and the improved version, highlighting the improvements in bold:

Translation 4 (Original)	Improved Translation 4
A maioria das mulheres, principalmente as que procuram o exame ginecológico para rastrede rotina, requer história e exame físico completos e avaliação ginecológica.	A maioria das mulheres, principalmente aquelas que procuram o exame ginecológico para **prevenção regular**, requer história e exame físico completos e avaliação ginecológica.
A avaliação ginecológica pode ser necessária para diagnosticar um problema específico, como dor pélvica, sangramento ou corrimento vaginal.	A avaliação ginecológica pode ser necessária para diagnosticar um problema específico, como dor pélvica, sangramento ou corrimento vaginal.
As mulheres também necessitam de avaliação ginecológica de rotina, que pode ser feita por ginecologista, médico generalista ou da família; este tipo de avaliação é recomendado anualmente para todas as mulheres sexualmente ativas ou com > 18 anos.	As mulheres também necessitam de **avaliação ginecológica de rotina**, que pode ser feita por ginecologista, médico generalista ou da família; este tipo de avaliação é recomendado anualmente para todas as mulheres sexualmente ativas ou com **mais de 18 anos**.
As avaliações obstétricas focalizam questões relacionadas à gestação.	As avaliações obstétricas **focam** questões relacionadas à gestação.
A avaliação ginecológica ou obstétrica pode incluir exame pélvico quando indicado de acordo com a anamnese ou sintomas e com o acordado com a paciente.	A avaliação ginecológica ou obstétrica pode incluir exame pélvico quando indicado de acordo com a anamnese ou sintomas e com o **acordo** da paciente.

Figure 1: Comparison table showcasing the original and improved translations, with improvements highlighted in bold.

We have showcased only two examples here, but the capability to engage in follow-up inquiries, receive suggestions, and improve translations using ChatGPT makes it an immensely useful tool for researchers and science communicators who aim to explore multilingual environments for knowledge sharing and production.

5. Discussion

In this study, we conducted evaluations of translations of an excerpt of a scientific text from English to Portuguese using both human assessments and ChatGPT. The results shed light on several important findings and provide insights into the strengths and limitations of using ChatGPT for translation evaluations.

Firstly, in terms of overall rankings, there was a consensus between human evaluations and ChatGPT assessments regarding the translation with the lowest score, which was the official translation from the MSD manual. However, variations were observed in the evaluations of the top-ranked translations. In scenario 1, when evaluated collectively, ChatGPT gives the highest rating to Google Translate's translation. However, in scenario 2, during individual evaluation, it assigns the highest rating to its own translation. Human evaluation in the second scenario aligns with ChatGPT's assessment, affirming ChatGPT's translation as the preferred choice.

The consistency of results was another aspect explored in this study. It was found that ChatGPT exhibited less variation in its evaluations compared to the human assessments. This consistency was evident across multiple evaluation criteria, including fluency, appropriateness, accuracy, and overall assessment. The ability of ChatGPT to provide more consistent evaluations highlights its potential as a valuable tool for individual users, such as researchers and science communicators, to assess translations of their scientific texts.

Furthermore, the interactive nature of ChatGPT proved to be advantageous in translation evaluations. By allowing follow-up questions and receiving suggestions from ChatGPT, users can actively engage in improving translations. This dynamic interaction empowers users to refine their translations and enhance the quality of their output. The feedback and improvement suggestions provided by ChatGPT serve as valuable guidance for users seeking to produce high-quality translations in the scientific domain.

However, it is important to note that while ChatGPT shows promise as a tool for individual translation evaluations, this methodology may not be suitable for formal and comprehensive assessments. The variations observed in the evaluations and the disparities between ChatGPT and human assessments indicate that caution should be exercised when relying solely on ChatGPT for translation evaluations in formal settings.

In our future research, we envision applying this methodology to a more diverse and extensive collection of texts, encompassing various scientific disciplines and covering a broader range of language pairs. Furthermore, we are eager to explore the potential applications of ChatGPT beyond translation evaluations in the realm of scientific texts. For instance, we plan to investigate its usability in tasks such as summarization, terminology extraction, and cross-lingual information retrieval, to facilitate and enhance the communication and dissemination of scientific knowledge across language barriers.

6. Acknowledgements

This work was carried out within the scope of the "PortLinguE" project (PTDC/LLT-LIG/31113/2017) financed by FEDER under Portugal 2020 and by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT,I.P.).

7. References

[1] Ermakova, L., Bellot, P., Kamps, J., Nurbakova, D., Ovchinnikova, I., SanJuan, E., Mathurin, E., Hannachi, R., Huet, S., Araujo, S.: "Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts." In: LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2022).

[2] Aslerasouli, P., & Abbasian, G. R.: "Comparison of Google Online Translation and Human Translation with Regard to Soft vs. Hard Science Texts." Journal of Applied Linguistics and Language Research 2(3), 169-184 (2015).

[3] Han, L., Jones, G., & Smeaton, A.F.: "Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods." ArXiv, abs/2105.03311 (2021).

[4] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J.: "Bleu: a method for automatic evaluation of machine translation." In: Proceedings of ACL 2002.

[5] Lin, C.-Y., & Hovy, E. H.: "Automatic evaluation of summaries using n-gram co-occurrence statistics." In: Proceedings of NAACL 2003.

[6] Banerjee, S., & Lavie, A.: "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments." In: Proceedings of the ACL 2005.

[7] Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J.: "Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 858-867."

[8] Lembersky, G., Ordan, N., & Wintner, S.: "Language Models for Machine Translation: Original vs. Translated Texts." Computational Linguistics 2012; 38(4), 799-825. doi: https://doi.org/10.1162/COLI_a_00111

[9] Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., & Chao, L. S.: "Learning deep transformer models for machine translation." arXiv preprint arXiv:1906.01787 (2019). [9]

[10] Kocmi, T., & Federmann, C.: "Large language models are state-of-the-art evaluators of translation quality." arXiv preprint arXiv:2302.14520 (2023).

[11]Radford, A., & Narasimhan, K.: "Improving Language Understanding by Generative Pre-Training" (2018).

[12] Thorp, H.: "ChatGPT is fun, but not an author." Science 379(6630), 313 (2023).

[13] MSD Manuals. "General Gynecologic Evaluation." Retrieved from https://www.msdmanuals.com/professional/gynecology-and-obstetrics/approach-to-the-gynecologic-patient/general-gynecologic-evaluation

[14] Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z.: "Is ChatGPT a good translator? A preliminary study." arXiv preprint arXiv:2301.08745 (2023).

[15] Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., & Tao, D.: "Towards making the most of ChatGPT for machine translation." arXiv preprint arXiv:2303.13780 (2023).

[16] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H.: "How good are GPT models at machine translation? A comprehensive evaluation." arXiv preprint arXiv:2302.09210 (2023).

[17] White, J. S., O'Connell, T., & O'Mara, F.: "The arpa mt evaluation methodologies: Evolution, lessons, and future approaches." In: Proceeding of AMTA (1994).

[18] Church, K., & Hovy, E.: "Good applications for crummy machine translation." In: Proceedings of the Natural Language Processing Systems Evaluation Workshop (1991).

[19] Blanchon, H., & Boitet, C.: "Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche." Trait. Autom. des Langues 48(1), 33-65 (2007).

[20] Liana Ermakova, Eric SanJuan, Stéphane Huet, Olivier Augereau, Hosein Azarbonyad, and Jaap Kamps. 2023. Overview of SimpleText - CLEF-2023 track on Automatic Simplification of Scientific Texts. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, Nicola Ferro (Eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)