ACCEPTANCE CODE: CO190

AI-POWERED PRECISION:
REVOLUTIONIZING COMPARATIVE REVIEW IN CLINICAL OUTCOME ASSESSMENTS

LIONBRIDGE

Authors: Stephanie Casale, Elisabet Sas Olesa, Juliana Coghi Jimenez, Melinda Johnson, Kathryn Nolte

INTRODUCTION

Linguistic Validation (LV) is the process by which Clinical Outcomes Assessments (COAs) are localized and validated for accurate and consistent data collection in target locales.

The process is lengthy and complex, by design, to ensure the highest quality and most thorough translations, but this complexity comes at a cost. In order to reduce the monetary and time burden of this process, this study's aim is to find ways to automate steps leveraging AI in the process that will reduce turnaround times and costs, while maintaining the high standards for which the LV process is designed. We focused on the Comparative Review (CR) step within the process.

Comparative Review is a key quality assurance step in the LV process, which compares source text to back translated text to determine conceptual equivalence. Because it is an intermediary step, the prior and subsequent steps are performed by trained, experienced linguists. This makes the CR step a prime candidate for automation, as it minimizes the risk of errors occurring without detection before finalization.

Our research aimed to develop a prompt that upheld, at a minimum, the existing quality of our current human suppliers for comparative review.

METHODOLOGY

We first spent time developing a prompt that produced the expected outcome of both a comparative review result and a comparative review comment, which gave further detail on the results. Comparative Review results would be divided into three categories:

Identical: Indicates the source text and back translation were exactly the same in every way, including capitalization and punctuation.

Equivalent: Indicates that while there may be differences in wording, sentence structure, or other details, the meaning of the segments remains conceptually equivalent. It would be understood by the reader to convey the same information.

Needs Review: Indicates that something in the two segments renders them conceptually inequivalent and could be misunderstood by a reader to mean something was not intended by the source text.

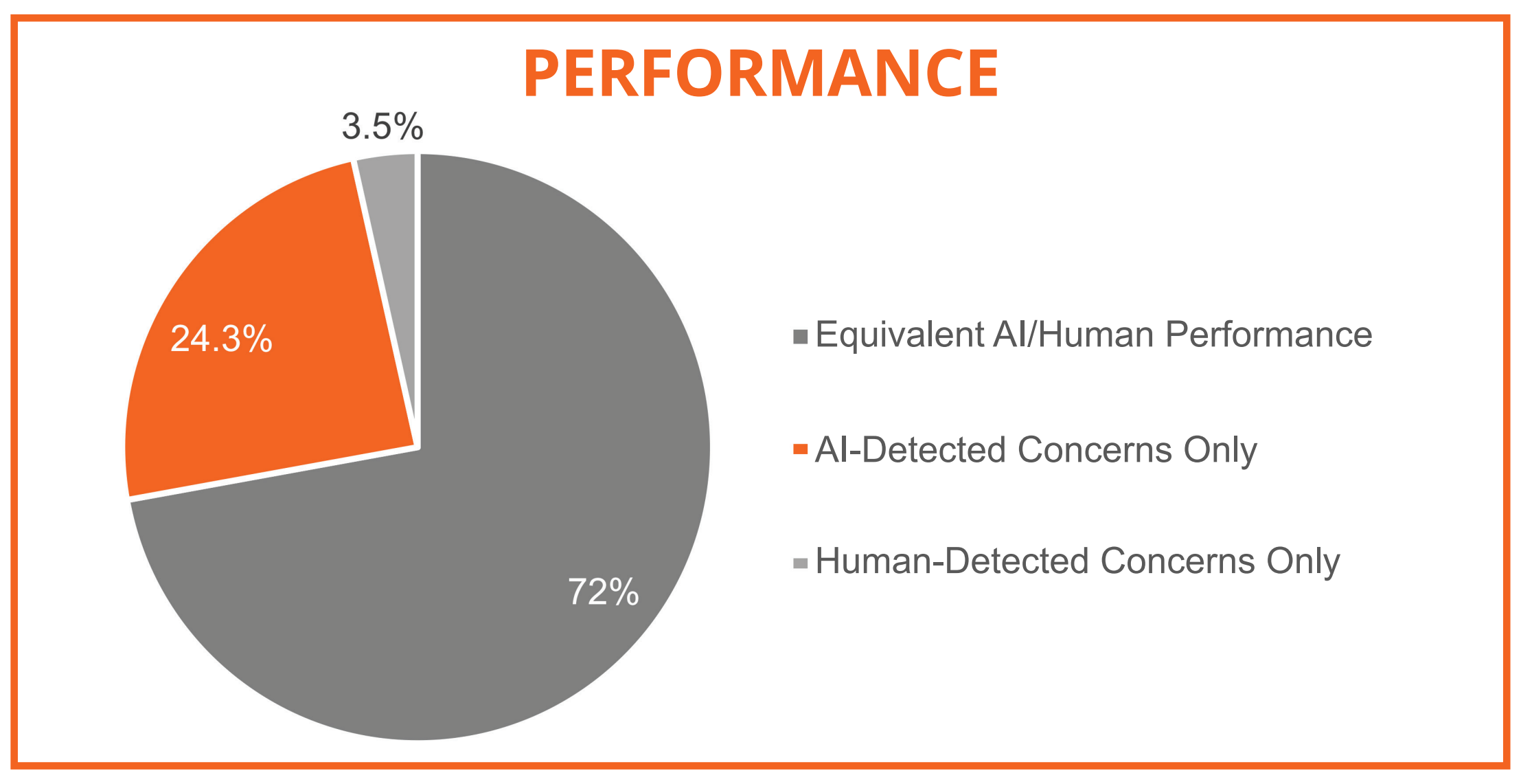
The prompt was then designed to produce a comparative review comment for any non-identical result. These comments should include an explanation of any conceptual differences between the two segments, including an elaboration on possible misinterpretations by a lay reader. The prompt was asked to ignore any punctuation and capitalization differences unless they were directly related to meaning and understanding, as well as to ignore any additional text not related to the meaning of the source text (i.e., formatting tags, etc).

Leveraging a sample size of ~1000 words, we conducted a pass/fail analysis on three sets of CR outputs in English, one set generated by a secure AI engine (leveraging Chat GPT-4o technology), and two sets generated by humans with 5+ years of CR experience in the COA industry.

A Rater with 15 years of CR experience in the COA industry then evaluated the 3 outputs, determining if they passed or failed task-specific expectations on each item ("segment").

RESULTS

The initial results are promising, with clear, concise descriptions of original assessment and back translation discrepancies at an overall preliminary accuracy rate of 96.4% by the AI engine. The average human score vs. the AI score can be seen in the chart below:



Of the total number of segments analyzed, 72% of findings were consistent as content that Needs Review by the AI engine and humans. Additionally, 3.5% of Needs Review findings were only flagged by the humans. AI detected 24.3% of Needs Review concerns that were not detected by the humans. All of these results were vetted by the Rater as true findings of potential issues.

Additional Notable Percentages:

INHERENT AI RISK	0.17%	This number included segments that might have been flagged by a human due to the non-Latin alphabet in the forward translation. They would not have been noted by the AI prompt.
INCONSISTENT RESPONSES	1.26%	During our review, the AI would occasionally give different responses for the same set of segments. This came to just over 1% of the total data.

CONCLUSIONS

Overall, this initial study showed that AI could not only perform at the human level of an expert with 5+ years of experience, but that it actually outperformed those humans within this small sample. **Due to this, AI has the potential to save significant time and costs in the Linguistic Validation process without reducing the quality standards defined by the industry.**

FURTHER RECOMMENDATIONS

Further study should be done to expand the data set and number of Raters, as well as the inclusion of a Proof of Concept that extends to the resolution steps. This study will examine effects of using this output with linguists. Additionally, further refinement of the prompt could help eliminate some of the inconsistencies and risks.

Disclaimer: AI was used to generate the poster's title and per the study methodology disclosed in the body of this poster.
The authors did not leverage AI in any other manner in the study nor the creation of this poster.

ACCEPTANCE CODE: CO149

TRANSFORMING CONCEPT ELABORATION IN COA LOCALIZATION: A GENAI-BASED APPROACH

LIONBRIDGE

Authors: Elisabet Sas Olesa, Stephanie Casale, Melinda Johnson, Kathryn Nolte, Fernando Agüero, Karolina Elizondo, Anca Bodzer, Laura O’Gara, Stefanie Costa

INTRODUCTION

Concept Elaboration (CE) is the process of creating clear explanations for concepts in an original Clinical Outcome Assessment measure (COA) to ensure the translation captures the conceptual meaning of the items.

This step is performed to allow a meaningful translation that’s conceptually equivalent to the source text and culturally and linguistically appropriate in the target country to facilitate pooling and comparison of data (Acquadro et al., 2017). The risk of not performing this preparatory step is misinterpretation (and thus potential mistranslation) of items or concepts present in the COA measure (Wild et al., 2005). This may render the collected patient results inaccurate.

CE is a thorough analysis of the source file, usually performed by an expert, which aims to:

- Understand and clarify the overall conceptual meaning of each item, term, and statement
- Identify and define key terms
- Identify potential semantic differences between the source and target languages
- Identify potential cultural differences between the source and target countries
- Identify segments that must or must not be localized

COA localization, particularly through Linguistic Validation, has traditionally been a discipline highly reliant on human activities (Williams, 2024). However, these authors recognized potential for progress that AI capabilities offer. We set out to streamline the CE process by automating it further with AI.

METHODOLOGY

We conducted quantitative and qualitative analyses on two sets of CE outputs in English.

One set was generated by a secure AI engine (leveraging Chat GPT-40 technology). One set was generated by a human for various types of COA measures. To mimic a real-life project, which has a variety of content in terms of length and complexity level, we selected a sample of ~9500 source words, divided as follows:

- 4 Clinician-Reported Outcome measures (ClinROs)
- 4 Observer-Reported Outcome measures (ObsROs)
- 4 Performance Outcome measures (PerfOs)
- 4 Patient-Reported Outcome measures (PROs)

We customized an AI prompt within the secured AI engine. It included a summary of the latest COA industry standards and established practices for preparation of Concept Elaborations. We proceeded to run AI on the sample content of each of the above mentioned measures. Next, we inserted this AI output into our usual Concept Elaboration report template so they’d be structurally comparable to any human-performed task.

As a parallel Control, we selected 2 of our linguists with different levels of expertise. We asked them to perform the Concept Elaboration task from scratch for all 16 files, replicating a regular translation request.

Then we chose a qualified **Rater** and requested they perform an in-depth analysis of two sets of CEs by rating under the following parameters, the **5 Cs**:

CHOICE
Defined as the appropriateness of selection of key terms within the context of the scale

CLARITY
Defined as the quality of the elaboration; how easily the concept elaboration is understood by the linguist working on the task

COMPLIANCE
Defined as the extent to which the concept elaboration adheres to the various instructions for the task

CITATION
Defined as the accuracy and appropriateness of references and citations used to support the concept definition process

CONSISTENCY
Defined as the uniformity and coherence of the concept elaboration throughout the document, especially where terminology repeats or refers to the same Core concept within the assessment. Also, overall consistency of the style of the CE comments

Each category was rated with a three-point scale with corresponding anchors, as described below:

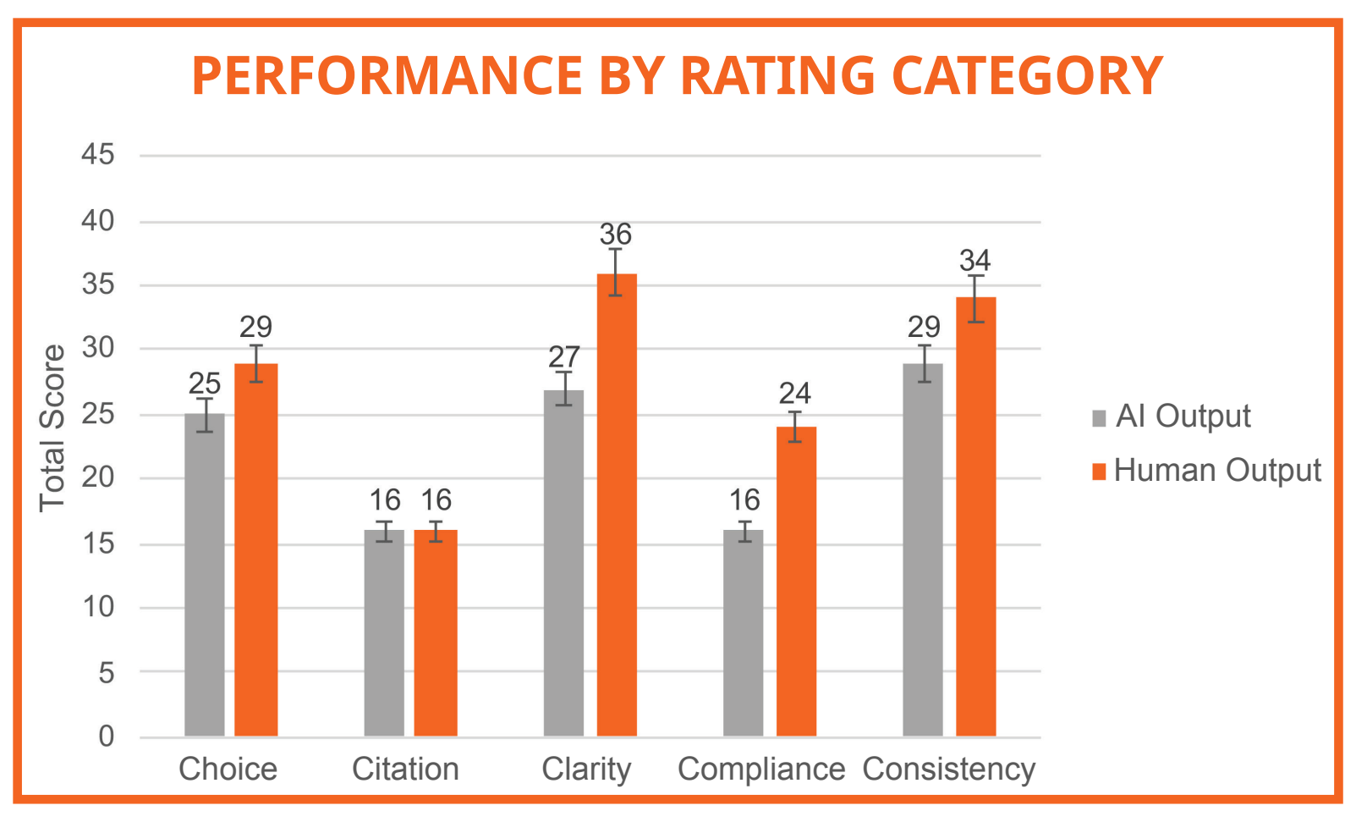
VALUE	DESCRIPTION	MEANING
1	Poor	The AI engine or Human failed and did not meet expectations in that area.
2	Average	The output was adequate, but may need improvement (i.e., prompt refinement or further training).
3	Excellent	Ideal performance by AI engine or Human.

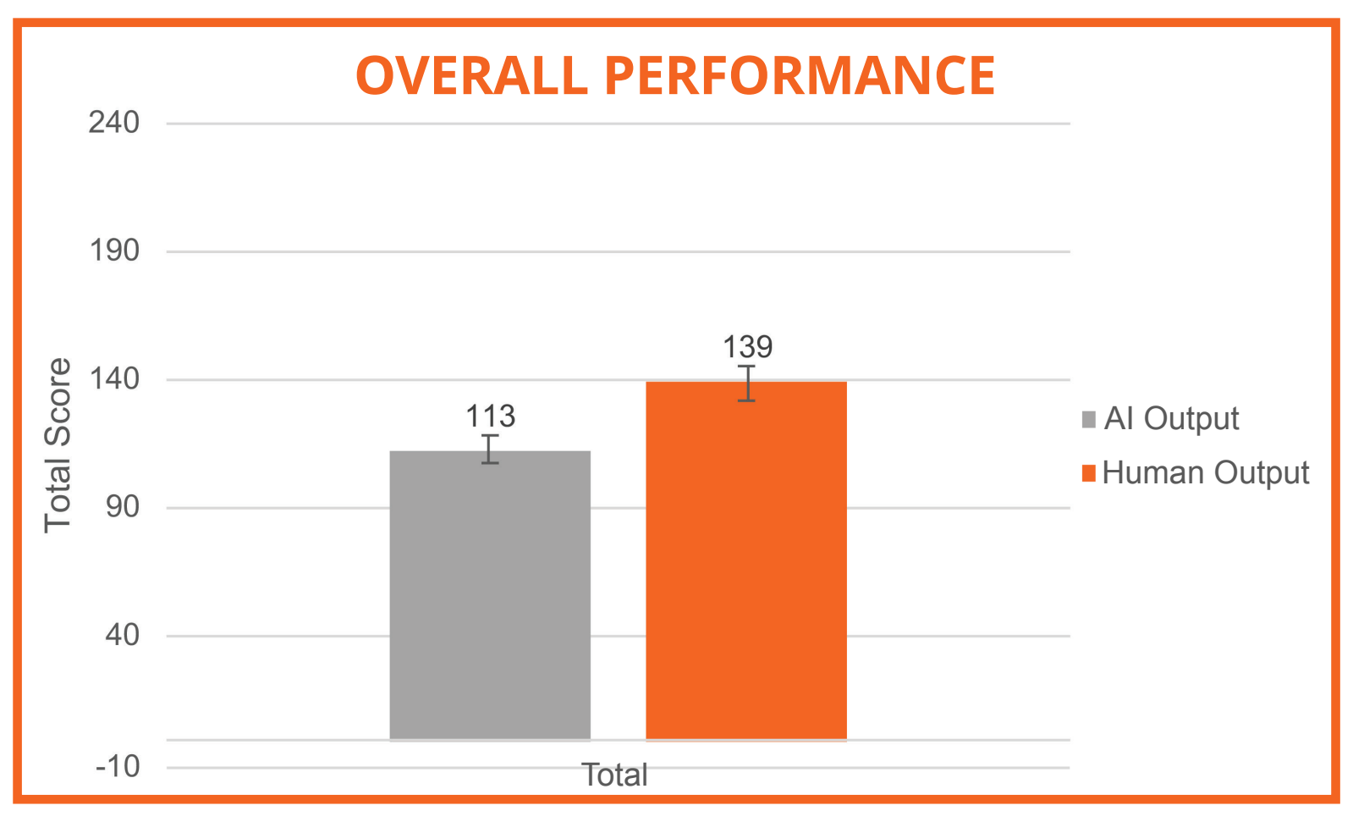
In addition to providing numeric values for the above categories for each COA, we also requested that our **Rater** provide their “Overall Impression” perception of all content to obtain a more subjective point of view. Using these categories and rating guidelines, we systematically compared and evaluated various aspects of CE quality provided by an AI Engine versus quality provided by one or more “human” linguists. This structured approach ensured the analysis was as objective and comprehensive as possible. It’s important to highlight that this comparison was blinded. Because we wanted to avoid bias stemming from preconceived notions about AI, we never disclosed our AI usage to create CEs to any participants.

RESULTS

Overall Performance

Within each of the **5 Cs**, the AI and Human could score 16 to 48 points.



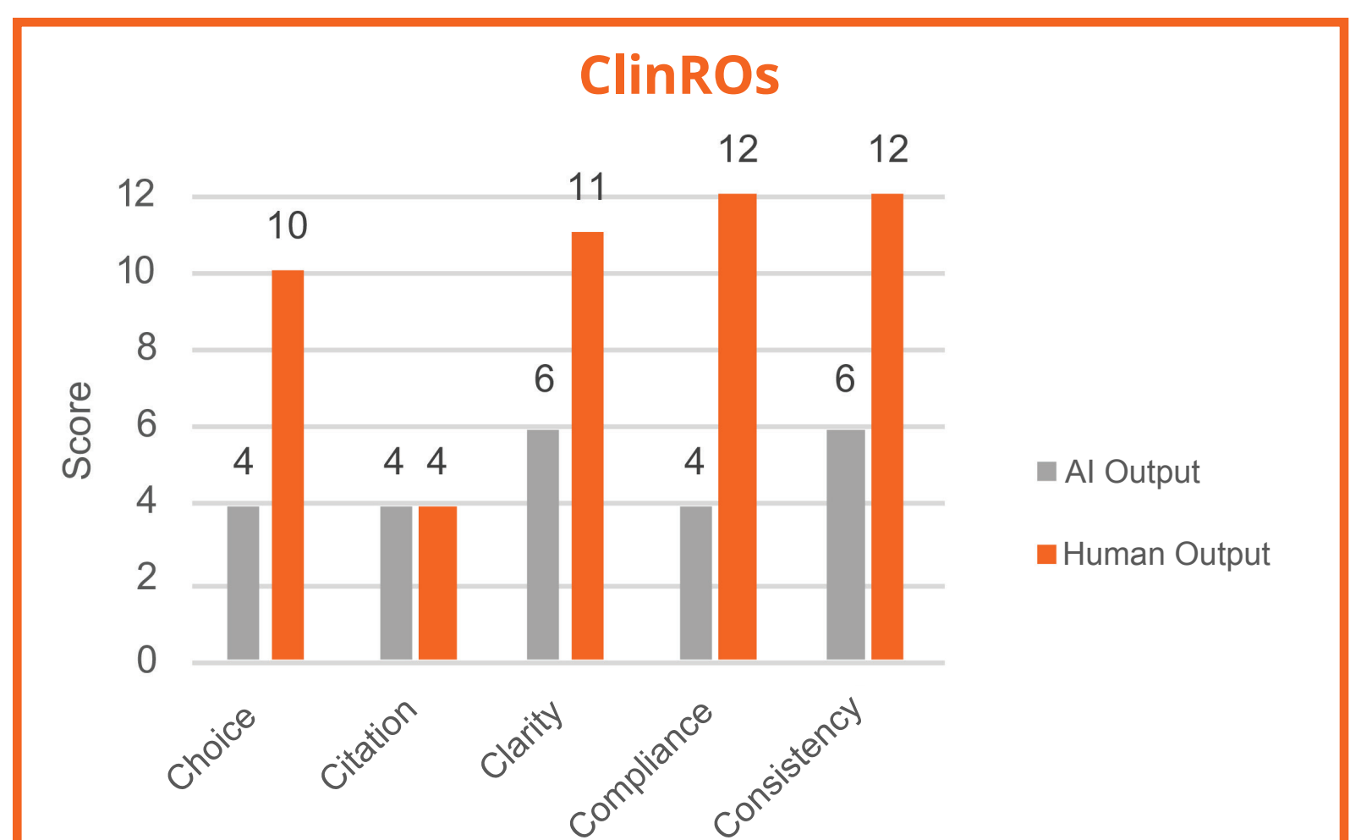


Scores for both outputs ranged from 16 to 36 per category, with the most comparable results between Choice, Citation, and Consistency. Clarity and Compliance showed the most significant difference between the AI and Human output scores. The results of our study indicate that, even though the Human did consistently better, their performance was not perfect. There was no total prevalence over the AI engine, which performed comparatively well in some categories.

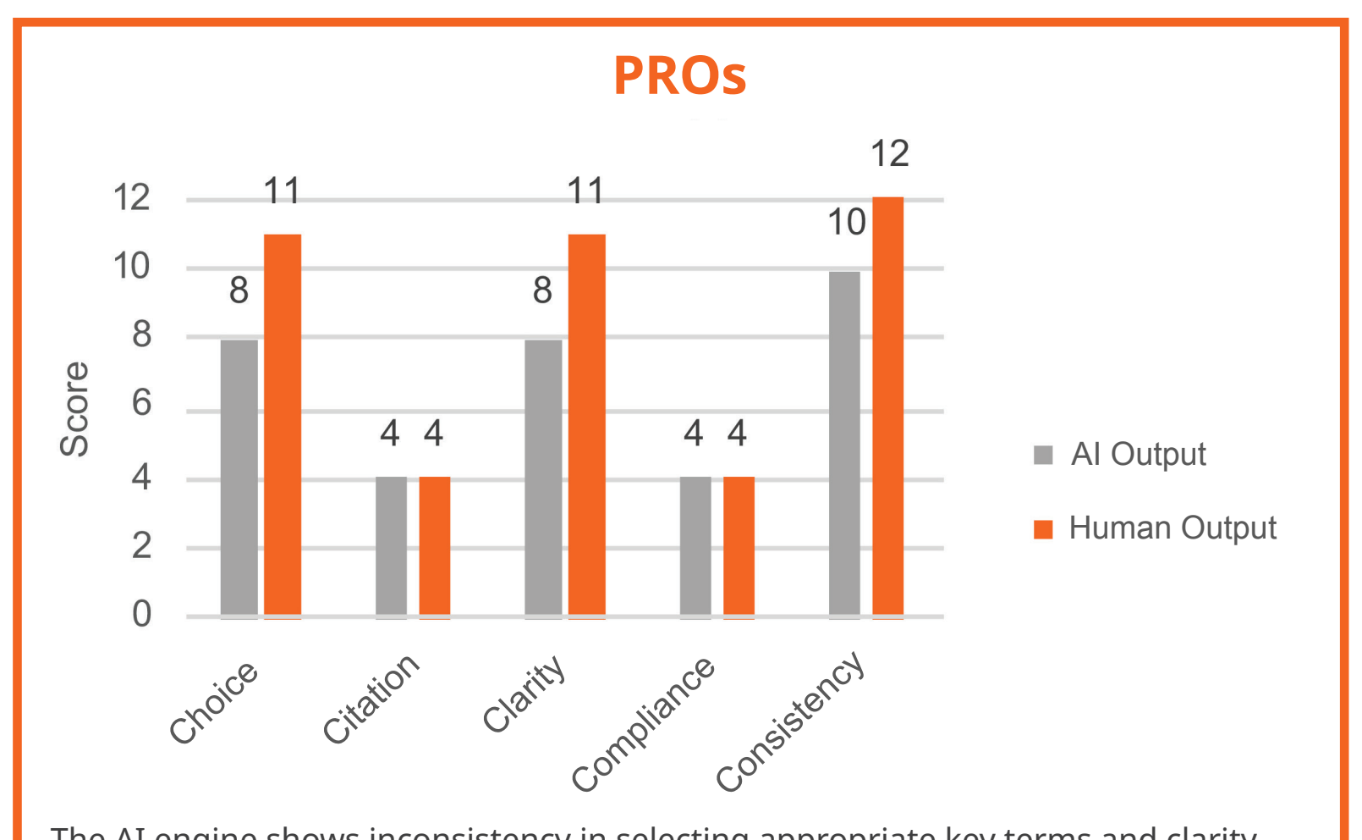
Across the **5 Cs**, the AI and Human could score 80 to 240 points. Our results show significant room for improvement in Human performance. This can be considered an excellent opportunity for responsible stakeholders to deliver focused training and guidance, as well as constructive feedback to experts performing the CE task. Alternatively, AI output doesn’t fall much further behind Human performance, but it may need further enhancement by refining prompt engineering to ensure optimal definition across the **5 Cs**.

Regarding overall performance impression, Rater highlighted that on multiple occasions, the AI engine wasn’t extracting and elaborating all necessary key terms (this may be due to data analysis limitations). It also became apparent that its Citation capability is unreliable because the engine lacks capacity to reference specific sources. This is essential to ensure key term definitions belong to the specific therapeutic domain of the COA in question.

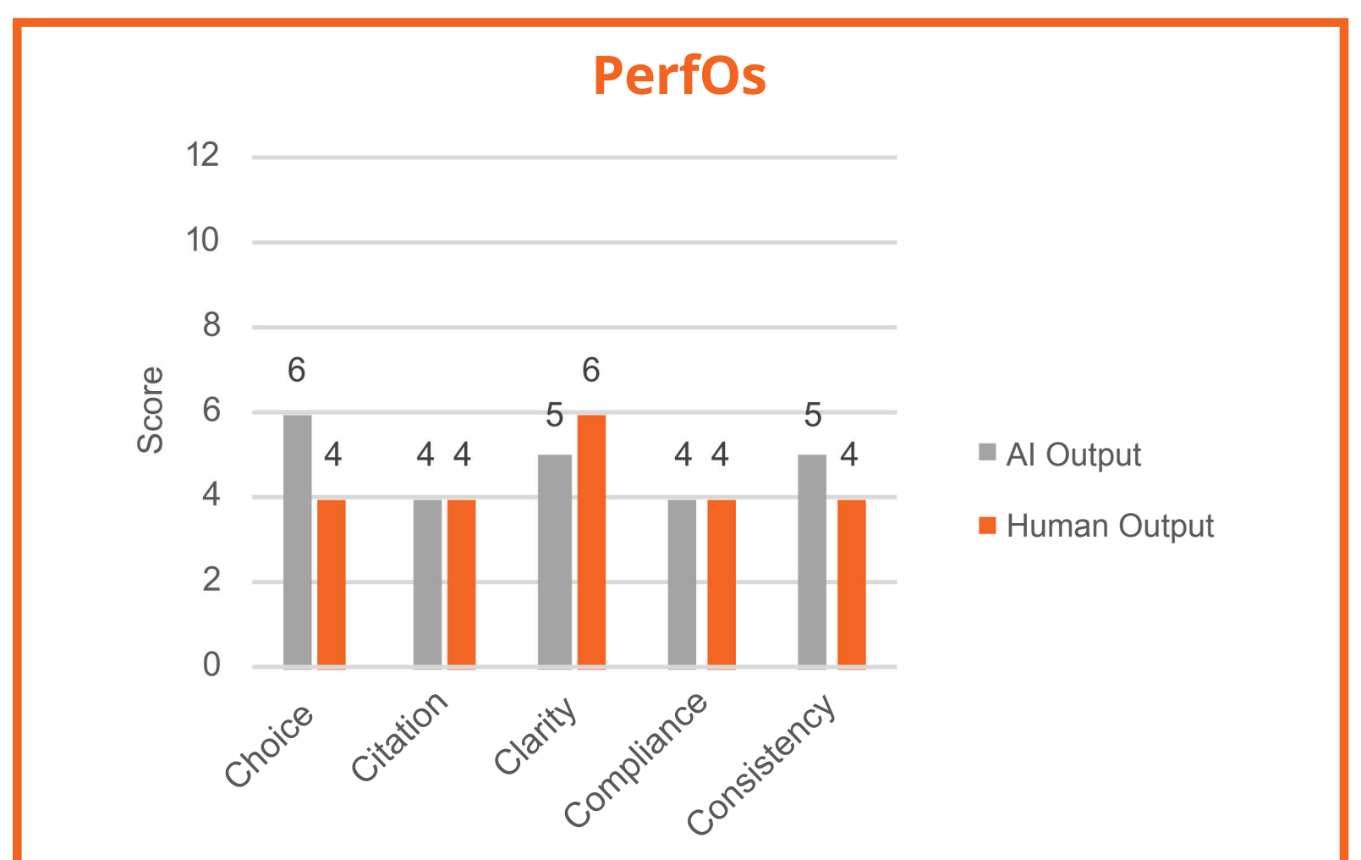
Performance by Assessment-Type



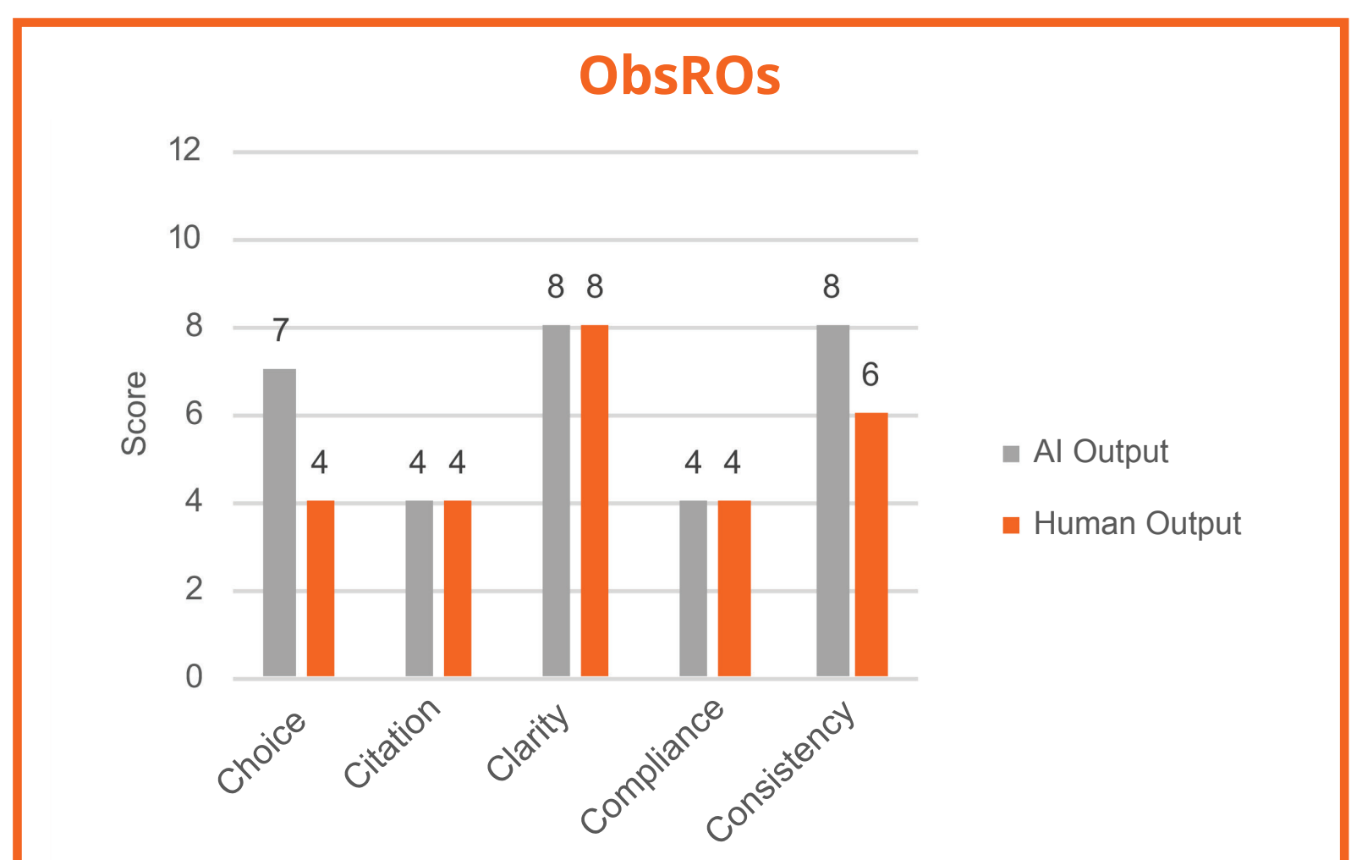
Based on the analysis of this COA type alone, the Human expert demonstrates a higher overall performance in the Concept Elaboration task than the AI engine. The Human outputs are more appropriate, clear, compliant, and consistent overall, making them more useful for localizing the scales into target languages. However, it’s vital to note that both parties need to improve in accuracy and appropriateness of Citations, which is the category with the lowest score.



The AI engine shows inconsistency in selecting appropriate key terms and clarity, with ratings ranging from “Excellent” to “Average” to “Poor.” The Human output can be considered generally better for this type of content, with mostly “Excellent” ratings in Choice, Clarity, and Consistency, indicating a strong ability to select relevant key terms and maintain internal coherence. However, both outcomes require significant improvement in Compliance and Citation (consistently rated as “Poor,” highlighting the need for better adherence to guidelines and more reliable research practices).



In this COA type, the AI engine shows variability in selecting key terms and consistency, with ratings ranging from “Poor” to “Average,” indicating occasional appropriateness but, more importantly, a general need for improvement or external review. The Human output is consistently rated “Poor” across most categories, highlighting significant issues in selecting relevant key terms and maintaining clarity and consistency. Both parties require substantial improvement in Compliance and Citation (both consistently rated “Poor”).



In the CE of this type of COA, we notice that generally, the AI engine clearly performs better than the Human expert in parameters such as the Choice of key terminology. The remaining categories, Clarity, Compliance, and Citation, are consistently awarded the same ratings, either *Average* or *Poor*. Neither of the subjects did well on Consistency.

CONCLUSIONS

Our goal was to test AI’s capability to improve efficiency and save overall project execution time and costs. While AI is undeniably faster than its human counterpart at the CE task, the results above suggest its output must be balanced via a “human-in-the-loop” to ensure completeness and final quality. In future studies, we would like to ascertain the exact time and resource savings when we use AI to perform the entire CE from scratch, then have an expert review and edit output. However, it should be noted a much larger sample of CE and quality evaluation requirements will be needed to make this assessment. We completely agree with Williams that AI integration into the Linguistic Validation process should be seen as an opportunity for progress, not a threat (Williams, 2024).

Therefore, adopting AI as part of the CE process proves to be a valuable tool for augmenting quality, increasing efficiency, and delivering more comprehensive results. It represents a significant step forward in merging AI capabilities within existing Linguistic Validation processes, paving the way for more robust and reliable outcomes in COA localization.

FUTURE RECOMMENDATIONS

- AI shows great potential in the identification of the **Therapeutic Area and Disease/Condition** being evaluated in the questionnaire.
- Interdependency of **Concept Elaboration** and **Source Analysis and Translatability Assessment** - potential to be merged into one single activity, further condensing the steps needed to perform the full COA localization process.

Citations:
Acquadro, C., Patrick, D. L., Eremenco, S., Martin, M. L., Kuliś, D., Correia, H. and Conway, K. “Emerging good practices for Translatability Assessment (TA) of Patient-Reported Outcome (PRO) measures.,” Journal of Patient-Reported Outcomes, vol. 2, no. 8, 2017.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health*, 8(2), 94–104. https://doi.org/10.1111/j.1524-4733.2005.04054.x

Williams, H. (2024, March 1). *Harmonising linguistic validation with AI: Precision, efficiency, and the human touch in patient-reported outcome translation.* Medical Writing. https://journal.emwa.org/translation/harmonising-linguistic-validation-with-ai-precision-efficiency-and-the-human-touch-in-patient-reported-outcome-translation/

Disclaimer: AI was used to generate the poster’s title and per the study methodology disclosed in the body of this poster. The authors did not leverage AI in any other manner in the study nor the creation of this poster.