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.

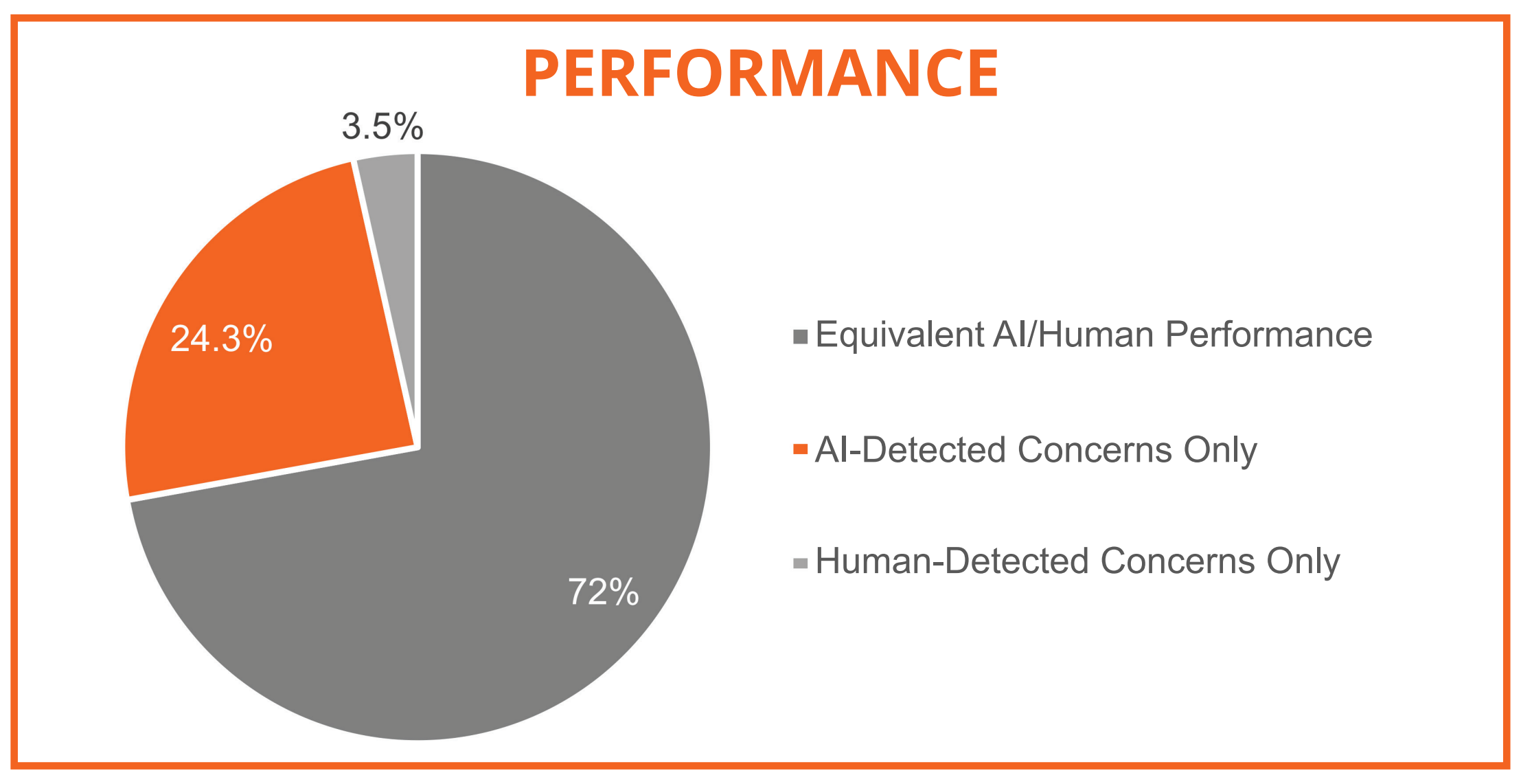
The prompt was then designed to produce a comparative review comment for any non-identical result. These comments should include an explanation of any conceptual differences between the two segments, including an elaboration on possible misinterpretations by a lay reader. The prompt was asked to ignore 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-40 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.