The Future of Linguistic Validation is Now:

Enhancing Reconciliation Process Compliance with GenAI

LIONBRIDGE | Pearson

Presented By Lionbridge: Elisabet Sas Olesa; Karolina Elizondo; Kathryn Nolte; Nathalie Azuaje; Melinda Johnson

INTRODUCTION

This study explores how Generative Artificial Intelligence (GenAI) could support the Linguistic Validation (LV) process, particularly focusing on its role in enhancing the Dual Translation Reconciliation step's efficiency, the stage where two or more independent forward translations are merged into one (Koller et al., 2012).

The primary objective of our investigation was evaluating GenAI's capacity to detect non-compliances within Reconciliations produced through conventional methods.

METHODS

We conducted a practical analysis using a PerfO sample (1,000-2,000 words) that included a mix of APAC, long-tail, and regional language variants. Based on these parameters, we selected various files from a previously completed localization of the Clinical Evaluation of Language Fundamentals®-Fifth Edition (CELF-5), a flexible system of individually administered tests. These tests are commonly used to assist clinicians in accurately diagnosing a language disorder in children and adolescents (NCS Pearson, Inc., 2013). The selected target languages for analysis were Spanish (Argentina), Spanish (Spain), French (France), Armenian (Armenia), Japanese (Japan), and Traditional Chinese (Taiwan).

PROMPT DESIGN EVOLUTION

The initial prompt was designed to operate alongside the traditional Reconciliation output by comparing Translation A against Translation B, thus generating a binary Pass-Fail outcome of the Reconciliator's final decision.

However, significant limitations were associated with this approach, particularly when applied to languages other than English and in highly specialized contexts — such as those defined by a specific therapeutic area or disease/condition commonly assessed in Quality-of-Life instruments. In response to these findings, we applied GenAI to evaluate the Reconciliation outcome, thus supporting our QA workflow.

The updated prompt was then integrated into *Aurora Clinical Outcomes*, Lionbridge's proprietary platform for end-to-end Linguistic Validation. Using GenAI to support internal decision-making also provides insight into the Reconciliator's thought process. It ensures the task was performed to the necessary standards and industry requirements. Such requirements are reflected in the GOALS OF RECONCILIATION and highlighted in the guidance provided to the Reconciliator in the task instructions:

1. **Conceptual Equivalence to the Original Measure**
2. **Cultural Adaptation**
3. **Accessibility for the Intended Study Population/Audience**
4. **Detection of Bias Trends**

RESULTS

Positive Outcomes

Analysis of GenAI outputs by Language Experts revealed three key areas of support:

- **Invalid or Incomplete Justifications:** The GenAI consistently flagged comments lacking sufficient linguistic or conceptual rationale (e.g., "Translation A is better" or "Translation B is preferred"). It also detected unresolved queries or unanswered questions. This prompted recommendations to refine the prompt, ensuring such cases are escalated for follow-up.
- **Missing Justifications:** All experts noted GenAI reliably identified absent, incomplete or unclear justifications, significantly accelerating QA review.
- **Efficiency:** The GenAI demonstrates high processing speed, analyzing up to 300 text segments in seconds.

This pre-screening allows Language Experts to filter results quickly and decide if a file is ready for the next step in the LV process.

Areas of Improvement

Despite its potential, GenAI's inherently unpredictable nature also introduced certain limitations during prompt execution:

- **Inconsistent Behavior:** GenAI results exhibit inconsistent behavior when evaluating identical comments across different segments, even within the same file. This undermines GenAI output's reliability as a standalone solution.
- **Contextual/Meta-reference Challenges:** The frequent use of vague comments and references like "See comment above" poses a serious challenge to GenAI in the current setup. GenAI cannot find the previous comment referenced. Experts suggest using concise and explicit references and repeated explanations across segments, as part of the expected deliverables from the Reconciliator.
- **Lack of Standardization of Viable Reconciliation Comments:** Our experts agree there needs to be a more precise, explicit definition and alignment between all parties on what constitutes acceptable reasoning and detailed justification.

CONCLUSION

This study demonstrates the significant potential of GenAI as a Quality Assurance support tool in identifying gaps, non-compliances, and missing justifications.

GenAI enables faster detection of rework needs and allows the Localization Team to deliver more targeted feedback to stakeholders, thus equipping them with insights to improve future performance. However, human expertise remains essential for interpreting results, validating nuanced decisions, and ensuring alignment with project and regulatory requirements. A "human-in-the-loop" approach (combining GenAI-driven analysis with expert review) enhances overall quality, improves Reconciliator performance through informed feedback, and supports continuous process refinement.

Ultimately, with ongoing prompt optimization and targeted training, this approach can deliver higher-quality outcomes and improved cost-efficiency across COA localization projects.

Citations:
Koller, M. et al. (2012). The process of reconciliation: evaluation of guidelines for translating quality-of-life questionnaires. Expert Review of Pharmacoeconomics and Outcomes Research 12(2), 189–197.
CELF-5 - Clinical Evaluation of Language Fundamentals | Fifth Edition. Copyright © 2013 NCS Pearson, Inc. All rights reserved. Web: CELF-5 - Clinical Evaluation of Language Fundamentals | Fifth Edition | Pearson Assessments US

GenAI in Action:

A Study on the Efficacy of AI Comparative Review Output

LIONBRIDGE | Pearson

Presented By Lionbridge: Stephanie Casale

OBJECTIVES

Linguistic Validation (LV) is the process by which Clinical Outcomes Assessments are localized and reviewed for accurate and consistent data collection across target locales. The process is lengthy and complex by design. This approach ensures the highest quality and most thorough translations, but the complexity comes at a cost. To reduce this process's monetary and time burden, our study aims to shorten turnaround times and outsourcing costs while maintaining the high standards the process requires.

This poster examines the feasibility of using Generative AI to perform Comparative Review (CR). Comparative Review is a key quality assurance step in the LV process. It compares source text and back-translated text to determine conceptual equivalence. Because it's an intermediary step, the prior and subsequent steps are performed by trained, experienced linguists, making CR a prime candidate for automation. This approach minimizes the risk of errors 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.

METHODS

We first developed a prompt that produced the expected outcome of a Comparative Review Result and a Comparative Review Comment, which provided further details related to the Result. Comparative Review Results would be divided into three categories:

✓ **Identical** – This result indicates the source text and back translation were precisely the same in every way, including capitalization and punctuation.

✓ **Equivalent** – This result indicates that while there may be differences in wording, sentence structure, or other details, the meaning of the segments remains conceptually equivalent. The reader would understand they convey the same information.

✓ **Needs Review** – This result indicates something in the two segments renders them conceptually inequivalent. A reader could misunderstand the translated text, thinking it conveyed something unintended by the source text.

The prompt was then designed to produce a Comparative Review Comment for any non-identical result. These comments explain conceptual differences between the two segments, including 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. The prompt should also have ignored any additional text unrelated to the meaning of the source text (i.e., formatting tags, etc.).

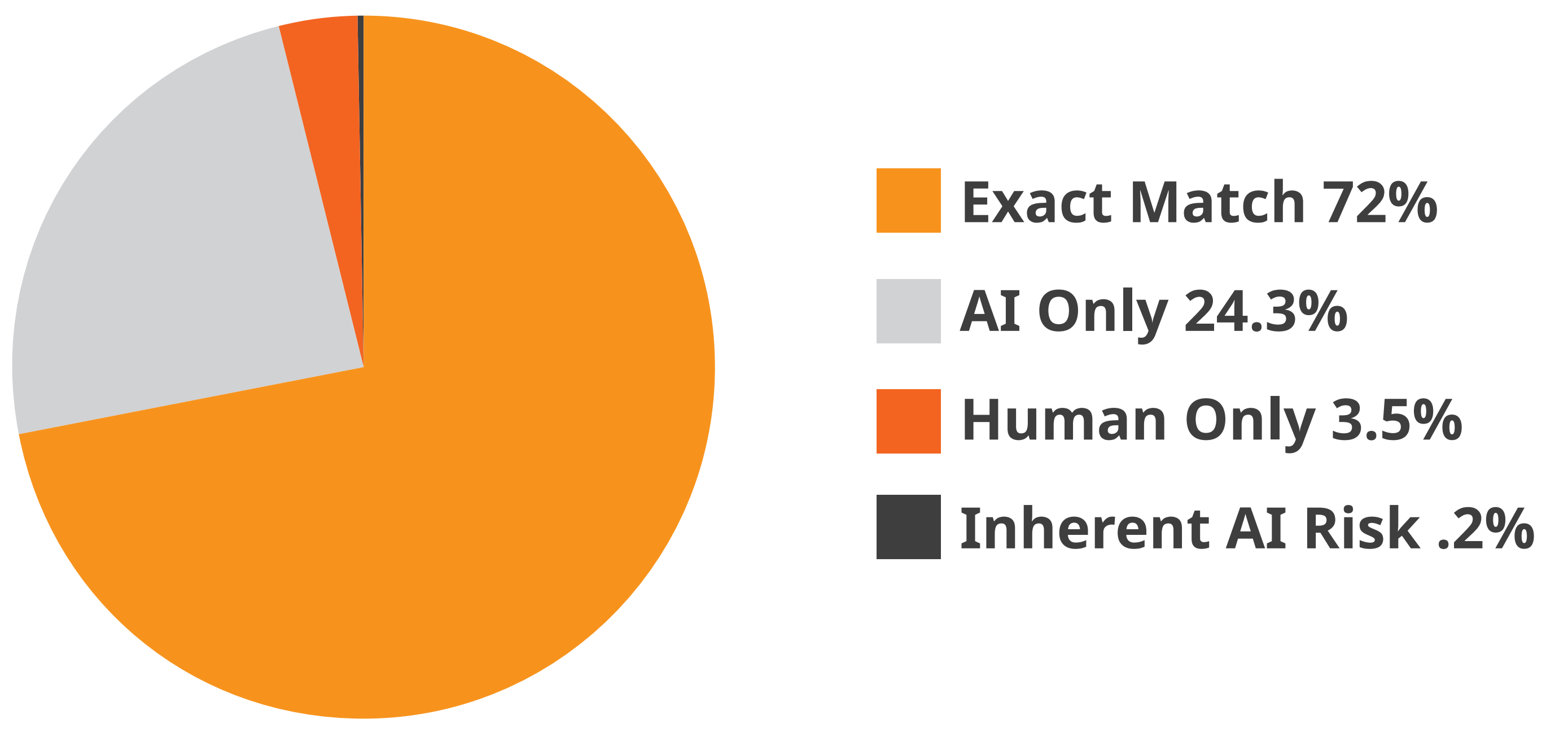
In compiling data, we used a previously localized Pearson assessment, Delis-Kaplan Executive Function System (D-KEFS) - Trail Making Test. It contains approximately 1000 source words. Back translations were performed by a variety of linguists who were native speakers in either English or the target language. Both a professional Comparative Reviewer and members of our COA Localization team conducted multiple Comparative Reviews. They had varying levels of Comparative Review experience and a mix of native languages, thus ensuring a wide range of outputs to compare.

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%.

This breakdown can be seen in the following chart. It includes 72.09% of exact matches between the human and AI outputs, with an additional 24.3% of segments flagged as discrepancies by the AI, but not by the human comparative reviewer. The results included 3.5% of human-flagged discrepancies the AI noted as equivalent. Results also included 0.17% of identical responses, which researchers noted as risky due to forward translation not being in Latin characters, and thus easily missed when not accounting for that text.

PERCENTAGE OF AI MATCH TO HUMAN OUTPUT



Notable Percentages:
The GenAI outputs analyzed by our Language Experts revealed 3 main areas GenAI can effectively support:

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 and would not have been noted by the AI prompt.

Inconsistent responses – 1.26%: During our review, the AI occasionally produced different responses for the same set of segments, accounting for just over 1% of the total data.

CONCLUSION

Generative AI has the potential to save significant time and money in the Linguistic Validation process.

Further studies should expand the data set. Next steps will need to include a Proof of Concept to examine the effects of using this output with linguists during the Comparative Review Reconciliation step. Further refinement of the prompt could help mitigate some inconsistencies and risks.