PCR79

EVALUATING GENAI VS. HUMAN SCREENSHOT REVIEW OUTPUTS IN ECOA LOCALIZATION:
Does GenAI Hold the Key to Improved Feedback?

LIONBRIDGE

Authors: Kathryn Nolte, Karolina Elizondo Jimenez, Rupali Kadam, Melinda Johnson

INTRODUCTION

The translation, migration, and screenshot review of Electronic Clinical Outcome Assessments (eCOAs) have traditionally relied heavily on manual human activities, due in part to the non-editable nature of on-screen content. During the screenshot review process (SSR), target language screen reports are checked against source master screen reports and, if applicable, the legacy content (i.e., the original paper questionnaires).

There are two types of screenshot review executed by Lionbridge: **Simple** and **Complex**. Complex SSR differs from Simple SSR in that it requires an explicit line-by-line check against all target legacy content. Lionbridge recognized and tested the potential of generative AI (GenAI) and Optical Character Recognition (OCR) technology to supplement traditional screenshot review methods, reduce costs and timelines (i.e., the number of review rounds), improve quality, and ultimately lead to better patient outcomes. This research was carried out leveraging both the Simple and Complex screenshot review functionalities of Lionbridge's proprietary Aurora AI Clinical Outcomes™ tool.

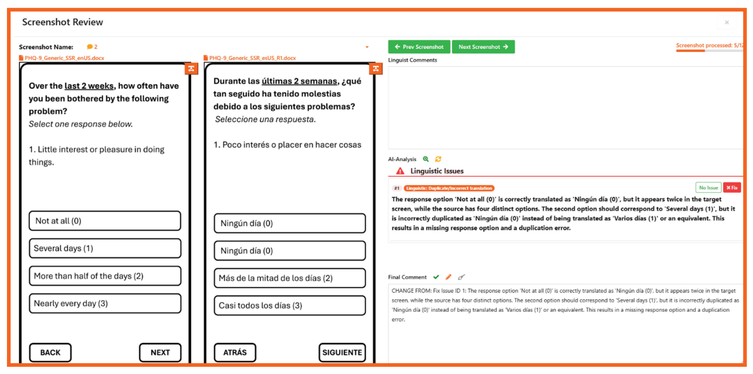
METHODS

Lionbridge leveraged a secure GenAI engine to generate quality assurance feedback for target screenshot reports for 5 patient-facing Electronic Clinical Outcome Assessments (eCOAs) of varying lengths and complexity levels with 11 types of errors intentionally incorporated into them for testing purposes: Missing Content, Untranslated Content, Incorrect Translation, Missing eCOA Edits, Version Number Mismatches, Formatting Issues, Morphology Issues, Tag Issues, Line Break Issues, Capitalization Issues, and Scale Anchor Issues. 16 target languages were tested in total, representing a variety of alphabets and language families: Bulgarian (Bulgaria), Polish (Poland), Romanian (Romania), Greek (Greece), Spanish (Argentina, Mexico, United States), French (Belgium, France), Portuguese (Brazil, Portugal), Hungarian (Hungary), Turkish (Türkiye), Korean (Korea), Traditional Chinese (Taiwan), and Thai (Thailand). Prompts were customized until a suitable output was obtained, in line with the latest eCOA industry standards and established practices. Simultaneously, we sent the same eCOA screen reports to Lionbridge-approved linguists for human review. An impartial reviewer was then tasked with validating both outputs (AI vs. human) and rating them for accuracy and completeness.

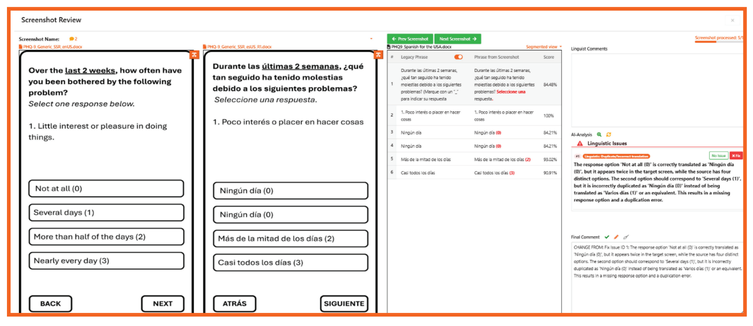
Prompt Design Evolution:

The first phase of our research utilized a GenAI prompt for Simple Screenshot Review, which is ideal for comparing source screen reports to target screen reports following the translation workflow (either translation/edit or forward translation/back translation). Once this feature of the tool was finalized, we turned our attention to developing and testing a more nuanced Complex Screenshot Review GenAI functionality, which is better suited for content that has undergone migration because it goes one step further than its Simple counterpart. In addition to comparing target screen content with the source master screen report, the Complex functionality also compares target screen content with target legacy content.

Simple Screenshot Review Interface:



Complex Screenshot Review Interface:



CONCLUSION

Our research demonstrated that integrating GenAI with human review is the most effective strategy for screenshot review in eCOA translation and migration. The prompt is being continuously refined to detect an increasing number of issues over time. However, it's still currently recommended to maintain human involvement in the screenshot review process to ensure optimal quality for this highly sensitive content. GenAI can significantly boost speed, efficiency, and accuracy by identifying preliminary errors, while human expertise remains essential for nuanced decision-making and identifying issues AI may have missed. This "human-in-the-loop" approach has the capacity to streamline timelines by reducing the number of review rounds needed, reducing costs by saving the linguists' time during their own quality checks, and, most importantly, strengthening our ability to capture the patient voice with greater precision and higher quality, thus ensuring a more reliable, patient-centered experience.

RESULTS AND INTERPRETATIONS

Analysis of the quality assurance feedback for the Simple SSR GenAI feature revealed that it is a very helpful QA check that can be used in collaboration with human feedback for both translation and migration projects. The Simple feature is particularly well-suited to support linguists with translation/screenshot review projects that do not require a cross-check against legacy content. By leveraging this feature, the Aurora AI Clinical Outcomes Tool captured several instances of all 11 error types reflected in the target screens. However, it was limited in some cases by its inability to check the legacy.

While the Complex SSR GenAI feature is still a work in progress (expected to deploy in Q1 2026), interim results show that it catches everything the Simple SSR Tool identified — and more. It also delivers more nuanced feedback compared to the Simple SSR Tool. The ability to cross-check against the legacy will enable the Complex SSR feature to eliminate some of the false positives currently produced by the Simple SSR Tool, as outlined in the chart below.

ERROR TYPE	SIMPLE SSR GENAI FEATURE: NOTES AND LIMITATIONS	COMPLEX SSR GENAI FEATURE: NOTES AND LIMITATIONS	EXAMPLES
Missing Content	Can flag if target content is missing	Can confirm if target text is missing intentionally because it was not present in the legacy	Portuguese (Portugal) #4 Inconsistency: Missing content Target Paragraph: 1 — No Issue / Fix **"The statement '[scale name] is subject to [copyright holder]'s Terms of Use.' is present in the source but missing in the target. All content should be retained and translated."**
Untranslated Content	Can flag text that has been left in English and suggest that this could be intentional (i.e., copyright text)	Can confirm if the text was left in English intentionally	Bulgarian (Bulgaria) Linguistic Issues #1 Linguistic: Untranslated content — No Issue / Fix **"The line '[scale name] - items H17, BP1, N6, GE6' remains in English in the Bulgarian screen. This should be translated or localized for Bulgarian users unless these are standardized item codes that should remain in English. If these are not standard codes, provide a Bulgarian translation."**
Incorrect Translation	Can identify deviations from the source	Can identify translations that are incorrect because they deviate from the legacy, even if they match the source	Portuguese (Portugal) #3 Linguistic: Inconsistent scale anchors — No Issue / Fix **The response options in the target text do not consistently match the gradation of the source. For example, '1 - Muito pouco' (Very little) is stronger than 'A little bit', and '2 - Mais ou menos' (More or less) does not directly correspond to 'Somewhat'. '4 - Muito' (A lot) is not equivalent to 'Very much'. The scale anchors should be consistently translated to preserve the intended gradation.**
Missing eCOA Edits	Equal functionality expected for Simple and Complex features		Portuguese (Portugal) #1 Linguistic: Incorrect translation Source Paragraph: 6 Target Paragraph: 4 — No Issue / Fix **The instruction 'Please select one number per statement to indicate your response as it applies to the past 7 days.' is translated as 'Faça um círculo ou marque um número por afirmação para indicar a sua resposta no que se refere aos últimos 7 dias.' The phrase 'Faça um círculo ou marque um número' introduces the instruction to circle or mark a number, which is not present in the source text. The source only instructs to select a number, not to circle or mark. The translation should not introduce additional instructions not present in the source.**
Version Number Mismatches	Can flag version number mismatches between source and target, with the caveat that the differences may be intentional	Can confirm if version number differences are intentional by checking the legacy	Korean (Republic of Korea) #3 Inconsistency: Version number mismatch — No Issue / Fix **The version number in the source is 'v1.1', while in the target it is 'v1.0'. The version number should match between source and target unless there is a justified reason for the difference.**
Formatting Issues	Can flag formatting that differs from the source master screen report	Can confirm if formatting deviates from the source intentionally, based on the legacy content	Portuguese (Portugal) Formatting/Layout Issues #1 Formatting/Layout: Inconsistent bolding/underlining Source Paragraph: 7 Target Paragraph: 5 — No Issue / Fix **In the source screen, the phrase 'past 7 days.' is bolded, while in the target screen only 'últimos' is bolded. The bolding should match the source, with 'últimos 7 dias.' in bold in the target.**
Morphology Issues	Can detect unexpected gender morphology for adjectives, with the caveat that it may be intentional (i.e., all adjectives could be in the feminine form due to a 100% female patient population)	Can provide additional context re expected gender morphology by checking against the legacy	Spanish (United States) Linguistic Issues #1 Linguistic: Inconsistent translation Source Paragraph: 7 Target Paragraph: 5 — No Issue / Fix **The source uses 'I feel fatigued,' which is gender-neutral. The target uses 'Me siento agotada,' which is feminine. For a general patient-facing questionnaire, the translation should be gender-neutral ('Me siento fatigado/a' or 'Me siento fatigada o fatigado') unless the instrument is specifically for female patients.**
Tag Issues	Equal functionality expected for Simple and Complex features		Thai (Thailand) Formatting/Layout Issues #1 Formatting/Layout: Underlining/HTML Tag Display Source Paragraph: 7 — No Issue / Fix **The target screen displays HTML tags (<u>) instead of rendering underlined text for 'การกลืนลำบาก', 'แย่ที่สุด', and 'ความรุนแรง'. The source uses underlining for emphasis, but the target shows raw tags, which is incorrect formatting. The underlining should be properly rendered in the target.**
Line Break Issues	Equal functionality expected for Simple and Complex features		French (France) Formatting/Layout Issues #1 Formatting/Layout: Line break Source Paragraph: 6 Target Paragraph: 4 — No Issue / Fix **The target screen introduces a line break in 'habi-tuelles' that splits the word across two lines with a hyphen, which is not present in the source screen. This disrupts readability and should be corrected so that 'habituelles' appears on one line without a hyphen.**
Capitalization Issues	Can flag unexpected capitalization patterns that deviate from the source, with the caveat that these may be intentional	Can confirm if unexpected target screen capitalization matches the legacy	Hungarian (Hungary) #2 Linguistic: Inconsistent capitalization Source Paragraph: 6 Target Paragraph: 6 — No Issue / Fix **The target text uses lowercase for 'magyar verzió', while the source uses title case 'English version'. For consistency and professionalism, the Hungarian should use 'Magyar verzió Magyarország részére'.**
Scale Anchor Issues	Equal functionality expected for Simple and Complex features		Portuguese (Portugal) #2 Formatting/Layout: Scale anchor placement Source Paragraph: 8 Target Paragraph: 7 — No Issue / Fix **The scale anchor text in the target is longer and may not sit directly under the endpoint of the scale, potentially causing layout misalignment. Ensure the anchor text does not spread into the center and remains directly under the endpoint.**

Man vs. Machine:

- GenAI is much faster than humans when evaluating target screen content. It generates feedback for all target screens for a given scale in a matter of seconds.
- Lionbridge quality assurance specialists noted time and efficiency gains with the GenAI process because the Aurora Clinical Outcomes Tool SSR interface enabled them to easily access legacy files, master screen reports, and target language screen reports simultaneously.
- GenAI caught several issues that human reviewers missed, particularly line breaks, capitalization, wrong translation (including duplicated text), and scale anchors. These issues may be easier for GenAI to catch because the tool uses Optical Character Recognition (OCR) to parse text discrepancies in non-editable image content that the human eye can easily miss.
- GenAI SSR sometimes caught errors inconsistently across languages. For example, while GenAI caught the missing translation of "items" in most languages, it was missed in Spanish. This was likely because the only difference was an accent mark ("ítems").
- Human reviewers were less likely to incorrectly flag context-dependent issues (i.e., female patient populations, intentional version number discrepancies, copyright text intentionally left in English, etc.).