语言验证的未来就在眼前：
利用 GenAI 增强对照流程合规性

LIONBRIDGE | Pearson

主讲人：来自 Lionbridge 的 Elisabet Sas Olesa、Karolina Elizondo、Kathryn Nolte、Nathalie Azuaje、Melinda Johnson

背景介绍

本研究探讨了生成式人工智能 (GenAI) 如何赋能语言验证 (LV) 流程，尤其关注 GenAI 在提高双重翻译对照步骤效率方面的作用，该步骤是将两个或多个独立的正向翻译合并为一个译文版本的阶段（Koller 等人，2012 年）。

我们研究的主要目标是评估 GenAI 在检测通过传统方法生成的对照结果中不合规之处的能力。

产品设计演进

最初的提示词被设计为与传统的对照输出协同运行，通过将译文 A 与译文 B 进行对比，从而将对照工作者的最终决定生成一个二元的“通过/未通过”结果。

然而，这种方法存在显著的局限性，尤其是在应用于非英语语言及高度专业化的语境时，例如那些由特定治疗领域或生活质量评估工具中的常见疾病/状况所定义的语境。针对这些发现，我们应用了 GenAI 来评估对照结果，从而支持我们的质量保证 (QA) 工作流程。随后，我们将更新后的提示词整合到了 *Aurora Clinical Outcomes* 中，它是 Lionbridge 专有的端到端语言验证平台。通过利用 GenAI 支持内部决策，还能对对照工作者的思维过程进行深入洞察，确保任务按照必要的标准和行业要求完成。此类要求反映在了对照目标中，并在任务说明中提供给对照工作者的指导中加以强调：

1. **与原始量表保持概念对等**
2. **适应当地文化**
3. **面向目标研究人群/受众的无障碍性**
4. **偏见趋势检测**

方法

我们使用一份 PerfO 样本（1,000 到 2,000 词）进行了实际分析，其中包含亚太地区 (APAC)、长尾及区域语言变体的混合内容。基于这些参数，我们从之前完成的语言基础临床评估®-第五版 (CELF-5) 的本地化内容中选取了多样化的文件。CELF-5 是一套灵活的个体化测试系统，常用于辅助临床医生准确诊断儿童和青少年的语言障碍（NCS Pearson, Inc.，2013 年）。分析所选的目标语言为西班牙语（阿根廷）、西班牙语（西班牙）、法语（法国）、亚美尼亚语（亚美尼亚）、日语（日本）和繁体中文（中国台湾）。

结果

积极成果

语言专家对 GenAI 输出的分析揭示了三个关键支持领域：

- **理由无效或不完整：**GenAI 持续标记出了缺乏充分语言或概念依据的评论（例如，“译文 A 更好”或“译文 B 更可取”），还检测到了未解决的疑问或未回答的问题。这促使相关人员提出优化提示词的建议，确保此类案例会进行上报以便跟进处理。
- **理由缺失：**所有专家都指出，GenAI 可靠地识别出了缺失、不完整或不清晰的理由，显著加快了 QA 审核流程。
- **效率：**GenAI 展示出了高速处理能力，可在数秒内分析多达 300 个文本句段。

这种预筛查功能可帮助语言专家快速过滤结果，并判断文件是否已准备好进入语言验证 (LV) 流程的下一步。

需要改善的方面

尽管 GenAI 潜力巨大，但其固有的不可预测性也在提示词执行过程中带来了一定限制：

- **行为不一致：**在评估不同句段的相同评论时，GenAI 的输出结果会表现出不一致的行为，即使在同一文件中也是如此。这削弱了 GenAI 输出作为独立解决方案的可靠性。
- **上下文/元数据引用挑战：**在当前设置下，频繁使用诸如“见上文评论”之类的模糊评论和引用对 GenAI 构成了严峻挑战，因其无法找到所引用的先前评论。专家建议使用简洁明确的引用，并在不同句段中重复说明，这应是对照工作者的预期交付成果之一。
- **可行的对照评论缺乏标准化：**我们的专家一致认为，各方需达成更精准、明确的定义与共识，确定哪些内容构成可接受的论证依据和详细理由。

结论

本研究证明，GenAI 作为质量保证支持工具，在识别漏洞、不合规项和缺失理由方面具有巨大潜力。

GenAI 能够更快地检测出返工需求，使本地化团队能够向利益相关者提出更具针对性的反馈，从而为他们提供改进未来表现的洞见。然而，在解读结果、验证具有细微差别的决策以及确保符合项目和监管要求方面，人类的专业知识仍然至关重要。采用“人工介入”方法（将 GenAI 驱动的分析与专家审查相结合）可以提高整体质量，通过有见地的反馈改进对照工作者的表现，并支持持续的流程改进。

最终，通过持续的提示词优化和有针对性的训练，这种方法可以在临床结果评估 (COA) 本地化项目中交付更高质量的成果，实现更高的成本效益。

引用：
Koller, M. 等人 (2012 年)。对照流程：生活质量问卷翻译指南评估。药物经济学和结果研究学会专家审查 12(2), 第 189-197 页。
CELF-5 — 语言基础临床评估 | 第五版。版权所有 © 2013 NCS Pearson, Inc. 保留所有权利。网页：CELF-5 — 语言基础临床评估 | 第五版 | Pearson Assessments US

在实践中运用 GenAI：
关于 AI 比较审核输出有效性的研究

LIONBRIDGE | Pearson

主讲人：来自 Lionbridge 的 Stephanie Casale

目标

语言验证 (LV) 是将临床结果评估进行本地化并审查的流程，用于确保跨目标地区准确且一致地收集数据。该流程设计得冗长且复杂，能够确保最高质量和最细致的翻译，但其复杂性也带来了相应成本。为减轻该流程的资金和时间负担，本研究旨在缩短周转时间并降低外包成本，同时保持流程所需的高标准。

本海报探讨了使用 GenAI 执行比较审核 (CR) 的可行性。比较审核是 LV 流程中的关键质量保证步骤，它将源文本与回译文本进行比较，以判定概念对等性。它是一个中间步骤，其前后步骤均由训练有素、经验丰富的语言学家执行，这使得 CR 成为自动化的首选步骤之一。这种方法可在最终定稿之前尽可能降低错误漏检风险。

本研究旨在开发一个提示词方案，该方案应至少维持当前人工供应商在比较审核步骤中的现有质量水平。

方法

我们首先开发了一个提示词方案，能够生成预期的比较审核结果和比较审核评论，评论可提供与结果相关的更多详细信息。比较审核结果将分为三类：

完全一致 — 此结果表明源文本和回译文本在各方面都完全相同，包括大小写和标点符号。

概念对等 — 此结果表明，尽管在措辞、句子结构或其他细节上可能存在差异，但句段的意思在概念上保持对等。读者能理解它们传达的是相同的信息。

需要审核 — 此结果表明两个句段中存在某些内容，导致它们在概念上不对等。读者可能会误解译文，认为它传达了源文本未意图表达的内容。

然后，我们对提示词进行了设计，要求对所有并非完全一致的结果生成比较审核评论。这些评论应解释两个句段之间的概念差异，包括详细说明普通读者可能产生的误解。我们要求提示词忽略所有标点符号和大小写差异，除非它们与含义和理解直接相关。提示词还应忽略与源文本含义无关的任何附加文本（例如格式标签等）。

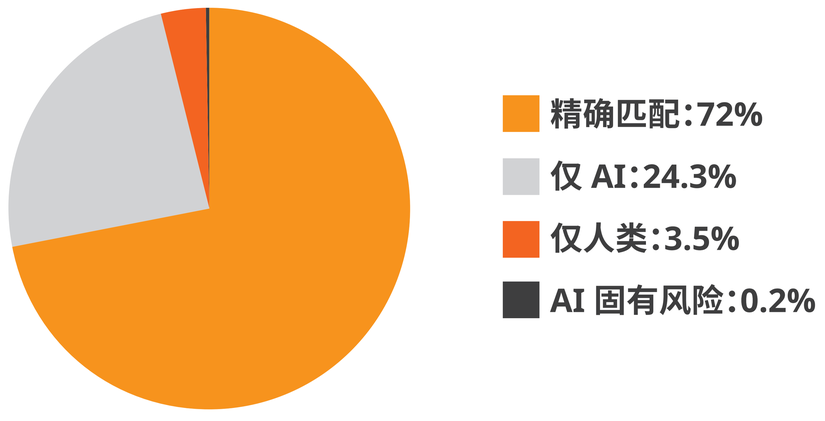
在编制数据时，我们使用了先前已本地化的 Pearson 评估工具 Delis-Kaplan 执行功能系统 (D-KEFS) 进行路径绘制测试。该测试包含约 1,000 个源词。回译工作由母语为英语或目标语言的多元化语言学家完成。专业比较审核员以及我们 COA 本地化团队的成员都进行了多次比较审核。这些人员具有不同的比较审核经验水平，且母语背景多样，从而确保输出结果的广泛性，以便进行比较。

结果

初步结果颇有发展潜力，原始评估和回译差异的描述清晰简洁，整体初步准确率为 96.4%。

具体细分数据如下图所示。结果显示，人类与 AI 输出之间完全匹配率为 72.09%，另有 24.3% 的句段被 AI 标记为差异内容，但未被人类比较审核员标记；3.5% 的人类标记差异被 AI 判定为等效内容；此外还有 0.17% 的完全一致响应，研究人员指出这部分内容存在风险，因为正向翻译未使用拉丁字符，若未考虑到对相应文本进行处理，这些字符容易被遗漏。

AI 与人类输出之间的匹配百分比



值得注意的百分比数据：
我们的语言专家对 GenAI 输出进行分析后揭示，GenAI 可在三个主要领域提供有效支持：

AI 固有风险 — 0.17%：此数据包含因正向翻译中存在非拉丁字母而可能被人工标记，但未被 AI 提示词识别到的句段。

响应不一致 — 1.26%：在我们的审核过程中，AI 偶尔会对同一组句段生成不同的响应，占总数据量略高于 1% 的比例。

结论

GenAI 有望在语言验证流程中显著节省时间和成本。

后续研究应扩大数据集。接下来需在研究中包含概念验证，以评估在比较审核对照步骤中使用该输出与语言学家协作的效果。进一步优化提示词可能有助于降低不一致性并减轻风险。