語言驗證的未來已然降臨：
運用 GenAI 增進核對流程的法規遵循

由 Lionbridge 發表：Elisabet Sas Olesa；Karolina Elizondo；Kathryn Nolte；Nathalie Azuaje；Melinda Johnson

簡介

本研究旨在探索生成式人工智慧 (GenAI) 可以如何支援語言驗證 (LV) 流程，尤其聚焦於它的應用，如何有助提升雙重翻譯核對 (Dual Translation Reconciliation) 這個步驟的效率。雙重翻譯核對這個階段是指核對兩個或更多個各自獨立完成的翻譯，並將它們統整合併為一份翻譯 (Koller et al., 2012)。

本研究的主要目的，是要評估 GenAI 能否從以傳統方法製作的核對結果中，偵 測到未能遵循法規的地方。

研究方法

我們使用一個內含 APAC、長尾及地區性語言變體的 PerfO 樣本 (1,000-2,000 字) 進行了一次實務分析。根據這些參數，我們從之前已本地化的《Clinical Evaluation of Language Fundamentals®-Fifth Edition》(臨床語言基礎評估量表第五版，簡稱 CELF-5) 中，挑選了多個不同的檔案。這是個靈活彈性且可單獨進行的測驗系統，通常是用來協助臨床醫師正確地診斷兒童及青少年的語言障礙 (NCS Pearson, Inc., 2013)。分析所選的目標語言則為：西班牙文 (阿根廷)、西班牙文 (西班牙)、法文 (法國)、亞美尼亞文 (亞美尼亞)、日文 (日本) 以及繁體中文 (台灣)。

提示設計演進

初始提示是要搭配傳統的核對輸出使用，它會將翻譯 A 與翻譯 B 相比較，就核對人員的最終決定生成「通過/不通過」的二元結果。

然而，這種做法伴隨著很大的侷限，尤其是應用在英文以外的語言或是高度專業的脈絡時，例如由特定治療領域定義的語境，或是生活品質工具經常評量的疾病/症狀。針對這些發現，我們使用 GenAI 來評估核對結果，藉此支援我們的 QA 工作流程。更新後的提示會接著整合到 Lionbridge 專有的全方位語言驗證平台 *Aurora Clinical Outcomes* 中。使用 GenAI 支援內部決策的另一個好處，是能為核對人員的思考過程提供一些深入見解，確保工作能遵循必要的標準和產業要求執行。這些要求除了會反映在「核對目標」(GOALS OF RECONCILIATION) 中，也會在供核對人員參考之指引的任務說明中特別強調：

1. **與原文評量的概念等同性**
2. **因應當地文化調適**
3. **對目標研究族群/對象的取用性**
4. **偏見趨勢的偵測**

結果

正面結果

語言專家分析 GenAI 輸出後，發現它能支援三個重要領域：

- **無效或不完整的論證：**GenAI 能一致地標記出在語言或概念上缺乏足夠論證的評論 (例如：「翻譯 A 更好」或「偏好翻譯 B」等)，也能偵測到未解決的查詢或未回答的問題，並進一步顯示精進提示的建議，確保此等情況能向上呈報以利後續跟進。
- **遺漏的論證：**所有的專家都指出 GenAI 能可靠地找出缺少、不完整或不清楚的論證，可大幅加快 QA 審閱的時間。
- **效率：**GenAI 展現了很高的處理速度，可在數秒內分析高達 300 個文字句段。

這種預先篩選的做法，讓語言專家得以快速篩選結果，並判斷檔案是否可以進入 LV 流程的下個步驟。

可改善的地方

儘管 GenAI 在應用上極具潛力，但它本身難以預測的特性，也會在執行提示時引入某些限制：

- **不一致的行為：**在評估不同句段的相同評論時，GenAI 結果會展現不一致的行為，即使是同一個檔案亦是如此。這削弱了 GenAI 輸出的可靠性，使它難以單獨做為一個解決方案使用。
- **脈絡/中繼資料參照上的挑戰：**經常使用像是「如上述評論」等含糊不清的評論和參照，對目前建置環境下的 GenAI 構成了嚴重的挑戰。GenAI 無法找到所參照的先前評論。專家建議應該要求核對人員在交付結果中，於不同句段間使用簡潔明確的參照並重複說明。
- **對可行的核對評論缺乏標準化做法：**我們的專家都同意，對於何謂可接受的推論以及詳細的理由說明，應該要提供簡潔明確的定義，方便所有參與的人士都能校準與配合。

結論

本研究證明 GenAI 能有效找出缺口、辨識出未能遵循法規的地方以及找出遺漏的論證，是極具潛力的品管支援工具。

GenAI 可以更快地偵測到重做的需求，並可讓本地化團隊提出更精準的意見回饋給相關利害人士，使他們能運用這些深入見解改善日後的成效表現。然而，人類的專業能力，仍舊是解讀結果、驗證細膩決策以及確保符合專案與法規要求的關鍵要素。這種「人機迴圈」的做法 (將採用 GenAI 的分析與專家審閱相結合)，不但可提升整體品質、透過有憑有據的意見回饋增進核對人員的表現，亦能支援持續不斷地精進流程。

最終，透過持續的提示最佳化以及精準的訓練，這種做法可以為 COA 本地化專案交出更高品質的結果以及更高的成本效益。

引用資料：
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.

GenAI 應用實例：
AI 比較審閱輸出之功效研究

由 Lionbridge 發表：Stephanie Casale

研究目標

語言驗證 (Linguistic Validation，簡稱 LV) 這個流程，是指審閱各個目標地區的臨床結果評估本地化版本，檢查其正確性以及資料收集的一致性。這個流程故意設計得既冗長又複雜，好確保能取得品質最佳也最完善的翻譯。但這樣的複雜性也伴隨著高昂的代價。為了降低這個流程的財務及時間負荷，本研究旨在設法一方面縮短交付時間並降低委外成本，一方面又能保持流程所需的高標準。

本文件探討的是使用生成式 AI 執行比較審閱 (Comparative Review，簡稱 CR) 的可行性。比較審閱是 LV 流程中很重要的品管步驟，是指將來源文本與回譯文本相比較，以判定其概念等同性。這是個介於中間的步驟，由於之前與之後的步驟都是由訓練有素且經驗豐富的語言專家執行，這使得 CR 成了自動化的首選對象。這種做法可以將未在最終定稿前發現錯誤的風險降至最低。

我們的研究目標是要開發提示，同時其表現至少要媲美我們目前人類供應商在比較審閱方面的現有品質。

研究方法

我們先開發了一個提示，可產出預期的比較審閱結果及比較審閱評論，後者可進一步提供有關結果的詳細資料。比較審閱結果可以分成三個類別：

- **相同** – 這個結果代表原文跟翻譯在各種方面都完全一樣，包括大小寫跟標點符號。
- **等同** – 這個結果代表雖然在用字遣詞、句型結構或其他細節上有些差異，但句段的意義在概念上仍舊是等同的，也就是說讀者能理解它們傳達的是相同的資訊。
- **需要檢閱** – 這個結果代表有些因素導致兩個句段在概念上不是等同的，亦即讀者可能會誤解譯文，認為它傳達了原文並未表達的意思。

這個提示接著又加以調整，為任何不相同的結果產出比較審閱評論。這些評論會說明兩個句段在概念上的不同之處，包括深入解釋一般讀者可能會產生的誤解。除非與意義及理解直接相關，否則提示應忽略任何標點符號與大小寫的不同。此外，提示也應該要忽略任何跟原文意義無關的附加文字 (例如格式化標籤等)。

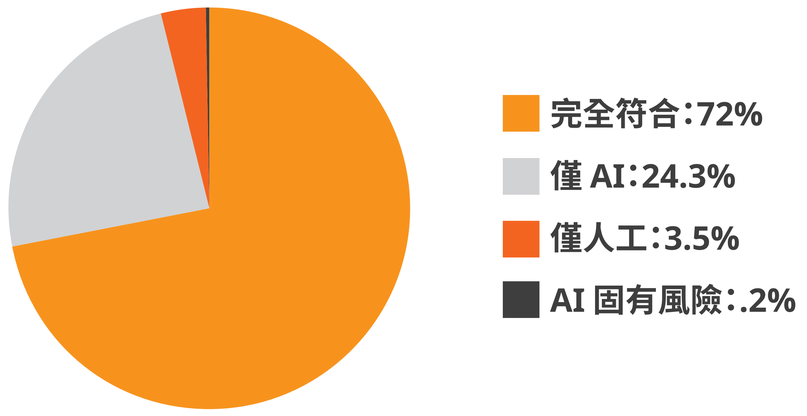
在資料彙集上，我們從已本地化的 Pearson 評量「Delis-Kaplan Executive Function System」(D-KEFS 執行功能測驗，簡稱 D-KEFS) 中，選用了「Trail Making Test」(軌跡標示測驗)。這個測驗的原文約有 1000 字。回譯則是由多個母語為英文或目標語言的語言專家進行。接著，由一位專業的比較審閱人員以及我們 COA 本地化團隊的成員完成了多個比較審閱。這些人員在比較審閱上的經驗多寡不一，並有多種母語，可確保有多種不同的輸出可供比較。

結果

初始結果顯示前景看好，能清楚簡潔地描述原始評估，同時回譯不一致的整體初步正確率為 96.4%。

下方表格是結果的詳細分析。其中，人類與 AI 輸出比對完全符合的比率為 72.09%，另外有 24.3% 的句段是被 AI 標記為不一致而非比較審閱人員。結果中有 3.5% 的句段是人工標記為不一致但 AI 認為是等同的。此外結果中也有 0.17% 的相同回應，是研究人員認為由於翻譯不是採用拉丁字元，如果不計入這些文字的話很容易會遭到忽略，因此有一定風險存在。

AI 與人工輸出比對的百分比



值得注意的百分比數據：
我們的語言專家分析 GenAI 輸出後，發現 GenAI 可以有效地支援 3 個主要領域：

AI 固有風險 0.17%：這個數據，是指翻譯中有非拉丁字母而被人類所標記，但沒有被 AI 提示注意到的句段。

不一致的回應 1.26%：在審閱時，AI 有時會對同一組句段產生不同的回應，這大概只佔整體資料的 1% 多一點。

結論

生成式 AI 具有潛力，能大幅節省語言驗證流程的時間與金錢。

後續的研究應該進一步擴大資料集。接下來必須要進行概念驗證 (Proof of Concept)，檢驗在比較審閱核對步驟中使用此輸出搭配語言專家的效果。進一步細微調整提示，應該也有助於減少一些不一致和風險。