LIONBRIDGE

AI 資料分析
模型評估適用

評比類別

想要有效地評估大型語言模型 (LLM) 的成效，必須採用結構化的做法，才能準確掌握語言生成多元面貌的本質。為此，我們特地挑選以下這些評估類別，以期能以完備的架構，來評量模型輸出在品質、可靠性及使用者相關性等方面的表現。這些類別納入了例如流暢度、完整性、領域專屬術語及文化相關性等層面，可確保進行真正的深入評估，而不是流於表面地確認正確性。這種做法也能讓組織信心十足地評量模型的輸出是否能滿足其使用者的期望、符合業務目標，以及遵循倫理標準。

正確性

評量回應的內容是否符合事實且沒有錯誤。

例：在法律相關回應中引用正確的法律。

完整性

檢查問題的所有部分是否都已充分回覆。

例：能在要求摘要的回應中涵蓋所有重點。

相關性

評量回應是否有直接回答詢問且沒有離題。

例：避免產品說明中出現不相關的資訊。

一致性

確認回應的內容符合邏輯，沒有矛盾之處。

例：在多段落的回應中不會有自相矛盾的問題。

流暢度

評估語言的文法正確與否以及是否自然流暢。

例：能使用正確的句型與標點符號。

幻覺 (反向)

評估模型是否會避免捏造事實。

(幻覺產生的次數越少 = 反向分數越高)

例：不會捏造不存在的產品功能。

術語

評量能否正確地使用領域所屬的術語和專業行話。

例：在醫療保健回應中使用精準的醫療術語。

可讀性

評量目標對象是否能輕易地閱讀和理解文本。

例：針對一般大眾採用簡明扼要的用字遣詞。

文化相關性

評估回應是否能敏感地察覺目標對象的文化規範並做出適切的反應。

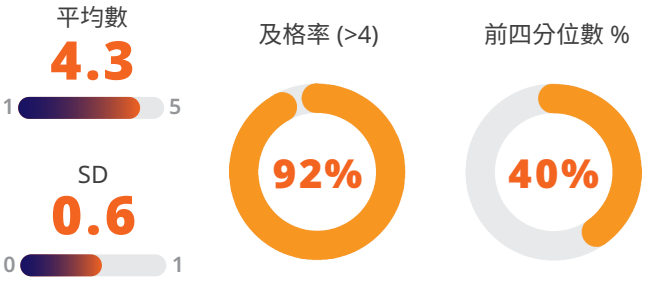
例：避免使用不適合全球目標對象的成語或指涉。

藉由系統性地使用這些類別，並採用李克特 (Likert) 五點量表評分，客戶便能取得可執行的深入見解，了解模型在哪些方面表現卓越，哪些地方又需要再微調、進一步訓練或人類監督。這樣一來，便可確保 AI 部署在例如客戶支援、內容生成、產品推薦或其他領域等不同應用上，都能提供一致且高品質的體驗。

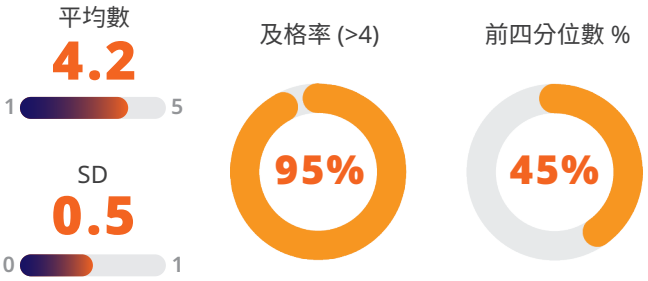
分析範例 | 取得更透徹的深入見解

Lionbridge 資料服務的分析功能，提供了非常重要的 AI 訓練流程可見度，可讓您清楚一窺流程的品質、一致性及效率。這些深入見解會有助我們找出資料多樣性中的模式、註解人員信度 (評估的一致程度)、模型優缺點以及輸出趨勢等，進而確保 AI 系統健全、無偏見且效能表現卓越。透過善用這些分析資料，組織便可據此做出明智的決策，進而信心十足地最佳化模型訓練、改善註解工作流程，以及加快 AI 開發腳步。

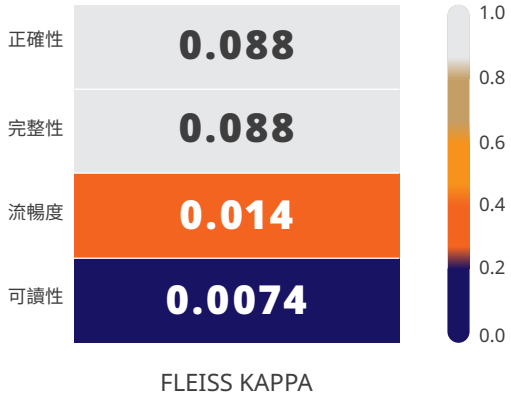
正確性



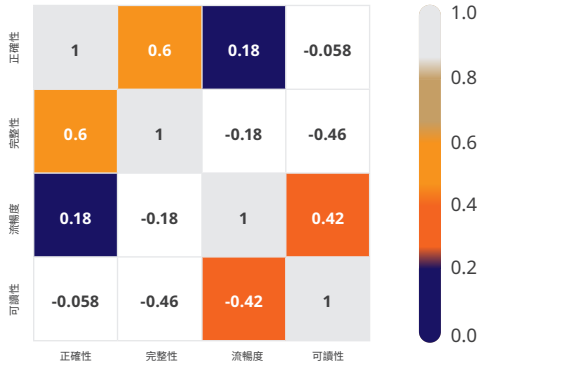
相關性



評分者間信度 (Fleiss' Kappa)



評估類別間的關聯性



如需深入了解，歡迎造訪
LIONBRIDGE.COM