SAMSUNG

三星發表 TRUEBench：有效評估實際 AI 模型應用生產力的基準

此獨創基準支援多語言生產力情境，消弭現有 AI 基準差距

三星電子日前發表 TRUEBench（Trustworthy Real-world Usage Evaluation Benchmark，真實場景可信度評估基準）- 此為三星研究院獨創之 AI 模型生產力評估基準。

TRUEBench 提供全面的衡量指標，評估大型語言模型（LLM）在實際工作應用中的表現。結合多元對話情境及多語言條件，確保評估結果具備可信度。

TRUEBench 借鑒三星內部運用 AI 提升工作效率的實際應用經驗，針對 10 個類別和 46 個子類別中常用的企業工作進行評估，例如內容生成、數據分析、摘要和翻譯等。該基準由人類與 AI 共同設計並不斷優化，透過 AI 自動評估，確保可靠的評分結果。

三星電子 DX 事業群技術長暨三星研究院負責人 Paul（Kyungwhoon）Cheun 表示：「三星研究院憑藉豐富的實際 AI 應用經驗，帶來深厚的專業知識和競爭優勢。期盼 TRUEBench 奠定生產力的評估標準，進而鞏固三星在科技產業的領導地位。」

近年來，隨著企業工作逐漸採用 AI，衡量大型語言模型效率的需求也日益增長。然而，現有基準主要用來衡量整體表現，且多半以英語為中心，並僅限於單回合問答結構。此方式間接限制其反映實際工作環境的能力。

為突破這些限制，TRUEBench 共提供 2,485 個測試集，其中涵蓋 10 個類別和 12 種語言(註一)，同時支援跨語言情境。測試集旨在檢驗 AI 模型的實際解決方案，三星研究院使用的測試集長度範圍，從最短 8 個字元到超過 20,000 個字元，不論簡單要求到冗長文件摘要皆可應用。

為評估 AI 模型的表現，必須訂定明確標準判斷 AI 回應是否正確。在現實世界的情境中，並非所有使用者都會在指令中明確說明其意圖。TRUEBench 的設計不僅考量答案的準確性，亦滿足使用者隱性需求的具體條件，進而使評估標準符合現實。

三星研究院透過人類與 AI 合作，針對評估項目進行驗證。首先，評估人員會建立一套標準，接著由 AI 進行審核，檢查是否有錯誤、矛盾或不必要的限制條件。然後，評估人員再次改良基準，重複此流程，使其逐漸精準。藉由上述交叉驗證的標準，對 AI 模型進行自動評估時，將能最小化主觀偏見，並確保一致性。此外，每一次測試皆須符合所有條件，模型才能通過評估基準。如此得以讓跨工作的評分更加詳盡且精確。

TRUEBench 資料範本與排行榜已於全球開放原始碼平台 Hugging Face 上發布，使用者最多可針對五種模型進行比較，一目了然地對照 AI 模型的表現。此外，平台亦公布回應結果的平均長度數據，方便同步對照各模型的表現與效率。詳細資訊請參閱 TRUEBench Hugging Face 頁面：https://huggingface.co/spaces/SamsungResearch/TRUEBench。

註一：中文、英文、法文、德文、義大利文、日文、韓文、波蘭文、葡萄牙文、俄文、西班牙文與越南文