AI 數據

企業為何需要在部署前評估 AI

將可靠AI與昂貴錯誤區分開來的隱秘步驟——以及為何企業匆忙跳過這一步是其最昂貴的捷徑。

編寫團隊

Lifewood Data Technology

知識文章

發佈時間

2026年6月1日 · 閱讀約 6 分鐘

在線閱讀

www.lifewoodtech.com/zh-hant/blogs/ai-evaluation-before-deployment

AI評估是指在模型進入生產環境前，將其與獨立保留的、針對特定任務的數據進行測試——測量其在真實輸入下的準確率、失敗模式和行為，而不是在演示條件下的表現。

當AI在未經評估的情況下上線會發生什麼?

一家中型保險公司花了八個月時間開發了一個AI助手，用於處理客戶理賠查詢。這項技術令人印象深刻，演示過程流暢，管理層已批准。它於週一早上正式上線。

到週三，客戶來電不斷。AI向客戶提供了關於其保單覆蓋範圍的錯誤信息——對一些客戶表示理賠會獲批，實際上並不會獲批；對另一些客戶則錯誤地報出了等待期。兩週內，該公司便將系統下線，並重新啟動。

技術本身並未出錯。AI能夠進行對話並流暢回應。但它從未在正式上線前接受過一次系統的評估。沒有人系統性地測試過其回答是否真正正確。這個故事每個月都在各個行業中上演——而幾乎完全是可以避免的。

什麼是AI評估?

AI評估是指在你信任該AI系統處理真實客戶或真實決策之前，對其進行測試。可以將其比作一場駕駛考試。你不會因為某人閱讀過手冊並觀看了一些視頻就將車鑰匙交給他。你必須在真實環境下進行測試，並由專業考官監督，確保其具備獨立駕駛的能力。

AI評估的工作方式與此相同。在部署前，你需要通過真實場景運行系統，讓專家審查其回答，識別其出錯的地方，並填補這些空白。只有在完成這些步驟後，系統才能真正面對真實用戶。大多數企業會急於推進這一步驟，因為它們渴望儘快上線——而這種捷徑的成本幾乎總是高於評估本身。

一個恰當評估的六個階段是什麼?

- **定義目標。** 明確AI需要完成的任務以及成功的標準。
- **準備測試數據。** 構建多樣化的、貼近現實世界的測試數據集，以充分考驗模型性能。
- **運行AI測試。** 在不同場景、邊緣情況和用戶類型上執行結構化的測試。
- **人工評審。** 專家評審員對輸出結果進行準確性、語氣、偏見和安全性的評估。
- **修復並重新訓練。** 將評估發現轉化為更優的訓練數據，使模型持續改進。
- **有信心地部署。** 僅在所有質量關卡均通過後才進行發佈。

第四階段是最常被跳過的，因為它需要時間且無法自動化。然而，這也是最有價值的階段：正是在這裡發現了自動化檢查永遠無法捕捉到的細微錯誤。

為什麼企業總是跳過評估環節?

這種模式是可以預測的。項目週期拉長，預算壓力逐漸增大。有人會問，為什麼評估階段不能縮短，因為演示效果看起來很好。內部只運行少量測試用例，顯而易見的缺陷被修復後，系統就上線了。

問題在於，內部團隊測試的是他們預期AI會遇到的問題。真實用戶提出的問題往往超出預想，提問方式各不相同，且來自不同的文化與語言背景。如果沒有由未參與系統構建的人員進行的結構化評估，這些失敗模式就會一直隱藏，直到演變為公眾問題。

2024年，一家航空公司在其聊天機器人向客戶提供了關於喪葬票價政策的錯誤信息後被法庭追責。客戶依賴該信息預訂了行程；當航空公司拒絕履行聊天機器人所述的票價時，法庭裁定航空公司需對其聊天機器人所陳述的內容負責。如果當時有結構化的評估流程，並由人工評審員檢查邊緣案例，這一問題本可以在觸及客戶之前就被發現。

一個好的評估測試應該包含什麼?

應關注失敗場景的覆蓋範圍，而非測試量的多少。一個從未包含邊緣案例的保留數據集，永遠無法發現這些案例，因此評估數據的準確性標準與訓練數據保持一致，達到95%以上，並需經過兩輪獨立審核，而非一輪。

多數人以為評估只是檢查AI能否給出正確答案，實際範圍遠不止於此。全面評估包括**準確性**：答案是否符合事實；**意圖理解**：系統是否理解用戶真正想表達的意思，而非僅看輸入的文字；**語氣**：是否符合品牌與場景；**安全性**：是否妥善處理敏感話題；**一致性**：在不同語言和地區是否表現可靠；以及**幻覺檢測**：識別模型自信地陳述不實內容的情況。

幻覺是這些風險中最危險的。模型不會主動標記自身的不確定性——無論答案是完全正確還是完全錯誤，它都會以同樣自信的語氣作答。如果沒有經過專門訓練的評審員來識別這類錯誤，這些錯誤將直接傳遞給客戶。

Lifewood是如何進行評估的?

評估數據集的構建標準與訓練數據一致：由客戶批准的黃金標準集、95%以上的準確率SLA以及95%以上的標註員間一致性閾值，覆蓋50多種語言，來自40多個交付中心。支撐這一評估能力的並非假設，而是實際人力投入——2025年期間，孟加拉國團隊累計投入了414,120小時的培訓工作量。保留的數據與訓練數據嚴格隔離，因為如果模型已經見過測試數據，那麼該測試集就無法衡量任何真實性能。

Lifewood在50多種語言和40多個交付中心構建評估數據集，並配備人工評審層對這些數據集進行評分。這項工作通過我們的QA流程和AI數據驗證服務直接反饋至訓練數據，因此一個評估週期不僅產出一份報告，更產出被修正的數據，使下一個模型版本得以優化。