AI 論文評分已經和「負擔過重」的老師一樣好，但研究人員表示它還需要更多研究

駐洛杉磯辦事處教育組

ChatGPT 能否減輕教師批改考卷的負擔？早期研究發現，大型語言模型的新人工智慧（也稱為生成人工智慧）在論文評分的準確性方面正在接近人類，並且可能很快就會變得更好。但我們仍然不知道使用 ChatGPT 進行論文評分最終會改善或是損害學生的寫作能力。

Tamara Tate 是加州大學爾灣分校的研究員，也是該校數位學習實驗室的副主任，她正在研究教師如何使用 ChatGPT 來改善寫作教學。最近，Tate 和她的 7 人研究團隊（其中包括亞利桑那州立大學的寫作專家 Steve Graham）比較了 ChatGPT 及人類在為中學生和高中生撰寫的 1,800 篇歷史和英語評分時的表現。

Tate 表示，ChatGPT 的表現「大致來說，可能與普通忙碌的老師一樣」、並且「確定與負擔過重、低於平均水平的老師一樣」。但是，她說 ChatGPT 還不夠準確，無法用於高風險測驗或影響最終成績的論文。

Tate 在 2024 年 4 月時於費城舉行的美國教育研究協會 2024 年年會上展示了她對 ChatGPT 作文評分的研究。（論文正在接受同行評審以發表，並且仍在修改中。）

最值得注意的是，研究人員從 ChatGPT 獲得了這些相當不錯的論文分數，而沒有先提供它樣本論文。這意味著任何老師都可以使用它以最少的費用和精力立即對任何論文進行評分。「教師可能有更多的經歷來指派更多的寫作作業。」Tate 表示。

泰特也警告，如果教師將過多的評分工作委託給 ChatGPT，寫作教學最終可能會受到影響。她說，看到學生的逐漸進步和了解學生常見錯誤對於決定下一步教什麼仍然很重要。例如，在學生的論文中看到大量連續句子可能會提醒教師教導學生如何分解句子。但如果教師沒有發現，你可能不會想到教它。

在這項研究中，泰特和她的研究團隊計算出，ChatGPT 打出的論文分數與訓練有素的人類評分者的分數處於「良好」到「中等」的一

致性。在 943 篇論文中，ChatGPT 在 89%的情況下與人類評分者的分數相差不遠。根據研究人員在研究中使用的六分評分標準，ChatGPT 經常給一篇文章 2 分，而人類專家評估員認為該文章實際上是 1 分；但在另外 344 篇英文論文中，這樣的一致性下降到 83%，且在其他 493 篇歷史論文中，更下降到 76%。這意味著在更多情況下，當老師將一篇文章評為 6 分時，ChatGPT 會給其 4 分。這就是為什麼 Tate 說 ChatGPT 僅能用於低風險測驗。

儘管如此，這種準確度仍然令人印象深刻，因為即使是老師對於評分也存在分歧，而且一分差異很常見。人類評分者之間只有一半的情況會出現完全一致的情況，而對於人工智慧來說，情況更糟，它只在大約 40%的情況下與人類評分完全一致。人類更有可能給出 6 分的最高分數或 1 分的最低分數。ChatGPT 則傾向給出中間分數（介於 2 至 5 分間）。

研究人員將人類使用的評分指南（稱為評分標準）的摘錄複製並貼上到 ChatGPT 中，並讓它「假裝」自己是一名老師，並按照 1 到 6 的等級對論文進行評分。

早期的評分機器

早期版本的自動論文評分器具有更高的準確率。但創建它們既昂貴又耗時，因為科學家必須針對每個論文問題使用數百篇人工評分的論文來訓練電腦。這僅在有限的情況下在經濟上可行，例如數千名學生回答相同論文問題的標準化考試。

一旦學生了解了電腦系統評分的功能，早期的評分機器也能被欺騙。在某些情況下，如果在無意義的文章中加入一些花俏的詞彙，就會獲得高分。但 ChatGPT 並非針對特定特徵進行評分，而是分析大量的語言資料庫。泰特表示她還沒有看過 ChatGPT 給一篇無意義的文章高分。

Tate 預計隨著新版本的發布，ChatGPT 的評分準確性將迅速提高。研究團隊已經發現，需要付費訂閱的新 4.0 版本的評分比免費的 3.5 版本更準確。Tate 懷疑對 ChatGPT 的評分說明或提示進行小幅調整可以改善現有版本。她也有意測試如果用已經評分的幾篇（也許是五篇）樣本文章來訓練 ChatGPT，它的評分是否會變得更加可靠。

許多教育科技新創公司，甚至知名的教育資源供應商，現在都在向學校推銷新的人工智慧作文機器人評分機。其中許多都是由 ChatGPT 或其他大型語言模型支援的。

人工智慧供應商的問題

目前，它對個別學生的評分並不那麼準確，但老師想確切了解每個學生的表現。Tate 建議正在考慮使用人工智慧論文評分者的教師和學校領導思考有關準確率的問題：人工智慧評分者和人類評分者對每篇論文的準確一致率是多少？它們彼此相差一分的頻率是多少？

Tate 的下一步研究是學生的寫作在經過 ChatGPT 評分後是否有所改善。她希望老師們嘗試使用 ChatGPT 對學生初稿進行評分，然後看看它是否鼓勵修改，這對於提高寫作水平至關重要。Tate 認為老師可以讓它「幾乎像一個遊戲：我如何提高分數？」

目前還不清楚如果沒有具體的回饋或改進建議，僅憑成績是否會激勵學生進行修改。學生可能會因 ChatGPT 給的低分而灰心並放棄。許多學生可能會忽略機器給出的成績。儘管如此，Tate 表示，有些學生還是不敢向老師展示自己的作品，而看到他們在 ChatGPT 上的分數提高可能正是他們需要的正向回饋。

撰稿人/譯稿人：駐洛杉磯辦事處教育組

資料來源：2024 年 5 月 20 日

The Hechinger Report

https://hechingerreport.org/proof-points-ai-essay-grading/