

Turnitin 的 AI 寫作檢測功能 常見問答 FAQ

- 2 Turnitin 的 AI 寫作檢測功能是如何運作的？
- 6 AI 檢測結果和說明
- 9 檢測的範圍
- 10 使用權限

編譯自英文版《FAQs for
Turnitin' s AI writing detection capabilities》：
<https://bit.ly/3IMNLUH>



Turnitin 的 AI 寫作檢測功能是如何運作的？

1. Turnitin 有無提供偵測 AI 寫作的解決方案？

為幫助教育工作者維護學術誠信，同時確保學生獲得公平對待，Turnitin 已發佈了英文 AI 寫作檢測功能試用。在相似度報告中新增了一個 AI 寫作指標，呈現文章中可能是以 AI 寫作工具，如 ChatGPT 所生成的文字比例。

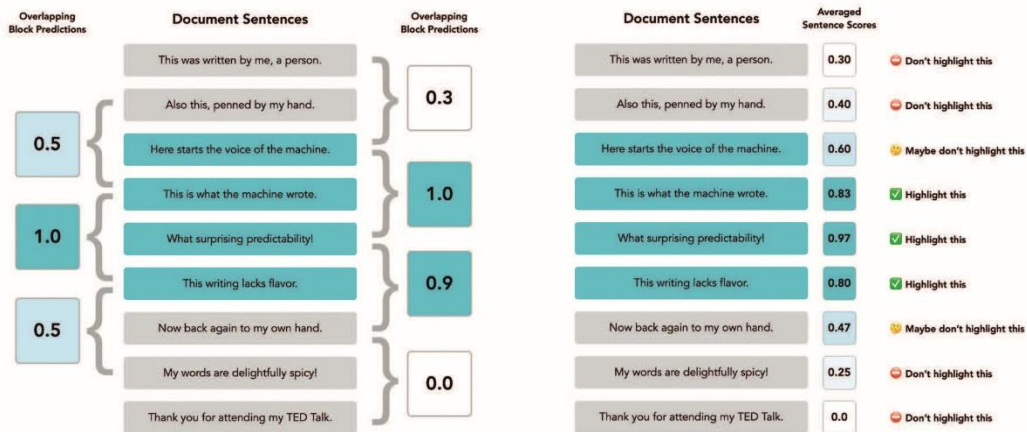
本指標會進一步連結到一份報告，在報告中會以底色套色顯示模型預測由 AI 生成的文字片段。**請注意，只有具有教師帳號權限者和管理者能夠看到這個指標。**

雖然 Turnitin 公司對發展的模型具備信心，但 Turnitin 並不會對不當行為做出判斷，而是做為向教育工作者提供資訊的工具，讓他們根據自己的學術和機構政策做出明智的決定。因此，Turnitin 必須強調，AI 寫作指標的百分比數值無法作為採取行動措施的唯一依據，也不應作為教職員決定性的評分標準。

2. 這項功能如何運作？

當一篇英文文章上傳至 Turnitin 時，會首先被切分成數個數百字的片段（大約五到十個句子）。然後這些片段會被相互重疊，以捕捉每個句子的上下文。





這些句子與 Turnitin 的 AI 檢測模型進行比對，為每一個句子給予 0 到 1 分的分數來確認它是由人類還是 AI 撰寫的。如果模型判定一個句子不是由 AI 生成，這個句子會獲得 0 分。如果模型判定整個句子是由 AI 產生的，這個句子會獲得 1 分。

透過文章所有片段的平均分數，該模型會生成一個整體預測值，即系統認為上傳的文稿中有多少文字是由 AI 生成的（根據 Turnitin 人工智慧創新實驗室收集和驗證的資料，可信度為 98%）。例如，當顯示一篇文章整體文字的 40% 是由 AI 生成，代表認為有 98% 的把握是這樣的情況。

目前，Turnitin 的 AI 寫作檢測模型被訓練用來檢測 GPT-3 和 GPT-3.5 語言模型中的內容，也包括 ChatGPT。Turnitin 正在積極擴大其模型，以利更完善地偵測來自其他 AI 語言模型的內容。

3. 在偵測 AI 寫作時，Turnitin 的模型會考量到哪些參數或標誌？

GPT-3 和 ChatGPT 是使用整個網際網路的文本為訓練對象，它們基本上是利用這些大量的文本，根據選擇下一個高概率單詞來生成詞彙序列。這意味著 GPT-3 和 ChatGPT 傾向以一種一致和高度可能的方式生成詞彙序列中的下一個單詞。

另一方面，人類的寫作往往是不一致和具有異質性的，導致對於人類寫作時在一個詞彙序列中的下一個字要使用什麼的可預測性是很低的。Turnitin 的分類器經過訓練，可以檢測出這些詞彙機率的差異，並熟練掌握人類作者產生的特定詞彙概率序列。

4. Turnitin 的模型是如何訓練的？

Turnitin 公司使用的模型是以具有代表性的資料樣本為訓練對象，其中包括 AI 生成的內容和真實的學術寫作內容。在建立樣本資料集的時候，亦考量到在統計學上代表性較為不足的群體，如第二語言學習者、非英語國家的英語使用者、來自多元招生體制學院和大學的學生，以及較不常見的學科領域如人類學、地質學、社會學等。

5. 我可以檢測過去已上傳文章中是否有 AI 寫作嗎？

以往已上傳的文章並不會自動進行檢測，請重新上傳該篇文章至 Turnitin。只有在 2023 年 4 月 4 日此功能推出之後，上傳至系統的文章才會自動檢查。

6. AI 文字檢測功能支援哪些語言？

目前**僅支援英文**。此為 Turnitin 的 AI 寫作檢測工具的初始試用版本，只能針對長篇的英文文章進行檢測。

7. 如果上傳非英文的文章進行檢測，會發生什麼事？

如您上傳的內容並非英文文章，將不會產生 AI 寫作內容檢測報告。指標將顯示為空白或錯誤狀態，且系統提示會向使用者說明此功能目前只適用於英文文章。



8. 我的機構還不是 Turnitin 用戶，可以提前試用這項新功能嗎？

與一般產品發布不同，因本次 Turnitin 是根據客戶回饋，提前讓這項技術推出面世，故無法提供任何客戶提前試用新功能。然而，Turnitin 已經在實驗室中進行了嚴格的 AI 檢測技術測試，並對結果充滿信心。為了幫助客戶更了解此功能，將提供說明文件和逐步指導以介紹如何使用它。

9. 如果不想看到新的 AI 寫作偵測指標及報告，教師或機構管理者可以隱藏它嗎？

此功能無法隱藏或關閉。在這個初始試用版本，Turnitin 無法為特定使用者關閉 AI 寫作檢測指標或報告。因 Turnitin 公司收到大量來自使用者的積極回饋，希望能盡快推出 AI 寫作檢測技術，以利檢閱學生是否使用 AI 寫作工具，並提供教育工作者深入見解，建置關閉功能可能會延後推出的時間。

為了幫助機構充分利用 AI 檢測工具，將會在系統中設置指引和常見問題頁面連結，解釋各種結果的含義以及該功能的原理。另外和往常一樣，Turnitin 提供了許多[教育資源](#)給教育工作者，以幫助他們了解如何在 AI 時代維護學術誠信。

10. 在相似度報告中增加 AI 檢測功能是否會改變使用系統的方式？

不會。這個額外的功能不會改變您原本使用相似度報告的方式、或您現有的工作流程。AI 檢測功能已自動被加入至到 Turnitin 相似度報告介面中，使用者無需再登入其他系統即可體驗。

11. 是否可以透過支援 LTI 的學習系統，如 Moodle、Blackboard、Canvas 等使用此功能？

是的，教師於原使用的 Moodle、Blackboard、Canvas 等學習系統開啟 Turnitin 相似度報告後，即可在報告視窗介面中查看 AI 檢測指標。



AI 寫作檢測結果和說明

1. AI 寫作檢測指標中的百分比數值代表的涵義為何？

該百分比代表 Turnitin 的 AI 寫作檢測模型判定上傳的英文文稿中，符合特定條件的文字中是由 AI 所生成的比例（根據在受控實驗室環境中精心收集並驗證的數據資料，具有 98% 的準確度）。

符合可分析條件的文字型態僅限散文句子，也就是只分析以標準文法句子所撰寫的文字區塊，不包括其他類型的寫作，例如列表、條列式重點或其他非句子結構。此處呈現的百分比數值未必是整個上傳文稿的比例，因為如果上傳文稿中的內容非屬於長篇散文文字，則不會被計算在內。

2. 顯示的 AI 寫作百分比有時與被底色套色的文字篇幅不一致，這是為什麼呢？

與 Turnitin 的相似度報告不同，AI 寫作百分比並不一定與上傳文稿的文字量有關。Turnitin 的 AI 寫作檢測模型僅會尋找英文長篇寫作中的散文句子。長篇寫作散文意指以個別文句組成的較長作品，例如小論文、碩博士論文、作文文章等。此模型無法檢測 AI 生成的詩句、劇本或程式碼，也不會檢測簡短或非常規的寫作形式例如條列式要點、表格或簡短的考試答案。

3. Turnitin 的 AI 寫作指標的準確度如何？

Turnitin 希望確保不會誤判非由 AI 生成的內容，故只有在 98% 確信某些內容是由 AI 生成時才將其標記為 AI 寫作。然而這表示可能會錯過多達 15% 由 AI 生成的文本；而把真人撰寫的文字誤判為 AI 生成的機率則小於 1%。



例如，如果系統判別一份文件中 50% 的內容是由 AI 寫作，Turnitin 有 98% 的把握至少有 50% 的內容是由 AI 生成，而且誤判率小於 1%。但實際上這篇文章有可能包含多達 65% 的 AI 寫作文字。

上述比例是 Turnitin 的 AI 寫作檢測模型使用在人工智慧創新實驗室中收集和驗證的資料測定的，但他們知道實際會與實驗室測試有所不同，因此調整了 AI 檢測器以利儘量減少對真實文章的誤判，即使這可能會讓偵測器錯過一些 AI 寫作的內容。

4. 如果我覺得 AI 寫作指標不正確，我可以做什麼？Turnitin 的指標如何處理「偽陽性」？

如果您發現遺漏某些由 AI 生成的內容，或者注意到誤將實際由學生寫作的作品判別為 AI 生成，請告訴我們！您的回饋對 Turnitin 進一步改善偵測技術至關重要。您可以藉由 AI 寫作報告中找到的「Feedback」按鈕與他們聯繫。

有時造成偽陽性（將真人撰寫的文字誤判為 AI 生成）的情況可能為：沒有太多句子結構變化的列表文字、字面上重複的內容，或者在沒有增加新觀點的情況下進行改寫的文字。如果在上述的文字型態中呈現較高比例的內容為 AI 撰寫，Turnitin 建議您在查看百分比時考量以上可能性。

在一份混合真人撰寫和 AI 生成文本的長篇文章中，要確定從哪開始是 AI 撰寫的文字、以及到哪邊為止是真人寫作的文字可能很困難，但 Turnitin 的模型應能為您提供可靠的指引，以利您與上傳文章的學生進行對話。

在只有幾百字的短篇文字中，檢測結果將近為「全有或全無」，因為僅能基於一個單一文字片段去偵測，而沒有機會將它與更多同篇文字內容進行重疊。這代表某些由 AI 生成和原創內容混合的文字可能被標記為完全由 AI 生成。在檢閱這些資料並與學生或他人進一步確認之前，請將以上情況納入考量。



5. 學生能看到結果嗎？

僅管理者帳號及教師帳號能查看，學生帳號中無法看到 AI 寫作檢測指標和報告。

6. 相似度百分比和 AI 寫作檢測百分比之間的區別是什麼？

「相似度百分比」與「AI 寫作檢測百分比」是兩項不同的工具，兩者之間並無關聯。

「相似度百分比」指的是進行相似度比對時，在上傳文稿中與 Turnitin 資料庫收錄內容文字相似的比例。另一方面，「AI 寫作檢測百分比」則顯示 Turnitin 的 AI 寫作檢測模型預測上傳文稿中由 AI 寫作工具生成文字的比例。

7. Turnitin 的模型是否考慮到 AI 寫作檢測技術可能對特定學科領域或第二語言的寫作者有偏見？

Turnitin 公司及 AI 團隊的指導原則之一就是要盡量減少可能為學生帶來傷害的風險，尤其是那些受到歷史和社會結構所壓迫或排斥的弱勢學生。因此在建立樣本資料集的時候，亦考量到在統計學上代表性較為不足的群體，如第二語言學習者、非英語國家的英語使用者、傳統黑人學院或大學的學生，以及較不常見的學科領域如人類學、地質學、社會學等。

8. 我如何在課堂上與學生一起使用 AI 寫作檢測工具？

Turnitin 的 AI 寫作檢測指標代表可能由 AI 寫作工具生成文字的百分比，而報告中則會顯示哪些是由 AI 生成的文字段落。是否具有任何不當行為的最終決定權取決於教師或審閱者。此工具並不會判定文稿中是否包含不當行為，而是提供教育者所需的資料，由他們根據其學術和機構政策做出明智的決定。



檢測範圍

1. Turnitin 的技術可以檢測哪些 AI 寫作模型？

Turnitin 的 AI 寫作檢測功能的第一個版本已經被訓練為能夠檢測 GPT-3、GPT-3.5 以及其變種的模型。此技術還可以檢測其他基於這些模型的 AI 寫作工具，例如 ChatGPT。Turnitin 未來亦計畫擴展檢測能力至其他模型。

2. 目前的模型能夠檢測 GPT-4 生成的文本嗎？

Turnitin 正在不斷改進和擴展 AI 寫作檢測能力。目前 AI 團隊正在使用現有的檢測器對 GPT-4 進行測試，以比較其表現並了解 GPT-3.5 (目前的模型所訓練的對象) 和 GPT-4 之間的差異，初步結果已能夠準確地檢測 AI 生成的文本。

Turnitin 的研究分析正在持續進行中，一旦建立可靠有效的指標，將會更新納入 GPT-4 至其模型。然而需要留意的是，目前 ChatGPT 的免費版本仍使用 GPT-3.5。

3. Turnitin 將如何為未來更高階版本的 GPT 和其他尚未出現的大型語言模型做好準備？

Turnitin 已知大型語言模型 (LLM) 正在快速擴展和演進，因此已致力開始構建其它的檢測系統。其初始的著眼點在於為 GPT-3 和 GPT-3.5、以及其他基於這些模型的寫作工具 (如 ChatGPT) 建立和發佈一個有效且可靠的 AI 寫作檢測器。

4. Turnitin 能否檢測 AI 寫作工具 (ChatGPT 等) 生成的文字是否有使用改寫工具(Paraphrasing Tool)進行進一步改寫？即使在這種



情況下，它是否會將內容標記為 AI 生成？

Turnitin 的檢測器是根據 GPT-3、GPT-3.5 和 ChatGPT 的輸出結果進行訓練，修改這些系統生成的文字會對檢測器識別 AI 寫作文本的能力產生影響。在人工智慧創新實驗室中，Turnitin 使用開源的改寫工具（包括不同的大型語言模型）進行測試，在大多數情況下檢測器保持有效性，即便使用改寫工具改變 AI 輸出文字的情況下，也能識別文字是 AI 生成的。

5. Turnitin 是否有計劃推出檢測學生自行改寫或透過 Quillbot 等工具改寫文字內容的工具？

Turnitin 投注心力於開發偵測改寫功能已有一段時間了 – 也就是檢測學生是否使用改寫工具或自行重寫文字，這樣的技術在 AI 創新實驗室中已經達到期望成果。當學生使用詞匯重組工具或線上改寫工具時，是利用詞匯重組工具來替換字詞，這類工具使用 AI 刻意阻礙相似度比對，而不是使用生成式 AI 工具（如 ChatGPT）來創建內容。Turnitin 計劃在 2023 年發布此工具的測試版本，需另付費加購。

使用權限

1. 誰能使用英文 AI 寫作檢測工具？

即日起至 2023 年 12 月 31 日，教師版帳號及管理者帳號可試用英文 AI 寫作檢測功能，以測試此工具並了解它如何運作。2024 年 1 月 1 日起是否仍能繼續使用此功能尚無法得知。



2. 何時可以使用這個工具？

現在即可開啟相似度報告視窗，開始使用 Turnitin 的英文 AI 寫作檢測功能。使用上述 Turnitin 產品的用戶應該能夠看到該指標並查看報告。

3. Turnitin 的 AI 寫作檢測工具是一個獨立的解決方案，還是其他產品的一部分？

Turnitin 的初始版本 AI 寫作檢測是相似度報告中的一個獨立功能，該指標會呈現一份報告，顯示上傳的英文文稿中偵測為 AI 生成的文字段落。

4. 為什麼 AI 寫作檢測沒有在其他 Turnitin 產品例如 Gradescope 和 iThenticate 中推出？

Turnitin 把資源集中在最大、最急迫的問題 - 高等教育和 K-12 的長篇作文之上，並正在研究如何向 iThenticate 使用者推出 AI 寫作檢測。目前並沒有計劃將這些功能添加到 Gradescope 中，因為 Gradescope 主要應用於手寫試卷作業，而 AI 文字檢測是著重於數位形式的文字內容。然而他們很高興能在這個產品中了解更多關於使用者對於 AI 寫作檢測的需求。另外目前開發的功能中並不包含 ChatGPT 程式碼檢測。

5. 在哪裡可以找到更多關於這個新工具的資訊？

您可以在[這個頁面](#)上找到關於 Turnitin 的 AI 寫作檢測功能的資訊。

