



The Use of Generative Artificial Intelligence Chatbots as a Learning Tool for Students in Medical Imaging and Radiological Sciences

I-Lin Wu¹, Malcolm Koo²

¹Radiology Diagnostic Center, Taitung MacKay Memorial Hospital, Taitung, Taiwan

²Department of Nursing, Tzu Chi University of Science and Technology, Hualien, Taiwan

Key words

artificial intelligence,
chatbots,
learning tools

ARTICLE INFO

Article history

Received March 10, 2023

Accepted December 6, 2023

Corresponding author

Malcolm Koo, Ph.D., Department of
Nursing, Tzu Chi University of Science
and Technology No. 880, Sec. 2, Chien-
kuo Road., Hualien City, Hualien County
970, Taiwan, R.O.C.

Tel: +03-8572158 ext.2206

E-mail: mkoo@ems.tcust.edu.tw

© SNM&AINOSCO PRESS

DOI:10.6332/ANMMI.202406_
37(2).0002

<http://www.ainoscopress.com>

<http://www.airitilibrary.com/>

Publication/alPublicationJournal?Pu
blicationID=P20141124001

ABSTRACT

Background: ChatGPT, an artificial intelligent chatbot with powerful language generation capabilities for simulating human-like text, was launched in late 2022 by OpenAI, an American artificial intelligence research company, and immediately became a global buzzword. ChatGPT's academic performance has also sparked discussion and analysis, but there is currently no research analyzing ChatGPT's performance in examinations based on Chinese language. Therefore, this study used GPT3.5 and GPT-4 to answer questions from the medical radiation technologist examination in Taiwan and analyzed its exam scores.

Methods: A total of 240 questions were selected from six subjects in the medical radiation technologist examination using the 2022 Second Senior Professional and Technical Examinations for Medical Personnel, and excluding questions that required answering based on image. The questions were then formatted and tested using ChatGPT application programming interface (API). The answers published on the Examination Yuan's examination platform were used as the criteria for determining correct answers.

Results: GPT-3.5 and GPT-4 were able to obtain a score of 39.4% and 68.1%, respectively. There were differences in performance across different subjects, which may be due to a lack of understanding of the questions and the ability to answer them. Therefore, for the time being, ChatGPT can be used for tasks such as translation, grammar correction, article summarization, and simplifying radiographic reports. However, any article revised through ChatGPT must be validated by human authors to ensure its correctness.

Conclusions: The development and application of artificial intelligence technology are inevitable trends that should be addressed positively and ethically to achieve a win-win situation.

1. 前言

人工智能 (artificial intelligence, AI) 一詞自 1950 年代提出來，受限於電腦運算效能，並未蓬勃發展。直至運用深度神經網路 (deep neural networks) 的 AlphaGo 在 2016 年戰勝韓國職業九段棋士李世乭後 [1,2]，才掀起大眾對 AI 的關注。2022 年 11 月底，美國 AI 研究公司 OpenAI 正式對外釋出 ChatGPT 聊天機器人測試版 [3]，提供公眾免費試用。由於 ChatGPT 具備強大模擬生成人類的語言文本的能力，應用場域貼近生活，可以翻譯語言、創作文學詩詞、摘要文本、提供關鍵字、回覆書信郵件及編寫程式碼等。因此，ChatGPT 服務於上線五天已累積 100 萬名用戶 [4]，更在上線 2 個月內突破 1 億用戶 [5]。微軟創辦人比爾蓋茲認為 ChatGPT 的發展重要性堪比網際網路和個人電腦的誕生 [6]。

當然，ChatGPT 的出現，亦帶來不少疑慮和道德上的爭議，不良份子可濫用作為詐騙工具 [7]、生成虛假學術論文 [8,9]、學生可能利用 ChatGPT 作弊或撰寫報告論文 [10]，這無疑對以產出文字為主的相關工作，如記者、撰稿人、翻譯人員等，帶來挑戰 [11]。而教育界方面的意見，則較為兩極。部分對 ChatGPT 持保留態度，認為學生可能養成過度依賴數位科技的情況，因而削弱基本語言表達和閱讀的能力。國外的一些教育部已經宣布禁止轄下學校的電子裝置和網路使用 ChatGPT [12]。同時，辨識 AI 生成文章的偵測工具已經相繼出現，如 OpenAI 所開發的 AI Text Classifier (<https://platform.openai.com/ai-text-classifier>) 和 GPTZero (<https://gptzero.me/>)，但其效用仍有待證實。

另一方面，亦有部分教育界人士對 AI 工具持接納態度，認為有必要將數位科技引入教育，讓學生掌握運用數位科技的能力，以配合數位化的時代，並應對未來的職場與生活需要。ChatGPT 可以作為一種輔助工具，幫助學生快速搜集資料和理清思路。此外，它可以提供個別化學習和即時反饋，讓學生自行安排學習時間和進度。當然，必須加強學生的自主學習能力和對資訊的分辨、思考和查證能力。況且要完全禁止學生在校外使用 AI 工具，更是難以實行，猶如過去維基百科的

爭議 [13]。使用 ChatGPT 本身並不同學不端的不行為，培養學生的學術誠信意識和強調原創思考亦是教育的關鍵。

ChatGPT 因其基於監督和強化學習的訓練方式和大型的預訓練數據集，具有優異的自然語言處理能力。它是基於 GPT-3.5 架構的大型語言模型上，再透過人類回饋的監督學習和強化學習微調。其大型語言模型的預訓練數據量達到 45TB，使用的參數量來到 1,750 億 [14]。GPT 是生成型預訓練變換模型 (Generative Pre-trained Transformer) 的英文縮寫，顧名思義，通過大量文本數據，包括互聯網、書籍文章、社交媒體、線上討論區等，進行預先監督式訓練，建立語言的知識和模式，並在後續任務中生成連貫且有上下文關聯的文本回應。有別於過去的聊天型 AI，變形器架構使 ChatGPT 能夠對輸入提示生成連貫且有上下文關聯的文本回應，因此，能夠生成連貫的回應。但 ChatGPT 也存在多種局限，首先 ChatGPT 是基於統計模型進行文本生成的，它缺乏常識和推理能力，在某些缺乏良好訓練數據的領域，表現就不如理想。此外，ChatGPT 也可能出現「人工智能幻覺」(artificial hallucination) 的情況，即回覆雖然看似合理亦充滿自信，但其實不正確或毫無意義 [15]。因此，使用者對 ChatGPT 的回應，須保持「信任，但要查證」(trust but verify) 的態度。再者，由於訓練時所用的資料庫目前僅更新至 2021 年，因此在此以後發生的事情，ChatGPT 就缺乏可以生成回覆的資料。目前的一個解決方法是透過瀏覽器的外掛套件，如 WebChatGPT，透過即時相關網頁搜索的結果，整合回覆到 ChatGPT 中。OpenAI 於 2023 年初進一步發表全新、推理能力更強及處理敏感問題表現最佳的 GPT-4 語言模型與「應用程式介面」(application programming interface, API)。

近期國外研究顯示，ChatGPT 在學術表現上有著不同的成績，引起了教育界的廣泛討論。一項分析 ChatGPT 是否能通過美國醫學執照考試 (United States Medical Licensing Exam) 的研究發現，ChatGPT 得分接近 60% 的及格線 [16]。另外一項研究則採用在四個不同的相關題庫探討 ChatGPT 對美國醫學執照考試第一和第二階段

考題上的表現，結果顯示該模型達到三年級醫學生的及格分數 [17]。在賓夕法尼亞大學沃頓商學院的營運管理課程期末考中，ChatGPT 的成績介於 B 和 B- 的水平，在高階分析問題上表現略遜 [18]。同樣地，ChatGPT 在明尼蘇達大學法學院的測試中，在憲法、聯邦制、雇傭和稅法四科的期末考中獲得相當於 C+ 成績水平，在申論題表現比選擇題好，但在有數學運算的選擇題則表現最差 [19]。然而，ChatGPT 並沒有通過美國心臟協會 (American Heart Association) 的基本救命術 (Basic Life Support) 和高級心臟救命術 (Advanced Cardiac Life Support) 考試，儘管如此，該文作者仍然認為 ChatGPT 對答案的解釋相當深入 [20]。此外，ChatGPT 挑戰號稱全球最難國家考試，錄取率只有 5% 的印度公務員考試，結果在 100 題當中只答對 54 題，結果亦不及格 [21]。至於在放射學國家考試方面的表現，一項在加拿大的研究，以 150 選擇題評估 ChatGPT 的表現。結果發現 ChatGPT 能正確回答 69% 的問題，該模型在要求低階思維的問題上的表現顯著優於要求高階思維的問題，而在涉及影像結果描述、計算和分類以及概念應用的高階思維問題上表現較差 [22]。該文作者並指出，放射科醫師須清楚 ChatGPT 的局限性，特別是它會傾向過於自信地回覆錯誤的答案。目前在臨床和教育上的應用，仍不能完全依賴 ChatGPT。整體而言，雖然 ChatGPT 在應付專業考試方面仍有很多進步空間，但部分學者仍對大型語言模型能夠助益學生學習持正面的看法 [23]。上述研究皆為使用英文進行考試的分析，目前尚未有探討

ChatGPT 在中文考試表現的相關研究。因此，本研究使用 ChatGPT 回答醫事放射師考試，分析其考試成績，並探討學生是否可以利用 ChatGPT 幫助學習。

2. 材料與方法

本研究利用民國 111 年第二次專門職業及技術人員高等考試，從醫事放射師考試 6 個科目，各 80 題，共 480 選擇題，從中選取奇數題目共 240 題，並排除需要根據圖像作答的題目。整理格式後，於 2023 年 5 月 9 日，透過 ChatGPT API，分別針對 GPT-3.5 和 GPT-4 模型進行測試。所得結果利用 Microsoft Excel，首先輸入考試院考選部考畢試題查詢平臺上公布的正確答案，然後利用公式自動判別 ChatGPT 的回答是否正確。

3. 結果

奇數題目共 240 題，排除當中 14 題需要根據圖像作答的題目，共分析 226 題。結果如 Table 1 所示，GPT-4 模型在 6 個考試科目中，有兩科（放射治療和醫學物理）未達到 60% 的及格水平，而最高分數的為基礎醫學的 87.5%，整體而言，分數為 68.1%，達到 60% 的及格水平。但 GPT-3.5 模型的表現在三個科目及整體分數則顯著較差，沒有任何一個的分數達到及格水平，最高分數的為放射器材的 51.3%，整體而言，分數為 39.4%，未達到 60% 的及格水平。

Table 1. The performance of GPT-3.5 and GPT-4 models in the six subjects of the medical radiation technologist examination

科目	題目數	答對題目數（百分比）		p 值
		GPT-4	GPT-3.5	
核子醫學	35	22 (62.9)	9 (25.7)	0.002
基礎醫學	40	35 (87.5)	19 (47.5)	< 0.001
放射診斷	34	23 (67.6)	13 (34.2)	0.088
放射治療	38	19 (50.0)	13 (34.2)	0.166
醫學物理	40	23 (57.5)	12 (30.0)	0.014
放射器材	39	32 (82.0)	20 (51.3)	0.230
總計	226	154 (68.1)	89 (39.4)	< 0.001

4. 討論

專門職業及技術人員高等考試是一項由考選部主辦的全國性考試，旨在評估參加考試者的專業知識和實務技能，以確保其符合相應職業的標準。本研究發現，目前 GPT-3.5 在醫療放射技術之相關方面的回覆，仍未能達到從事醫事放射師工作的及格水平，而 GPT-4 雖然總分有達及格水平，但放射治療和醫學物理的得分仍不到 60%。造成這個結果有幾個可能的因素。首先，雖然 ChatGPT 擁有優異自然語言表現，但對於相對簡短的選擇題問題，未必能在缺乏「情境」(context) 的脈絡下，充分瞭解題目的意思。事實上，使用明確而且具有意義的提示詞 (prompt) 才能與語言模型溝通，而充分發揮 ChatGPT 的功能和潛力。因此，衍生出「提示工程」(prompt engineering) 的新技能和職位 [24]。舉例說，放射治療考試其中的一個問題是：「根據 Withers 在 1985 年的敘述，晚期反應組織受何種因素影響較大？(A) 分次劑量大小、(B) 總照射時間、(C) 劑量率、(D) 總劑量」。GPT-3.5 和 GPT-4 皆回覆錯誤的答案 D。並且提供錯誤解釋：「晚期反應組織主要受總劑量的影響。高總劑量可能對正常組織造成較大的晚期副作用。在放射治療中，醫生需在控制腫瘤和降低晚期副作用之間取得平衡，通過調整劑量分配、照射時間和劑量率等參數來達到治療目標。」由此可見，ChatGPT 會可能提供一些看似頭頭是道，但其實暗藏似是而非的內容而誤導學生。

此外，生成文本的素質與背後訓練資料的質量和數量息息相關，GPT-3 的訓練數據主要是英文（以字數計算占 93%）、其他語言只有 7% [14]。在本文中，作者向 ChatGPT 提問：「請問臺灣考選部主辦的專門職業及技術人員高等考試，全國性的醫事放射師，它的考試科目包括哪 6 科？」ChatGPT 回覆為基礎醫學、醫用放射學、放射病理學、放射線生物學、放射線防護學及專業討論。顯示這方面的資料可能不足。若是使用英文提問，其結果亦是錯誤。回到前面的題目，慢速電腦斷層掃描的問題，當利用 WebChatGPT 擴充套件，由 ChatGPT 整合網頁搜索的結果，回覆則正確：「AAPM TG-76 的報告中提到，慢速電腦斷層 (slow

computed tomography [CT]) 是一種影像模擬攝影技術，用於減輕呼吸運動對影像質量和治療計劃的影響。報告中指出，慢速電腦斷層技術僅適用於肺部的模擬攝影 (simulation CT)，尤其是對於呼吸運動影響較大的肺部腫瘤模擬攝影。」這個結果顯示在專業的部分，ChatGPT 仍有明顯盲點，仍須等待未來更龐大的訓練資料庫提高其智能水平。事實上，ChatGPT 並非真正理解人類語言的含義，它只是基於神經網路架構生成文字。就如 Open AI 的主席 Sam Altman 認為，現在將重要事務依賴於它是一個錯誤 [25]。

在研究限制方面，本研究的結果只能代表 ChatGPT 基於在發問當天的能力。由於 ChatGPT 會定期更新，它除了使用大型語言模型的預訓練數據外，亦透過「人工回饋的強化式學習」(reinforcement learning from human feedback) 由人工標記員調教，以降低提供偏見或錯誤結果的機率。此外，ChatGPT 透過「一次性學習」(one-shot learning) 的方式讓它可以根據使用者前次的文字即時調整回應訊息。因此，本研究在連續詢問問題的情況下，ChatGPT 可能會因為前面題目的結果間接影響後面的作答。如果 ChatGPT 的回應能夠引起人類用戶的積極回應，那麼 ChatGPT 就會獲得積極的獎勵回饋。反之，ChatGPT 就會受到負面的獎勵回饋。通過這種方式，ChatGPT 不斷學習如何生成更加自然和有用的對話回應。再者，本研究只有針對分析 ChatGPT 在技術層面上的表現，至於學生對融入 ChatGPT 在傳統學習的可用性 (usability) 和成效仍需要未來研究以確定。最後，本研究只有探討 ChatGPT 生成式人工智慧聊天機器人，其他類似功能的聊天機器人的表現，如 Google 的 Bard [26]，仍需要後續的研究。

綜合而言，目前的 ChatGPT 仍未能針對學生準備國家考試，提供在解題上的協助，這方面仍然有待更強大的自然語言處理能力和訓練資料庫的支持。在目前的階段，ChatGPT 可以用於協助翻譯 [27]、矯正拼字和文法錯誤、改善文本的流暢度等。此外，亦可以創建文章的摘要，讓文章主要內容易於理解，甚至簡化放射診斷報告 [28]。當然，透過 ChatGPT 所修改的文章，必須經由作

者確認其正確性才能符合學術倫理 [29]。AI 技術如 ChatGPT 的發展和應用已經是必然之勢，我們應該積極面對，在符合社會的需求和倫理價值下，與 AI 協作，促進人類的福祉，實現雙贏的局面。

參考文獻

- Toosi A, Bottino AG, Saboury B, Siegel E, Rahmim A. A brief history of AI: how to prevent another winter (a critical review). *PET Clin* 2021; 16:449-469. doi:10.1016/j.cpet.2021.07.001
- Gibney E. Google AI algorithm masters the ancient game of Go. *Nature* 2016;529:445-446. doi:10.1038/529445a
- ChatGPT. OpenAI. <https://chat.openai.com/chat>. Updated Feb. 13, 2023. Accessed Oct. 14, 2023.
- Altman S. ChatGPT launched on Wednesday. today it crossed 1 million users! Twitter. <https://twitter.com/sama/status/1599668808285028353>. Updated Dec. 5, 2022. Accessed Oct. 14 2023.
- Milmo D. ChatGPT reaches 100 million users two months after launch. *The Guardian*. <https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app>. Updated Feb. 2, 2023. Accessed Oct. 14, 2023.
- Kay G. Bill Gates calls ChatGPT “every bit as important as the PC” or the internet. *Business Insider*. <https://www.businessinsider.com/bill-gates-chatgpt-ai-artificial-intelligence-as-important-pc-internet-2023-2>. Updated Feb. 3, 2023. Accessed Oct. 14, 2023.
- Patel A, Sattler J. Creatively malicious prompt engineering. <https://labs.withsecure.com/content/dam/labs/docs/WithSecure-Creatively-malicious-prompt-engineering.pdf>. Updated Jan. 2023. Accessed Oct. 14, 2023.
- Else H. Abstracts written by ChatGPT fool scientists. *Nature* 2023;613:423. doi:10.1038/d41586-023-00056-7
- Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature* 2023;614:214-216. doi:10.1038/d41586-023-00340-6
- Stokel-Walker C. AI bot ChatGPT writes smart essays—should professors worry? *Nature* 2022. [Online ahead of print] doi:10.1038/d41586-022-04397-7
- Castelvecchi D. Are ChatGPT and AlphaCode going to replace programmers? *Nature* 2022. [Online ahead of print] doi:10.1038/d41586-022-04383-z
- Johnson A. ChatGPT in schools: here’s where it’s banned—and how it could potentially help students. *Forbes*. <https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/?sh=67259f476e2c>. Updated Jan. 18, 2023. Accessed Oct. 14, 2023.
- Anderson N. Banning Wikipedia at school: good idea or missed opportunity? *Ars Technica*. <https://arstechnica.com/uncategorized/2007/11/banning-wikipedia-at-school-good-idea-or-missed-opportunity>. Updated Nov. 28, 2007. Accessed Oct. 14, 2023.
- Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 2020;33:1877-1901. doi:10.48550/arXiv.2005.14165
- Alkaiissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 2023;15:e35179. doi:10.7759/cureus.35179
- Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198. doi:10.1371/journal.pdig.0000198
- Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312. doi:10.2196/45312
- Terwiesch C. Would Chat GPT get a Wharton MBA? A prediction based on its performance in the operations management course. Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, 2023. <https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24.pdf>. Accessed Oct. 14, 2023.
- Choi JH, Hickman KE, Monahan A, Schwarcz DB. ChatGPT goes to law school. *Journal of Legal Education* 71:387-400. doi:10.2139/ssrn.43

35905

20. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma MJ. Can ChatGPT pass the life support exams without entering the American Heart Association course? *Resuscitation* 2023;185:109732. doi:10.1016/j.resuscitation.2023.109732
21. Bordoloi P. When ChatGPT attempted UPSC exam. *Analytics India Magazine*. <https://analyticsindiamag.com/when-chatgpt-attempted-upsc-exam/>. Updated Feb. 28, 2023. Accessed Oct. 14, 2023.
22. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023;307:e230582. doi:10.1148/radiol.230582
23. Lourenco AP, Slanetz PJ, Baird GL. Rise of ChatGPT: it may be time to reassess how we teach and test radiology residents. *Radiology* 2023;307:e231053. doi:10.1148/radiol.231053
24. McFarland A. What is prompt engineering in AI & why it matters. *Unite.AI*. <https://www.unite.ai/what-is-prompt-engineering-in-ai-why-it-matters/>. Updated Feb. 8, 2023. Accessed Oct. 14, 2023.
25. Altman S. Sam Altman on Twitter. *Twitter*. <https://twitter.com/sama/status/1601731295792414720>. Updated Dec. 11, 2022. Accessed Oct. 14, 2023.
26. Pichai S. An important next step on our AI journey. *Google Blog*. <https://blog.google/technology/ai/bard-google-ai-search-updates/>. Updated Feb. 6, 2023. Accessed Oct. 14, 2023.
27. Jiao W, Wang W, Huang J, Wang X, Tu Z. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv* 2023. doi:arXiv:2301.08745
28. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol* 2023. [Online ahead of print] doi:10.1007/s00330-023-10213-1
29. Koo M. The importance of proper use of ChatGPT in medical writing. *Radiology* 2023; 307:e230312. doi:10.1148/radiol.230312



生成式人工智慧聊天機器人作為醫學影像暨放射科學系學生的學習工具

吳宜霖¹ 辜美安²

¹ 台東馬偕紀念醫院 放射科醫學影像中心

² 慈濟科技大學 護理系

摘要

背景：美國人工智能研究公司 OpenAI 於 2022 年底推出具備強大模擬生成人類的語言文本能力的人工智能聊天機器人 ChatGPT，即時成為全球熱門話題。ChatGPT 在學術表現方面，亦引起討論和分析，然而目前尚未有分析 ChatGPT 在中文考試表現的相關研究。因此，本研究分別使用 GPT-3.5 和 GPT-4 模型回答醫事放射師考試，分析其考試成績。

方法：本研究利用民國 111 年第二次專門職業及技術人員高等考試，從醫事放射師考試 6 個科目，從中選取基數題目共 240 題，並排除需要根據圖像作答的題目。整理格式後，透過 ChatGPT 「應用程式介面」（API）方式進行測試。根據考試院考選部考畢試題查詢平臺所公布的答案作為判定正確的準則。

結果：GPT-3.5 和 GPT-4 模型在醫事放射師考試的整體得分分別為 39.4% 和 68.1%。在不同的科目的表現也有明顯的差別，可能是因為其在理解題意和回答問題方面仍有不足。因此，目前而言，ChatGPT 可以應用於翻譯、文法糾正、文章摘要和簡化放射診斷報告等。但是，透過 ChatGPT 修改的文章都必須經過作者的確認，以確保其正確性。

結論：人工智能技術的發展和應用已是必然之勢，應該積極面對，並遵循倫理原則，以實現雙贏的局面。

關鍵詞：人工智慧、人工智能、學習工具、聊天機器人