

## ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review

Philip Newton & Maira Xiromeriti

To cite this article: Philip Newton & Maira Xiromeriti (17 Jan 2024): ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review, Assessment & Evaluation in Higher Education, DOI: [10.1080/02602938.2023.2299059](https://doi.org/10.1080/02602938.2023.2299059)

To link to this article: <https://doi.org/10.1080/02602938.2023.2299059>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 17 Jan 2024.



[Submit your article to this journal](#)



Article views: 2201





[View related articles](#)



[View Crossmark data](#)

# ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review

Philip Newton  and Maira Xiromeriti 

Swansea University Medical School, Swansea University, Swansea, UK

## ABSTRACT

Media coverage suggests that ChatGPT can pass examinations based on multiple choice questions (MCQs), including those used to qualify doctors, lawyers, scientists etc. This poses a potential risk to the integrity of those examinations. We reviewed current research evidence regarding the performance of ChatGPT on MCQ-based examinations in higher education, along with recommendations for how educators might address challenges and benefits arising from these data. 53 studies were included, covering 114 question sets, totalling 49014 MCQs. Free versions of ChatGPT based upon GPT-3/3.5 performed better than random guessing but failed most examinations, performing significantly worse than the average human student. GPT-4 passed most examinations with a performance that was on a par with human subjects. These findings indicate that all summative MCQ-based assessments should be conducted under secure conditions with restricted access to ChatGPT and similar tools, particularly those examinations which assess foundational knowledge.

## KEYWORDS


Artificial intelligence; academic integrity; cheating; evidence-based education; MCQs; pragmatism

## Introduction

If a university cannot provide a reasonable guarantee that an assessment measures the learning of a named student, then the basic legitimacy of that university is undermined. Many assessments use multiple choice questions (MCQs), where students are presented with a problem scenario and are asked to select the single best answer from 4-5 options. MCQs are objective, having a correct answer, which can be revealed immediately if desired, thus giving instant feedback on learning to students and educators. MCQs can offer a broad coverage of the curriculum and, if written appropriately, can assess higher order learning (Newton 2023a), and so are used for professional licensing examinations for doctors, lawyers, social workers and others.

ChatGPT is an artificial intelligence (AI) 'chatbot', whose underlying architecture is a large language model (LLM) known as a generative pre-trained transformer (GPT). At time of writing (November 2023) a free version of ChatGPT uses a GPT called 'GPT3.5' which is an updated version of GPT3 (OpenAI 2023a). An updated subscription-only version of ChatGPT running GPT-4 was released in March 2023 (OpenAI 2023b). Henceforth, versions of ChatGPT running GPT-4 will be referred to as ChatGPT(4), earlier versions as ChatGPT(3).

**CONTACT** Philip Newton  [p.newton@swansea.ac.uk](mailto:p.newton@swansea.ac.uk)

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/02602938.2023.2299059>.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group  
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Early indications are that ChatGPT will significantly disrupt education systems (Farazouli et al. 2023). Initial media reports suggested that ChatGPT(3) could pass university MCQ-based examinations in law, medicine and business (e.g. Hammer 2023). This raised concerns about the fraudulent use of ChatGPT to complete MCQ-based examinations, particularly online examinations, which are already subject to significant misconduct (Newton and Essex 2023).

These concerns were heightened when OpenAI released ChatGPT(4). The accompanying technical document appeared to show a significant improvement on standardized tests in USA post-16 education, most of which are based upon MCQs (OpenAI 2023b). However, the findings have been criticized for a lack of transparency, and a potential contamination of the test materials with questions that could have been found in the ChatGPT training materials (Narayanan and Kapoor 2023).

This uncertainty over the power of ChatGPT, and the potential implications, has been reflected in extremes of media coverage. Some have warned of a forthcoming 'apocalypse' in assessment (Mollick 2023), while others propose that the concerns are largely unfounded and that ChatGPT represents more opportunities than threats (Ceres 2023). This disparity reflects an urgent need to fully understand the performance of ChatGPT on existing assessments.

A scoping review is a broad exploratory approach, aimed at quickly characterizing and understanding a new field (Tricco et al. 2018). We used the pragmatic research paradigm to inform the design and execution of a scoping review. Pragmatic research prioritises findings which are of practical use, and the asking of research questions designed to generate such findings (Kaushik and Walsh 2019), which are then more relevant when generating evidence-based policy (Newton, Da Silva and Berry 2020). Here we are interested in the performance of ChatGPT on assessments in higher education, from the perspective of future assessment design. Given the novelty of the topic, the speed with which it is developing and the proposed 'assessment apocalypse', we also analysed the broader perspective of the authors of the reviewed studies, since they will have had the chance to critically reflect on the implications of their findings.

This scoping review addresses the following research questions:

1. How does ChatGPT perform on MCQ-based examinations in higher education, including
  - a. Is there a difference in the performance of ChatGPT(3) and ChatGPT(4)?
  - b. Comparison to the pass-mark
  - c. Comparison to the average human student
2. How can we mitigate the challenges to assessment security posed by ChatGPT?
3. How can we harness the positive potential of ChatGPT for higher education?

## Methods

The study was conducted according to the PRISMA-extension protocol for scoping reviews (Tricco et al. 2018).

### *Information sources and search strategy*

ChatGPT is a very new research topic; our research questions did not exist in November 2022. The average time to publish an academic paper is over a year (Björk and Solomon 2013) and so the majority of the research on ChatGPT is currently available only as preprints. We used Google Scholar to identify relevant papers, since it provides the broadest coverage of preprints and the earliest indication of papers citing them (Haddaway et al. 2015; Wang, Glänzel and Chen 2020). Unfortunately, the limited user-interface for Google Scholar means it is not currently possible to report metrics that might be available from traditional academic databases, such as the numbers

of unique search results returned from each term, although the numbers were generally low given the novelty of the topic.

The following search terms were used; 'ChatGPT' AND 'MCQ', 'ChatGPT' AND 'multiple choice', 'ChatGPT' AND 'exam', 'ChatGPT' AND 'examination', 'ChatGPT' AND 'single best answer'. Searching was undertaken by both authors. We also searched the reference lists of included papers, and for papers citing included papers. Searches were conducted up until July 20 2023 and results were identified manually by scrolling through all search results.

### ***Inclusion criteria***

- Tested ChatGPT using MCQs in a summative assessment, meaning that the results carry course credit and/or are used for qualification/admissions decisions. These could be in-person or online examinations, proctored or unproctored.
- higher education or above
- study published in English language, although the examination itself may be in a different language
- zero-shot testing (i.e. copying the question verbatim into ChatGPT with no extra prompts)

### ***Data charting process; quantitative data***

All quantitative data were extracted by one author (PN). A majority of the extraction (covering 57.7% of items) was checked by a second author (MX). No discrepancies were found.

### ***Quantitative data items***

Where possible the following items were extracted for each iteration of each examination.

- **Sample size.** Number of questions tested.
- **ChatGPT performance.** Number/percentage of questions answered correctly by ChatGPT.
- **Pass Mark.** The mark required for a human student to achieve a 'pass'. If no pass mark was reported, then we attempted to identify it from external sources as reported in S1.
- **Human Performance.** The average and percentiles range of scores achieved by human examinees. One study (Wood et al. 2023) included a range of different question formats from a dataset of over 27000 items from multiple examinations: >80% of these were MCQs (N=22004), but student performance was not broken down by question format, and so the average score across all assessment types (76.7%) was used. For calculating Biomedical Admissions Test (BMAT) percentiles, we used the conversion table published by Medify (2023), and then approximated percentiles from the BMAT website (Cambridge Admissions 2021)
- **Language of testing.** Unless otherwise specified this was recorded as 'English'.
- **Year the questions were published/generated**
- **Contamination check.** We determined whether study authors took steps to ensure that the MCQs evaluated in their study were not part of the ChatGPT training materials. This could include writing *de novo* questions, providing assurances that the test questions were not in the public domain prior to Sept 2021, or directly testing for contamination using a memorization effects levenshtein detector (MELD) test (Nori et al. 2023), wherein each test item is split in half, with ChatGPT shown the first half and then asked to generate the second half itself. The generated second half is then compared to the original

second half. High similarity scores are indicative of the test content having been in the training materials. Studies were default scored as 'no' if they did not report such steps and/or where their questions were in the public domain prior to Sept 2021.

### **Summary measures**

We are not aware of any previous reviews on this topic and so these measures were agreed between the authors and informal discussions with colleagues. We calculated the performance of ChatGPT, expressed as (a) the percentage of questions answered correctly, (b) the average of the percentage correct, from each study (c) the mean difference between the score of ChatGPT and the pass mark, (d) the mean difference between the score obtained by ChatGPT and that scored by the average human, and (e) the average percentile achieved by ChatGPT.

### **Synthesis and analysis of quantitative data**

Data were tested for normal distribution using a Kolmogorov-Smirnov test prior to analysis. Non-parametric analyses were used if data were not normally distributed. Data are reported as mean  $\pm$  standard error. Effect sizes are reported as Cohen's *d*. To compare the performance of ChatGPT to the pass mark for an examination, and to the average human mark, the distribution of difference scores was compared to a hypothetical median of zero using a one-sample Wilcoxon signed-rank test. A significant finding indicates that ChatGPT on average is better (where the sum of rank is positive) or worse (where the sum of ranks is negative) than the human pass mark/average performance. Analysis of a particular metric was made only on the basis of the examinations which reported that metric. The numbers of examinations included in a particular analysis is reported in the relevant results section. All findings were considered significant where  $p < 0.05$ .

### **Data charting and analysis; qualitative data**

These were analysed using the principles of top-down thematic analysis (Braun and Clarke 2006), with a pragmatic focus aimed at identifying useful recommendations for practitioners, as in similar work (Marano et al. 2023). Each paper was analysed for recommendations, based upon Research Questions 2 and 3. These recommendations could be explicit or implied. A list of recommendations was made for each paper by one author (MX). These lists were then analysed for common themes and summarized into generalisable recommendations. Both the initial lists and the summary recommendations were reviewed, discussed and agreed with the second author (PN) who had independently read the reviewed papers.

## **Results**

### **Summary of study characteristics**

114 examinations were identified from 53 studies, totalling 49014 MCQs. Seven different languages were represented although most were in English (85 examinations). The majority of the examinations were from medical subjects, including postgraduate specialty/board examinations (21), medical licensing examinations (33), medical school admissions examinations (14), medical pharmacology (1), pharmacy (4), cardiac life support (4), medical parasitology (1), dentistry (1) and anatomy (1). Other disciplines were law (8), computer science (3), economics (2), business (1), mathematics (6), 'thinking skills' (3), physics (2), chemistry (1), engineering (3), social work (3) and accounting (2). A summary of the reviewed studies and key raw data is in Table 1. Full details

extracted data are in [supplementary material S1](#). For full bibliographic information of included studies see [supplementary material S2](#).

### **Research question 1a. ChatGPT performance**

ChatGPT(3) was tested on 48596 items, answering 26068 (53.6%) correctly. ChatGPT(4) was tested on 19280 items, answering 14480 (75.1%) correctly. Studies which compared both versions used a total of 18862 items, of which ChatGPT(3) answered 49.5% correctly, while ChatGPT(4) answered 75.5% correctly. ChatGPT(4) outperformed ChatGPT(3) on every comparison study, by an average of  $25.0 \pm 1.2$  percentage points. A paired  $t$ -test on the examinations which tested both versions showed that this difference was significant ( $t_{(35)} = 20.89$ ,  $p < 0.0001$ ) with a substantial effect size ( $d = 3.1$ ) (Figure 1).

### **Research question 1b. Comparison of ChatGPT to pass mark**

ChatGPT(3) achieved a passing mark on 13/64 examinations (20.3%). ChatGPT(4) achieved a passing mark on 26/28 examinations (92.9%). For ChatGPT(3) the mean difference score was  $-11.0 \pm 1.6$  percentage points ( $W = -1273$ ,  $p < 0.0001$ ). For ChatGPT(4) this figure was  $7.9 \pm 1.5$  ( $W = 332$ ,  $p < 0.0001$ ). The distribution of difference scores is shown in Figure 2.

### **Research question 1c. Comparison of ChatGPT to human examinees**

ChatGPT(3) surpassed the mean score of human students on 5/46 examinations (10.9%), with a mean difference score of  $-20.7 \pm 2.8\%$  ( $W = -953$ ,  $p < 0.0001$ ). For ChatGPT(4) this figure was 7/20 (35%), with a mean difference of  $-4.8 \pm 3.2$  ( $W = -152$ ,  $p = 0.0808$ ), although most fails (compared to humans) were from the same study. The distribution of difference scores is shown in Figure 3. For ChatGPT(3) the average percentile was  $26.8 \pm 5.5$  ( $N = 27$ ), whereas for ChatGPT(4) the average percentile was  $81.5 \pm 10.3$  ( $N = 4$ ).

### **Evidence that test materials were present in the training materials?**

Only 28/114 (24.6%) analyses, from 15 studies, reported steps to ensure that their MCQs were not part of the ChatGPT training data. Most of these used ChatGPT(3). A Mann-Whitney U-test showed there was no difference in the mean performance of ChatGPT(3) in these studies ( $56.21 \pm 2.4$ ) compared to studies which did not take any steps ( $51.44 \pm 1.9$ ) ( $U = 1010$ ,  $p = 0.3046$ ).

### **Research question 2; How can we mitigate the challenges to assessment security posed by ChatGPT?**

Most studies expressed concerns about how ChatGPT can be used to cheat, particularly in unsupervised assessments. Specific recommendations were:

- a. *Stop using summative online assessments unless they are efficiently monitored.* Various suggestions were made about how to achieve this, including large-scale phasing out of online assessments and moving back to in-person formats, including oral presentations (Ali et al. 2023; Wood et al. 2023).
- b. *Further development and use of detection tools to identify AI generated text.* Many studies recognized that the routine use of these tools requires further research to optimize their validity and precision due to concerns about their accuracy (Ali et al. 2023; Kortemeyer 2023; Pursnani, Sermet and Demir 2023; Wood et al. 2023).

Table 1. Included studies.

References	Discipline	Level of learners	Language	Year Qs published	Contamination check?	N	Question	ChatGPT(3) % score	ChatGPT(3) Percentile	ChatGPT(4) % score	ChatGPT(4) Percentile	Pass Mark	Ave Human perf
(Alberts et al. 2023)	Med (Spec)	P	Eng	2009	N	50	34.0	c	-	-	-	-	-
(Ali et al. 2023)	Med (Spec)	P	Eng	ns	N	500	73.4	-	-	83.4	-	69	73.7
(Ali et al. 2023)	Dentistry	U	Eng	2023	Y	10	90.0	-	-	-	-	-	-
(Al-Shakarchi and Haq 2023)	Med (Lic)	P	Eng	ns	N	191	73.3	-	-	-	-	63.5	-
(Antaki et al. 2023)	Med (Spec)	P	Eng	ns	Y	260	42.7	-	-	-	-	-	61
(Azizoglu and Okur 2023)	Med (Spec)	P	Eng	ns	N	260	55.8	-	-	-	-	-	74
(Bommarito II and Katz 2022)	Law	U	Eng	2019	N	210	43.8	-	-	76.2	-	-	-
(Bommineni et al. 2023)	Med (Entry)	U	Eng	2022	Y	107	50.0	-	-	-	-	60.0	68
(Bordt and von Luxburg 2023)	Med (Entry)	U	Eng	2022	Y	46	50	-	-	-	-	-	-
(Carrasco et al. 2023)	Med (Lic)	U	Spa	2022	Y	53	75	-	-	-	-	-	-
(Choi et al. 2023)	Law	U	Eng	Ns	N	51	75	-	-	-	-	-	-
(Choi 2023)	Med Pharm	U	Eng (T)	2019-21	Y	56	75	-	-	-	-	-	-
(Cuthbert and Simpson 2023)	Med (Spec)	P	Eng	2022	N	8	75.0	-	-	100.0	-	-	-
(Farajollahi and Modaberi 2023)	Med (Spec)	P	Eng (T)	2022	Y	210	51.4	-	-	-	-	-	-
(Fijačko et al. 2023)	Cardiac LS	-	Eng	2016	N	60	40	2	-	-	-	-	80
(Freedman and Nappier 2023)	Med (Spec)	P	Eng	2022	Y	10	60	14	-	-	-	-	80
(Friederichs, Friederichs, and März 2023)	Med (Lic)	U	Ger	2021-22	N	25	84	6	-	-	-	-	92
(Geerling et al. 2023)	Economics	U	Eng	ns	N	312	76.0	-	-	-	-	65.8	44
(Georgakopoulos 2023)	Business	U	Eng	ns	N	134	35.8	-	-	-	-	70.0	-
					N	100	40.0	-	-	-	-	84	-
					N	25	64	-	-	-	-	84	-
					N	25	68	-	-	-	-	84	-
					N	38	68.4	-	-	-	-	84	-
					N	38	76.3	-	-	-	-	84	-
					Y	250	53.9	3.0	97.0	75.3	97.0	-	-
					N	250	51.2	8.0	88.0	77.9	88.0	-	-
					N	395	65.5	-	-	-	-	60.0	52.7
					N	30	63.3	99.0	-	-	-	-	43
					N	30	86.7	91.0	-	-	-	-	47
					N	10	30.0	-	-	-	-	-	-

(Continued)

Table 1. Continued.

References	Discipline	Level of learners	Language	Year Qs published	Contamination check?	N Question	ChatGPT(3) % score	ChatGPT(3) Percentile	ChatGPT(4) % score	ChatGPT(4) Percentile	Pass Mark	Ave Human perf
(Giannos and Delardas 2023)	Law	U	Eng	2010	N	42	35.7	—	—	—	—	42.1
				2010		42	52.4	—	—	—	—	42.1
	Med (Entry)			2019		16	50.0	62.0	—	—	—	—
				2019		17	17.6	7.0	—	—	—	—
				2020		26	65.4	73.0	—	—	—	—
				2020		20	45.0	62.0	—	—	—	—
				2021		25	56.0	51.0	—	—	—	—
				2021		22	4.5	1.0	—	—	—	—
				2019		18	22.2	18.0	—	—	—	—
				2020		19	10.5	3.0	—	—	—	—
			2021		20	15.0	5.0	—	—	—	—	
			2019		20	20.0	8.0	—	—	—	—	
			2020		18	11.1	6.0	—	—	—	—	
			2021		18	11.1	3.0	—	—	—	—	
			2019		37	59.5	42.0	—	—	—	—	
			2020		45	44.4	9.0	—	—	—	—	
			2021		43	41.9	15.0	—	—	—	—	
(Giannos 2023)	Med (Spec)	P	Eng	ns	N	69	57.0	—	64.0	—	58.0	—
(Glison et al. 2022)	Med (Lic)	U	Eng	ns	N	100	42	20.0	—	—	60	77.6
						100	44	30.0	—	—	60	77.6
						102	57.8	—	—	—	60	77.6
						87	64.4	—	—	—	60	77.6
(Giunti et al. 2023)	Med (Entry)	U	Eng	2021	N	25	64.0	44.9	—	—	—	—
			Eng	2021		22	31.8	29.6	—	—	—	—
			Ita	2022		60	61.7	—	—	—	51.7	—
(Holmes et al. 2023)	Med (Spec)	P	Eng	2023	Y	100	53.0	—	75.0	—	—	68
(Huang et al. 2023)	Med (Spec)	P	Eng	2021	N	293	63.1	—	74.0	—	60.0	—
(Huh 2023)	Parasitology	U	Kor/Eng	2022	N	79	60.8	—	—	—	60.8	90.8
(Huynh et al. 2023)	Med (Spec)	P	Eng	2022	Y	135	28.2	—	—	—	—	—
(Kaneda et al. 2023)	Med (Lic)	U	Jap	2023	Y	389	55.0	—	—	—	70	—

(Continued)





Table 1. Continued.

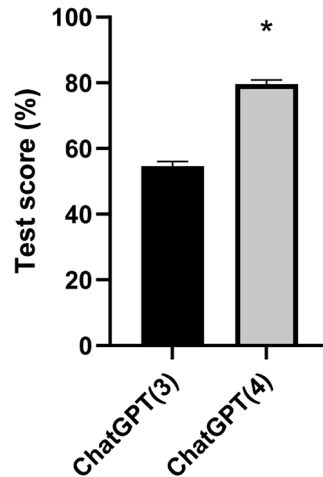
References	Discipline	Level of learners	Language	Year Qs published	Contamination check?	N Question	ChatGPT(3) % score	ChatGPT(3) Percentile	ChatGPT(4) % score	ChatGPT(4) Percentile	Pass Mark	Ave Human perf
(Kasai et al. 2023)	Med (Lic)	U	Jap	2019 2020 2018 2023 2021 2022 2018 2022 2023 2020 2021	N	296 299 299 295 300 297 200 197 200 200 200 200 197 200 200 202	50.7 49.5 47.8 47.5 51.3 54.9 61.5 62.9 83.2 85.0 60.0 59.9 59.9 71.5 49.2 46.7	— — — — — — — — — — — — — — — 10.0	72.6 73.2 73.9 74.9 75.0 76.8 80.5 83.2 85.0 85.0 85.3 86.5 75.7	— — — — — — — — — — — — — — 90.0	70.6 72.6 69.6 74.6 69.7 72.1 80.0 79.7 80.0 80.0 80.2 80.0 60.0	92.6 92.3 92.3 — 92.3 96.6 98.0 99.0 98.0 — 99.0 100.0 68
(Katz et al. 2023)	Law	U	Eng	2022	N	200	49.2	—	—	—	—	—
(Kortemeyer 2023)	Physics	U	Eng	ns	N	30	46.7	—	—	—	—	—
(Kumah-Crystal et al. 2023)	Med (Spec)	P	Eng	2018	N	254	74.8	—	—	—	60	77.6
(Kung et al. 2022)	Med (Lic)	U	Eng	2022	Y	119	36.1	—	—	—	60	75.8
(Leon and Vidhani 2023)	Chemistry	U	Eng	2022	N	122	55.7	—	—	—	60	82
(Liu et al. 2023)	Med (Entry)	P	Chi	2022	Y	165	37.0	—	—	—	70.0	—
(Naser et al. 2023)	Engineering	U	Eng	ns	N	39	51.5	—	46.2	—	50	—
(Nori et al. 2023)	Med (Lic)	U	Eng	Ns	N	79	—	—	70.9	—	65	—
(Oh, Choi, and Lee 2023)	Med (Spec)	P	Eng + Chi	2020-22	N	4183	50.1	—	69.5	—	—	—
(Passby, Jenko, and Werhham 2023)	Med (Spec)	P	Eng + Chi	ns	N	1273	44.6	—	74.7	—	—	—
(Pursnani, Seimet, and Demir 2023)	Engineering	U	Eng	2021	N	2173	49.1	—	83.8	—	60	77.6
(Savelka et al. 2023)	Coding	—	Eng	Ns	Y	376	56.9	—	84.3	—	60	77.6
(Shope 2023)	Law	U	Chi	2022	N	3426	40.3	—	71.1	—	—	—
(Skalidis et al. 2023)	Med (Spec)	P	Eng	2018/22	N	1413	50.6	—	82.2	—	—	—
(Takagi et al. 2023)	Med (Lic)	U	Jap	2023	N	280	46.8	—	76.4	—	—	—
					N	84	63.1	—	90.5	—	71.0	—
					N	122	42.5	—	66.4	—	—	—
					Y	106	67.9	—	89.6	—	70.0	—
					N	107	56.1	—	85.0	—	70.0	—
					N	300	—	—	57.0	51.0	57.3	62
					N	340	58.8	—	—	—	60.0	—
					N	176	48.8	—	76.7	—	74.6	83
					N	78	55.1	—	87.2	—	80.0	89.2

(Continued)

Table 1. Continued.

References	Discipline	Level of learners	Language	Year Qs published	Contamination check?	N Question	ChatGPT(3) % score	ChatGPT(3) Percentile	ChatGPT(4) % score	ChatGPT(4) Percentile	Pass Mark	Ave Human perf
(Talan and Kalinkara 2023)	Anatomy	U	Eng	2023	N	40	67.5	-	-	-	-	52.8
(Teabagy et al. 2023)	Med (Spec)	P	Eng	ns	N	167	57.0	-	81.0	-	-	74
(Thirunavukarasu 2023)	Med (Spec)	P	Eng	ns	N	49	42.9	-	-	-	60.2	-
						43	48.8	-	-	-	63.0	-
(Victor et al. 2023)	Social Work	P	Eng	2023	N	50	64.0	-	-	-	70.0	-
		U				50	76.0	-	-	-	70.0	-
		P				50	80.0	-	-	-	70.0	-
(Wang, Shen, and Chen 2023)	Pharmacy	P	Chi	2023	Y	210	53.8	-	-	-	60.0	-
			Chi			240	54.4	-	-	-	60.0	-
			Eng			240	56.9	-	-	-	60.0	-
			Eng			210	67.6	-	-	-	60.0	-
(Wang et al. 2023)	Med (Lic)	U	Chi	2022	N	600	36.5	-	-	-	60	68.7
				2021		600	45.8	-	-	-	60	67.9
(Weng et al. 2023)	Med (Spec)	P	Eng + Chi	2022	Y	125	41.6	-	-	-	60.0	75
(West 2023)	Physics	U	Eng	ns	Y	30	50.0	-	93.3	-	-	56.3
(Wood et al. 2023)	Accounting	U/P	Var	ns	N	20084	56.5	-	-	-	-	76.7
						1919	65.2	-	-	-	-	76.7

For 'discipline', 'Med (Entry)' = Admissions examination for Medical School, 'Med (Lic)' = medical licensing examination, 'Med (Spec)' = medical specialty examinations, 'Cardiac LS' = cardiac life support. Languages, Eng = English, Eng(T) = translated into English, Jap = Japanese, Chi = Chinese, Kor = Korean, Ita = Italian, Spa = Spanish, Ger = German, Var = Various. U = undergraduate, P = postgraduate. For full data extracted from each examination and calculations see [supplementary Material S1](#). For full bibliographic information of included studies see [supplementary material S2](#).



**Figure 1.** Raw test scores of ChatGPT(3) vs ChatGPT(4) from studies which directly tested both. \*=  $p < 0.005$  (paired  $t$ -test).

- c. *Use 'higher order' questions, with images.* Many studies proposed that ChatGPT struggles with higher order questions and yet does well with simple factual recall (Ali et al. 2023; Al-Shakarchi and Haq 2023; Choi 2023; Cuthbert and Simpson 2023; Friederichs, Friederichs and März 2023; Geerling et al. 2023; Huynh et al. 2023; Wang, Shen and Chen 2023; Wood et al. 2023).
- d. *Promote Academic Integrity.* With a focus on promoting the learning aspects, now and in future professional environments, rather than concentrating on cheating (Kortemeyer 2023)

### **Research question 3: how can we harness the positive potential of ChatGPT for higher education?**

There was widespread recognition that LLMs have the potential to aid the educational process, but that more research is necessary, in particular on their power and accuracy. Specific recommendations included:

- a. *Study Tool.* ChatGPT could be used by students as a study tool, complementary to traditional learning, and also as a revision aid when preparing for examinations (Liu et al. 2023). This included speculation that the technology could prepare personalised revision schedules and extra practice material (Pursnani, Sermet and Demir 2023). Students could ask ChatGPT to explain the answers it gives, and generate practice test questions (Bommineni et al. 2023; Cuthbert and Simpson 2023; Friederichs, Friederichs and März 2023; Fijačko et al. 2023; Kung et al. 2023)
- b. *Teaching Assistant.* ChatGPT could be used both in-real time when teaching and also during the preparation process (Naser et al. 2023; West 2023). Academics could also generate practice material and update their course content (West 2023), or even use ChatGPT as a facilitator in small group learning (Gilson et al. 2022). There was caution that ChatGPT could only complement, rather than replace, human instructors (Teebagy et al. 2023).
- c. *Test Subject.* ChatGPT(4) can aid in the testing of the validity and fairness of standardised examinations and provide useful feedback, by acting as a mock student (Passby, Jenko

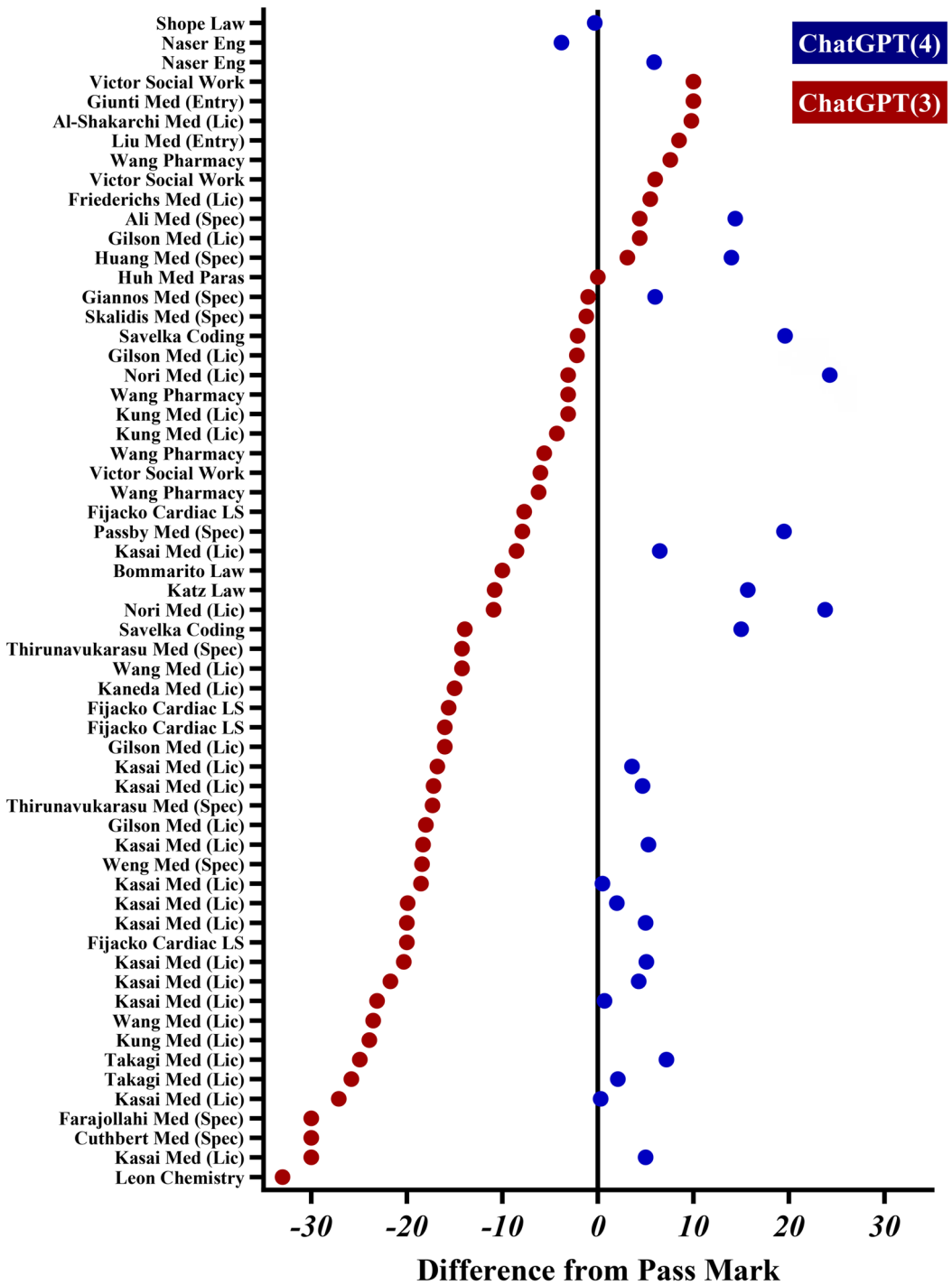


Figure 2. Difference between ChatGPT performance and passing score. Data are normalised to the pass mark, which varies between examinations. Thus, where ChatGPT scored below the pass mark, the data points are to the left of the central line at zero. Where ChatGPT exceeded the pass mark, data points are to the right. Not all studies tested both versions of ChatGPT.

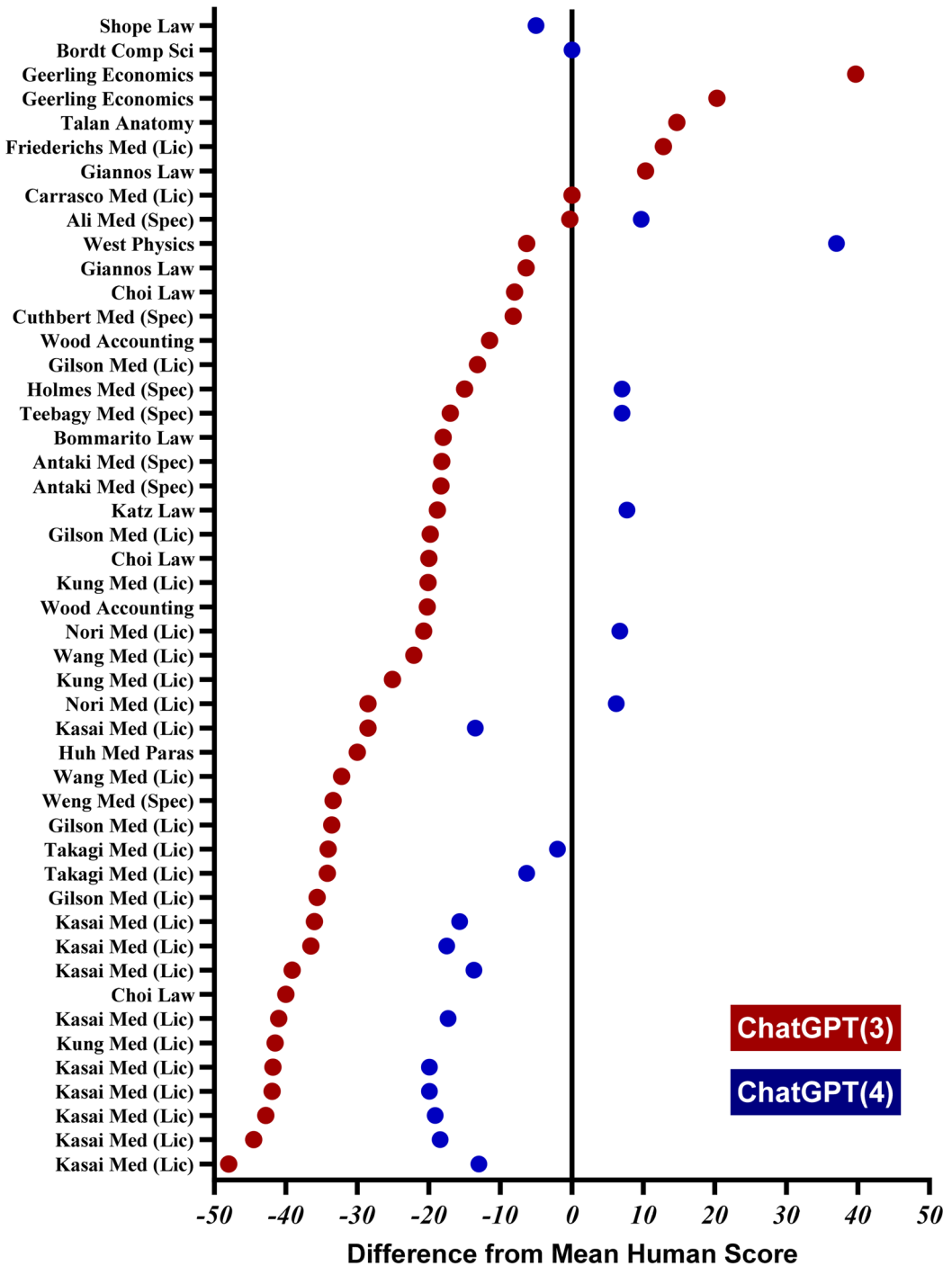


Figure 3. Difference between ChatGPT performance and the average human score. Data are normalised to the average score achieved by human test takers, which varies between examinations. Thus where ChatGPT scored below the mark achieved by humans, the data points are to the left of the central line at zero. Where ChatGPT scored higher, data points are to the right. Not all studies tested both versions of ChatGPT.

and Wernham 2023). This is supported by some studies showing that ChatGPT struggled to correctly answer questions that are also challenging for humans (Antaki et al. 2023; Friederichs, Friederichs and März 2023; Kasai et al. 2023).

- d. *Educating students.* These recommendations included developing and understanding of the risks and ethical implications of using ChatGPT for academic work, generally, as well as with a specific focus on academic integrity (Kortemeyer 2023; Pursnani, Sermet and Demir 2023; Victor et al. 2023).
- e. *Share best/emerging practice* about the uses of LLMs in education (Ali et al. 2023; Giannos 2023).
- f. *Further mapping and development of ChatGPT.* Many studies expressed cautious optimism, particularly those from the clinical field (Ali et al. 2023; Antaki et al. 2023; Cuthbert and Simpson 2023), with specific clarity needed on liability when using ChatGPT (Antaki et al. 2023; Carrasco et al. 2023) and the need to develop standards for ethical use (Liu et al. 2023; Passby, Jenko and Wernham 2023). One study suggested that future iterations and applications of the technology could include an 'indicator of uncertainty' to aid with clinical decision-making when using ChatGPT (Thirunavukarasu 2023), since ChatGPT currently does not indicate how confident it is in an answer, and does not appear to have 'insight' into its own limitations. These features are a cause for concern when considering the educational use of ChatGPT, especially in clinical settings (Cuthbert and Simpson 2023; Friederichs, Friederichs and März 2023).

## Discussion

### Summary of evidence

Older/free versions of ChatGPT based on GPT3 or GPT3.5 failed most examinations and underperformed compared to the average student, but still answered approximately half the questions correctly. ChatGPT(4) performed significantly better, passing most examinations. These findings have serious implications for the use of current MCQ-based assessments in higher education. In particular they provide clear evidence that unproctored online examinations are no longer a meaningful summative assessment method, in contrast to recent findings from research conducted before the emergence of ChatGPT (Chan and Ahn 2023; Newton 2023b).

### Limitations

This is a review of studies whose research questions were not plausible just a few months ago, and for which there is not currently established best-practice. One potential future example of good practice is to distinguish between the performance of LLM tools on questions which are *de novo* versus those which were present in the training data, or which are linguistically very similar. The distinction is important from a basic perspective because it, at least simplistically, represents the difference between 'reasoning' and simply reproducing an answer *via* memorization. We found no evidence of reduced performance on *de novo* questions, supporting studies which conclude that ChatGPT is not simply regurgitating findings from its training materials (Freedman and Nappier 2023; Nori et al. 2023; OpenAI 2023b).

There is an important pragmatic consideration here as well. If ChatGPT showed reduced performance on *de novo* questions, then theoretically one way to increase the security of examinations might be to generate a completely new set of questions for each sitting of the examination. However, from a pragmatic perspective this seems problematic. ChatGPT is not simply

regurgitating verbatim content from its training materials. Instead, it has been designed to demonstrate 'contextual understanding' meaning that it recognises similar patterns of language. Thus, even a completely *de novo* question would be easily answered if a question addressing the same learning outcome had been part of the training materials. Most higher education courses include core concepts and some use a specified curriculum. There are only so many ways in which questions on these topics can be asked, and by prioritising the use of novel language in the construction of every question it seems likely that many important aspects of item creation will be lost.

Many early studies proposed designing MCQ items that are harder for ChatGPT to answer. A feature stated by numerous papers was that ChatGPT performed better on recall/memorisation questions which assess factual knowledge, when compared to higher order questions which assess problem solving or transfer of knowledge (Ali et al. 2023; Al-Shakarchi and Haq 2023; Choi 2023; Cuthbert and Simpson 2023; Friederichs, Friederichs and März 2023; Geerling et al. 2023; Huynh et al. 2023; Wang, Shen and Chen 2023; Wood et al. 2023). This pattern was reflected in some studies where ChatGPT substantially outperformed the average human; the sample MCQs were of a 'lower order' type (Geerling et al. 2023; Giunti et al. 2023; Talan and Kalinkara 2023). However the much-improved performance of ChatGPT(4) seems to have quashed the idea that ChatGPT can be thwarted by the use of 'higher order' questions, particularly given its performance on examinations used for medical licensing and postgraduate medical qualifications. These MCQs are specifically designed to assess higher order learning and problem solving, and are written to a very high standard (Billings et al. 2020). Thus, it seems futile to try and 'outwit' ChatGPT by designing examination questions which it cannot answer, especially given that the models will likely improve.

Instead, we need to find other ways to mitigate the challenges to academic integrity posed by ChatGPT. One approach is to eliminate closed-book examinations and instead allow students access to ChatGPT, thus making assessment more of an authentic reflection of the real-world, since graduates are likely to be using these tools in their future jobs. However, foundational, basic knowledge is essential (Willingham 2006); it is the basis by which a graduate would know what to ask ChatGPT. Thus, assessments of foundational knowledge need to be undertaken securely, without access to ChatGPT or related tools. This does not necessarily mean in-person traditional examinations; it can include practical assessments, skills and presentations, vivas etc, as long as they are 'in-person', real time and proctored/invigilated.

Two studies were unconcerned about the impact of ChatGPT on academic integrity (Alberts et al. 2023; Thirunavukarasu 2023). ChatGPT showed modest performance in these studies, but both these studies utilized ChatGPT(3) and it is highly likely that ChatGPT(4) would show an improved performance. Even the seemingly modest performance of ChatGPT(3) is a significant cause for concern. For example, consider a struggling student who can correctly answer (without ChatGPT) only 40 questions on a 200-item examination. The data suggests that, using ChatGPT, they could correctly answer ~40% of the remainder (=64), substantially improving their score from 20% to 52%. It seems reasonable to conclude, based on the studies reviewed here, that a struggling student will almost certainly improve their score substantially if using ChatGPT in an unauthorized way.

Thus, we need to improve the security of assessments based on MCQs. An obvious common use of MCQs, to which ChatGPT is a threat, is in online examinations, where misconduct was already high before ChatGPT (Newton and Essex 2023). ChatGPT offers a more powerful, accurate tool to students when compared to simply searching the internet (Schultz et al. 2022). A more nuanced terminology and practice is needed, for example where ChatGPT is considered as a standalone tool (Dawson, Nicola-Richmond, and Partridge 2023). One oft-cited approach to increasing the security of online examinations is to use some form of remote monitoring system, but there are currently substantial challenges with the student experience of these systems, with students reporting concerns about fairness, technology, access and anxiety (Marano et al. 2023).

Some studies suggested that ChatGPT would perform worse in examinations that were not developed in English (Kaneda et al. 2023; Liu et al. 2023), and one concluded that *If the exam is given in English, ChatGPT scores 10.4% higher than if it is given in a different language* (Wood et al. 2023), but this study included a range of different question formats and the language effect was not broken down by format. English-language results dominated the examinations we reviewed. The next largest representation was Japanese with 15 examinations, but these were all from the same discipline (Medical Licensing). More work is needed to understand the role of language in the performance of ChatGPT on MCQs.

There is also a need to evaluate the performance of other LLM-based chatbots, including those developed in other languages. At the time of writing (November 2023) a number of other systems are available or in development, and these can now be 'customised' to produce chatbots that are designed for specific tasks, for example a 'ChatDoctor' based on a fine-tuned version of the LLaMA LLM from Meta, the owners of Facebook (Li et al. 2023). It seems reasonable, based on the evidence and trajectory to date, to propose that all models will be improved.

## Conclusion

ChatGPT is a serious threat to the use of MCQ-based assessments, which should now either (a) incorporate student use of LLMs into their design or (b) be in-person and invigilated. LLM-based Chatbots offer considerable opportunity for the future of higher education, but these challenges to academic integrity are here now, and pose a very substantial threat which must be addressed first.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Notes on contributors

**Philip Newton** is a neuroscientist at the Swansea University Medical School. His research focuses on evidence based approaches to learning, teaching and assessment in higher education.

**Maira Xiromeriti** is a medical student at the Swansea University Medical School.

## ORCID

Philip Newton  <http://orcid.org/0000-0002-5272-7979>

Maira Xiromeriti  <http://orcid.org/0000-0002-2975-184X>

## References

- Alberts, I. L., L. Mercolli, T. Pyka, G. Prenosil, K. Shi, A. Rominger, and A. Afshar-Oromieh. 2023. "Large Language Models (LLM) and ChatGPT: What Will the Impact on Nuclear Medicine Be?" *European Journal of Nuclear Medicine and Molecular Imaging* 50 (6): 1549–1552. doi:10.1007/s00259-023-06172-w.
- Ali, K., N. Barhom, F. T. Marino, and M. Duggal. 2023. "The Thrills and Chills of ChatGPT: Implications for Assessments in Undergraduate Dental Education." Preprints. <https://www.preprints.org/manuscript/202302.0513/v1>.
- Ali, R., O. Y. Tang, I. D. Connolly, P. L. Z. Sullivan, J. H. Shin, J. S. Fridley, W. F. Asaad, et al. 2023. "Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations." medRxiv. doi:10.1227/neu.0000000000002632.
- Al-Shakarchi, N. J., and I. U. Haq. 2023. "ChatGPT Performance in the UK Medical Licensing Assessment: How to Train the Next Generation?" *Mayo Clinic Proceedings: Digital Health* 1 (3): 309–310. doi:10.1016/j.mcpdig.2023.06.004.
- Antaki, F., S. Touma, D. Milad, J. El-Khoury, and R. Duval. 2023. "Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings." medRxiv <https://www.medrxiv.org/content/10.1101/2023.01.22.23284882v2>.



- Billings, M., K. DeRuchie, K. Hussie, A. Kulesher, J. Merrell, A. Morales, M. Paniagua, J. Sherlock, K. Swygert, and J. Tyson. 2020. *Constructing Written Test Questions for the Health Sciences*. Philadelphia, PA: National Board of Medical Examiners. [https://www.nbme.org/sites/default/files/2020-11/NBME\\_Item%20Writing%20Guide\\_2020.pdf](https://www.nbme.org/sites/default/files/2020-11/NBME_Item%20Writing%20Guide_2020.pdf).
- Björk, B.-C., and D. Solomon. 2013. "The Publishing Delay in Scholarly Peer-Reviewed Journals." *Journal of Informetrics* 7 (4): 914–923. doi:10.1016/j.joi.2013.09.001.
- Bommineni, V. L., S. Z. Bhagwagar, D. Balcarcel, C. Davatzikos, and D. L. Boyer. 2023. "Performance of ChatGPT on the MCAT: The Road to Personalized and Equitable Premedical Learning." medRxiv. <https://www.medrxiv.org/content/10.1101/2023.03.05.23286533v2>.
- Braun, V., and V. Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3 (2): 77–101. doi:10.1191/1478088706qp063oa.
- Cambridge Admissions. 2021. "Explanation of BMAT Results." <https://www.admissionstesting.org/Images/535824-bmat-test-specification.pdf>.
- Carrasco, J. P., E. García, D. A. Sánchez, E. Porter, L. D. L. Puente, J. Navarro, and A. Cerame. 2023. "Is 'ChatGPT' Capable of Passing the 2022 MIR Exam? Implications of Artificial Intelligence in Medical Education in Spain." *Revista Española de Educación Médica* 4 (1): 55–69. <https://revistas.um.es/edumed/article/view/556511>
- Ceres, P. 2023. "ChatGPT Is Coming for Classrooms. Don't Panic." *Wired*. <https://www.wired.com/story/chatgpt-is-coming-for-classrooms-dont-panic/>.
- Chan, J. C. K., and D. Ahn. 2023. "Unproctored Online Exams Provide Meaningful Assessment of Student Learning." *Proceedings of the National Academy of Sciences of the United States of America* 120 (31): e2302020120. doi:10.1073/pnas.2302020120.
- Choi, W. 2023. "Assessment of the Capacity of ChatGPT as a Self-Learning Tool in Medical Pharmacology: A Study Using MCQs." <https://www.researchsquare.com>.
- Cuthbert, R., and A. I. Simpson. 2023. "Artificial Intelligence in Orthopaedics: Can Chat Generative Pre-Trained Transformer (ChatGPT) Pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) Examination?" *Postgraduate Medical Journal* 99 (1176): 1110–1114. doi:10.1093/postmj/qgad053.
- Dawson, P., K. Nicola-Richmond, and H. Partridge. 2023. "Beyond Open Book versus Closed Book: A Taxonomy of Restrictions in Online Examinations." *Assessment & Evaluation in Higher Education*. doi:10.1080/02602938.2023.2209298.
- Farazouli, A., T. Cerratto-Pargman, K. Bolander-Laksov, and C. McGrath. 2023. "Hello GPT! Goodbye Home Examination? An Exploratory Study of AI Chatbots Impact on University Teachers' Assessment Practices." *Assessment & Evaluation in Higher Education*. doi:10.1080/02602938.2023.2241676.
- Fijačko, N., L. Gosak, G. Štiglic, C. T. Picard, and M. J. Douma. 2023. "Can ChatGPT Pass the Life Support Exams without Entering the American Heart Association Course?" *Resuscitation* 185. [https://www.resuscitationjournal.com/article/S0300-9572\(23\)00045-X/fulltext](https://www.resuscitationjournal.com/article/S0300-9572(23)00045-X/fulltext).
- Freedman, J. D., and I. A. Nappier. 2023. "GPT-4 to GPT-3.5: 'Hold My Scalpel' – A Look at the Competency of OpenAI's GPT on the Plastic Surgery In-Service Training Exam." arXiv. <http://arxiv.org/abs/2304.01503>.
- Friederichs, H., W. J. Friederichs, and M. März. 2023. "ChatGPT in Medical School: How Successful is AI in Progress Testing?" *Medical Education Online* 28 (1): 2220920. doi:10.1080/10872981.2023.2220920.
- Geerling, W., G. D. Mateer, J. Wooten, and N. Damodaran. 2023. *Is ChatGPT Smarter than a Student in Principles of Economics?* SSRN Scholarly Paper. Rochester, NY. <https://papers.ssrn.com/abstract=4356034>.
- Giannos, P. 2023. "Evaluating the Limits of AI in Medical Specialisation: ChatGPT's Performance on the UK Neurology Specialty Certificate Examination." *BMJ Neurology Open* 5 (1): e000451. doi:10.1136/bmjno-2023-000451.
- Gilson, A., C. Safraneck, T. Huang, V. Socrates, L. Chi, R. A. Taylor, and D. Chartash. 2022. "How Does ChatGPT Perform on the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment." medRxiv. <https://www.medrxiv.org/content/10.1101/2022.12.23.22283901v1>.
- Giunti, M., F. G. Garavaglia, R. Giuntini, S. Pinna, and G. Sergioli. 2023. *Chatgpt Prospective Student at Medical School*. SSRN Scholarly Paper. Rochester, NY. <https://papers.ssrn.com/abstract=4378743>.
- Haddaway, N. R., A. M. Collins, D. Coughlin, and S. Kirk. 2015. "The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching." *PLoS One* 10 (9): e0138237. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4574933/>. doi:10.1371/journal.pone.0138237.
- Hammer, A. 2023. "ChatGPT Can Pass the US Medical Licensing Exam and the Bar Exam." *Mail Online*. <https://www.dailymail.co.uk/news/article-11666429/ChatGPT-pass-United-States-Medical-Licensing-Exam-Bar-Exam.html>.
- Huynh, L. M., B. T. Bonebrake, K. Schultis, A. Quach, and C. M. Deibert. 2023. "New Artificial Intelligence ChatGPT Performs Poorly on the 2022 Self-Assessment Study Program for Urology." *Urology Practice* 10 (4): 409–415. doi:10.1097/UPJ.0000000000000406.
- Kaneda, Y., T. Tanimoto, A. Ozaki, T. Sato, and K. Takahashi. 2023. "Can ChatGPT Pass the 2023 Japanese National Medical Licensing Examination?" Preprints. <https://www.preprints.org/manuscript/202303.0191/v1>.
- Kasai, J., Y. Kasai, K. Sakaguchi, Y. Yamada, and D. Radev. 2023. "Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations." arXiv. <http://arxiv.org/abs/2303.18027>.
- Kaushik, V., and C. A. Walsh. 2019. "Pragmatism as a Research Paradigm and Its Implications for Social Work Research." *Social Sciences* 8 (9): 255. doi:10.3390/socsci8090255.

- Kortemeyer, G. 2023. "Could an Artificial-Intelligence Agent Pass an Introductory Physics Course?" arXiv. <http://arxiv.org/abs/2301.12127>.
- Kung, T. H., M. Cheatham, A. Medenilla, C. Sillos, L. D. Leon, C. Elepaño, M. Madriaga, et al. 2023. "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models." *PLOS Digital Health* 2 (2): e0000198. doi:10.1371/journal.pdig.0000198.
- Li, Y., Z. Li, K. Zhang, R. Dan, S. Jiang, Y. Zhang, Y. Li, et al. 2023. "ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge." *Cureus* 15 (6): e40895. <https://www.cureus.com/articles/152858-chatdoctor-a-medical-chat-model-fine-tuned-on-a-large-language-model-meta-ai-llama-using-medical-domain-knowledge>. doi:10.7759/cureus.40895.
- Liu, X., C. Fang, Z. Yan, X. Liu, Y. Jiang, Z. Cao, M. Wu, et al. 2023. "Performance of ChatGPT on Clinical Medicine Entrance Examination for Chinese Postgraduate in Chinese." medRxiv. <https://www.medrxiv.org/content/10.1101/2023.04.12.23288452v1>.
- Marano, E., P. M. Newton, Z. Birch, M. Croombs, C. Gilbert, and M. J. Draper. 2023. "What is the Student Experience of Remote Proctoring? A Pragmatic Scoping Review." EdArXiv. <https://edarxiv.org/jrgw9/>.
- Medify. 2023. "What is a Good BMAT Score? | Blog | Medify UK." <https://www.medify.co.uk/blog/good-bmat-score>.
- Mollick, E. 2023. "The Homework Apocalypse." <https://www.oneusefulthing.org/p/the-homework-apocalypse>.
- Narayanan, A., and S. Kapoor. 2023. "GPT-4 and Professional Benchmarks: The Wrong Answer to the Wrong Question. Substack Newsletter." *AI Snake Oil*. <https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks>
- Naser, M. Z., B. Ross, J. Ogle, V. Kodur, R. Hawileh, J. Abdalla, and H.-T. Thai. 2023. "Can AI Chatbots Pass the Fundamentals of Engineering (FE) and Principles and Practice of Engineering (PE) Structural Exams?" arXiv <http://arxiv.org/abs/2303.18149>.
- Newton, P. M. 2023a. "Guidelines for Creating Online MCQ-Based Exams to Evaluate Higher Order Learning and Reduce Academic Misconduct." In *Handbook of Academic Integrity*, edited by S. E. Eaton, 1–17. Singapore: Springer Nature. doi:10.1007/978-981-287-079-7\_93-1.
- Newton, P. M. 2023b. "The Validity of Unproctored Online Exams is Undermined by Cheating." *Proceedings of the National Academy of Sciences* 120 (41): e2312978120.
- Newton, P. M., A. Da Silva, and S. Berry. 2020. "The Case for Pragmatic Evidence-Based Higher Education: A Useful Way Forward?" *Frontiers in Education* 5: 5. <https://www.frontiersin.org/articles/10.3389/feduc.2020.583157/full>. doi:10.3389/feduc.2020.583157.
- Newton, P. M., and K. Essex. 2023. "How Common is Cheating in Online Exams and Did It Increase during the COVID-19 Pandemic? A Systematic Review." *Journal of Academic Ethics*. doi:10.1007/s10805-023-09485-5.
- Nori, H., N. King, S. M. McKinney, D. Carignan, and E. Horvitz. 2023. "Capabilities of GPT-4 on Medical Challenge Problems." <http://arxiv.org/abs/2303.13375>.
- OpenAI. 2023a. *New GPT-3 Capabilities: Edit & Insert*. <https://openai.com/blog/gpt-3-edit-insert>.
- OpenAI. 2023b. *GPT-4 Technical Report*. <http://arxiv.org/abs/2303.08774>.
- Passby, L., N. Jenko, and A. Wernham. 2023. "Performance of ChatGPT on Dermatology Specialty Certificate Examination Multiple Choice Questions." *Clinical and Experimental Dermatology* 2: llad197. doi:10.1093/ced/llad197.
- Pursnani, V., Y. Sermet, and I. Demir. 2023. "Performance of ChatGPT on the US Fundamentals of Engineering Exam: Comprehensive Assessment of Proficiency and Potential Implications for Professional Environmental Engineering Practice." arXiv. <http://arxiv.org/abs/2304.12198>.
- Schultz, M., K. F. Lim, Y. K. Goh, and D. L. Callahan. 2022. "OK Google: What's the Answer? Characteristics of Students Who Searched the Internet during an Online Chemistry Examination." *Assessment & Evaluation in Higher Education* 47 (8): 1458–1474. doi:10.1080/02602938.2022.2048356.
- Talan, T., and Y. Kalinkara. 2023. The Role of Artificial Intelligence in Higher Education: ChatGPT Assessment for Anatomy Course. *Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi* 7, no. 1: 33–40. doi:10.33461/uybisbbd.1244777.
- Teebagy, S., L. Colwell, E. Wood, A. Yaghy, and M. Faustina. 2023. "Improved Performance of ChatGPT-4 on the OKAP Exam: A Comparative Study with ChatGPT-3.5." medRxiv <https://www.medrxiv.org/content/10.1101/2023.04.03.23287957v1>.
- Thirunavukarasu, A. J. 2023. "ChatGPT Cannot Pass FRCOphth Examinations: Implications for Ophthalmology and Large Language Model Artificial Intelligence." *Eye News*. <https://www.eyenews.uk.com/features/ophthalmology/post/chatgpt-cannot-pass-frcophth-examinations-implications-for-ophthalmology-and-large-language-model-artificial-intelligence>.
- Tricco, Andrea C., Erin Lillie, Wasifa Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, et al. 2018. "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation." *Annals of Internal Medicine* 169 (7): 467–473. <https://www.acpjournals.org/doi/10.7326/M18-0850>. doi:10.7326/M18-0850.
- Victor, B. G., S. Kubiak, B. Angell, and B. E. Perron. 2023. "Time to Move beyond the ASWB Licensing Exams: Can Generative Artificial Intelligence Offer a Way Forward for Social Work?" *Research on Social Work Practice* 33 (5): 511–517. doi:10.1177/10497315231166125.

- Wang, Y.-M., H.-W. Shen, and T.-J. Chen. 2023. "Performance of ChatGPT on the Pharmacist Licensing Examination in Taiwan." *Journal of the Chinese Medical Association: JCMA* 86 (7): 653–658. doi:10.1097/JCMA.0000000000000942.
- Wang, Z., W. Glänzel, and Y. Chen. 2020. "The Impact of Preprints in Library and Information Science: An Analysis of Citations, Usage and Social Attention Indicators." *Scientometrics* 125 (2): 1403–1423. doi:10.1007/s11192-020-03612-4.
- West, C. G. 2023. "AI and the FCI: Can ChatGPT Project an Understanding of Introductory Physics?" <http://arxiv.org/abs/2303.01067>.
- Willingham, D. 2006. *How Knowledge Helps*. Washington, DC: American Federation of Teachers. <https://www.aft.org/periodical/american-educator/spring-2006/how-knowledge-helps>.
- Wood, D., A. M. P. Achhpilia, M. T. Adams, S. Aghazadeh, K. Akinyele, M. Akpan, K. D. Allee, et al. 2023. "The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions?" *Issues in Accounting Education* 38 (4): 1–28.