

ChatGPT and Medical Education: A Double-Edged Sword

¹Urfa Khairatun Hisan, ²Muhammad Miftahul Amri*

Corresponding Author: *muhammad.amri@te.uad.ac.id

¹ Faculty of Medicine, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

² Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

ARTICLE INFO

Article history

Received 6 March 2023

Revised 9 March 2023

Accepted 11 March 2023

Keywords

ChatGPT

medical education

exam

research

abstract

ABSTRACT

ChatGPT has gained attention worldwide. In the medical education field, ChatGPT, or any similar large language model, provides a convenient way for students to access information and practice their skills. ChatGPT can simulate patient interactions, allowing students to practice their diagnostic and communication skills in a safe and controlled environment. ChatGPT can also answer questions and provide explanations for complex medical concepts. There have been attempts involving ChatGPT to assist researchers in writing journal articles. Due to its capabilities, ChatGPT has the potential to be abused and sparking ethical concerns. Unwise researchers can now employ ChatGPT to write academic articles. Similarly, irresponsible students, might utilize ChatGPT to cheat during exams. We investigated whether ChatGPT, in its current state, can answer Indonesian medical doctor examination problems. Among the 3 problems that we have picked, ChatGPT can only correctly answer 1 question. We also examine whether ChatGPT-generated abstracts can fool professionals and educators. We have brought 6 abstracts, 3 of which are taken from actual published papers, while the other 3 were generated by ChatGPT. We recruited 12 participants with either a medical doctor (M.D.), a doctorate (Ph.D.), or an M.D., Ph.D. degree from various institutions in Indonesia. Surprisingly, 4 of the participants couldn't guess a single abstract correctly, 6 could only identify 1 abstract accurately, one correctly guessed 2 abstracts, and one correctly identified 3 abstracts. Therefore, it is safe to say that ChatGPT, in its current state, has been able to fool professionals and educators.

This is an open-access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Introduction

Since ChatGPT was first made available to the public in November 2022, there has been an enormous number of people who have tried it and are amazed by its capability [1-3]. It is safe to say that ChatGPT has created quite an uproar. According to the ChatGPT itself, ChatGPT is a large language model developed by OpenAI. It is a variant of the GPT (Generative Pre-trained Transformer) model and is trained on a massive dataset of conversational text to generate human-like responses in natural language understanding and generation tasks. It can be fine-tuned for various natural languages processing tasks such as question answering, language translation, and text summarization.

Following these capabilities, various attempts have been made to exploit ChatGPT's abilities. ChatGPT can be used to do diverse things, such as mathematical calculations, model simulations, challenge arguments, and even write music and essays. ChatGPT can act as a brainstorming partner, product manager, Quality Assurance (QA) tester, marketing head, and many other professionals. In [4], the authors employed ChatGPT to generate a literature review of Digital Twin in Healthcare. In [5], the authors published an editorial paper titled "*A conversation with ChatGPT on the role of computational systems biology in stem cell research*," where ChatGPT has been used as some kind of 'conversation partner'. ChatGPT might also replace conventional methods to assess students' knowledge [6, 7]. In [8], the authors showed that ChatGPT, which has been utilized for stance detection tasks, can achieve SOTA or similar performance for commonly used datasets, including SemEval-2016 and P-Stance. The authors in [9] employed ChatGPT in their article titled "*ChatGPT Talks About Animal Rights*".

The utilization of ChatGPT in the finance sector has also been explored. In [10], the findings suggest that ChatGPT has the potential to improve NLP-based financial applications. However, as in [11], the commercial use of ChatGPT, indeed, requires caution. Security risk and ethical issues are among two factors that need to be carefully considered.

Apparently, ChatGPT can also be used to aid programmers in designing their code. A command as simple as "*write me a [programming language] code for [task]*" will trigger ChatGPT to generate the appropriate code. This situation has sparked discussion in the programming field about whether ChatGPT will replace or will not replace programmers. There have been mixed opinions on this. Many people said that the current (or the future) ChatGPT might be able to replace programmers. On the other side, as in [12], they believe that ChatGPT will not be able to replace coders as programming requires complicated thinking, which only humans can do.

While ChatGPT seems novel and offers various capabilities that have never been available before, it raises a lot of discussion regarding the ethical aspects and its potential to be abused, especially in the field of education and academics. Among those concerns were

raised by the author in [13, 14]. The authors were concerned that ChatGPT might be used to cheat on exams and evaluations. In [15], the author investigated whether ChatGPT is able to pass high school exams on English Language Comprehension in the Netherlands. The result shows that ChatGPT can achieve a mean grade as high as 7.18, which is similar to the mean grade of all students in the Netherlands. The author concludes that while ChatGPT provides notable opportunities, it also raises concerns about the impact of such technologies.

Nevertheless, it is currently unclear how far ChatGPT could perform in academic tasks and exams and, thus, whether it constitutes an opportunity or a threat. This is because, while the above authors have demonstrated the relatively-decent performance of ChatGPT, other authors have argued that in its current state, ChatGPT-generated text is easily identifiable by attentive readers. The current ChatGPT-generated paper tends to contain incorrect statements, grammatically error texts, repeated and monotone sentences, a cliché-like writing style, and wrong (even nonexistent) references [16-19]. In this article, we aim to investigate whether the current ChatGPT (15 January 2023 version) is able to generate a convincing academic abstract to the extent that it can fool academia and experts in its field. The ChatGPT's ability to answer standard medical examination questions is also investigated in this study.

Finally, the rest of this article is organized as follows. In Section II, we briefly explain how ChatGPT works. The explanation is made simple such that it is understandable for those readers who do not have computer science backgrounds. Section III presents the results of our first experiment: *ChatGPT to Generate Academic Journals: Can It Fool Experts?*, followed by the arguments that ChatGPT introduces potential abusive usage in academic writing in Section IV. In Section V, we present the findings of our second experiment: *ChatGPT to Answer Exam and Homework Questions: Is it Possible?*, while its related discussion on whether Exams and Homework are still needed in the Medical Education Standard is presented in Section VI. Finally, we conclude this study in Section VII. Note that in this article, the term ChatGPT can also be interpreted as other language models. We use the term ChatGPT as it is the language model that is currently attracting the most interest worldwide.

How does ChatGPT work?

ChatGPT works by using a deep learning technique called unsupervised pre-training. This means that the model is first trained on a large dataset of conversational text without being given any specific task to perform. During this pre-training stage, the model learns to understand and generate natural language by analyzing patterns and relationships in the data.

Once pre-trained, the model can be fine-tuned for specific natural languages processing tasks such as question answering, language translation, and text summarization. During fine-tuning, the model is given a smaller, task-specific dataset and "learns" to perform the task by

adjusting the parameters in its neural network. In more detail, ChatGPT is a transformer-based language model, which means that it uses attention mechanisms (self-attention) to understand the context of the input text. It can be fine-tuned by training it on a smaller dataset with specific task-related labels while the pre-trained weights of the model remain fixed. This enables the model to perform the specific task with better accuracy and speed.

ChatGPT to Generate Academic Journals: Can It Fool Experts?

While ChatGPT's capability has amazed many people, its full potential in the research and academic sector is not yet explored and, thus, intriguing. However, as mentioned in the introduction section of this manuscript, several researchers have investigated its potential to aid them in writing research articles. These capabilities have, indeed, raised many people's eyebrows. In fact, many authors have listed ChatGPT as one of the co-authors [20-23]. ChatGPT was listed as one among 12 authors in an article titled "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models" [20]. In [21], ChatGPT was listed as an author alongside O'Connor in an editorial in the journal *Nurse Education in Practice*. Although, not everyone agrees with this kind of practice [24]. While Nature and Science have stated that ChatGPT does not meet the authorship eligibility, Taylor & Francis and ArXiv are still reviewing this matter.

In this section, we investigated whether, at its current state, ChatGPT is able to fool experts. To do that, we recruited professionals with academic backgrounds of a medical doctor (M.D.), a doctorate degree (Ph.D.), or a medical doctor with a doctorate degree (M.D., Ph.D.) from various institutions in Indonesia. Among them were medical doctors working in clinical setting (hospitals) or university professors (lecturers) in medicine or health-related majors. We then presented 4 academic abstracts, 2 of which were dummy abstracts generated by the ChatGPT. The participants were then asked to identify which abstracts were real and which ones were dummies. We did not disclose how many of those abstracts were real.

The participants' responses are gathered through questionnaires via Google Forms. The Google Forms links are distributed to several academic institutions and hospitals in Indonesia. We set the data-collecting period as only three days. This study is conducted in a voluntary manner, meaning that no respondents are 'forced' to fill out the questionnaire. As a result, 12 respondents returned the questionnaire. Other than the latest academic degrees and occupation (medical doctor or university lecturer), no other information was collected during the study nor presented in this manuscript. This study has not been reviewed by institutional review boards (IRB) or ethical committees since the identity of the participants cannot be readily ascertained, directly or indirectly, through identifiers linked to the subjects. Thus, this study falls within one of the eight criteria for Human Subjects Research Exemptions [25-27].

Still, before filling out the questionnaire, all the participants are asked to fill out the written consent statement through Google Forms.

The 4 abstracts used in our study are presented below. As a reader, you may also test your own ability to identify by guessing whether the abstracts below are legit or generated by ChatGPT.

A. Abstract 1

"The present study aimed to investigate the effect of vibration therapy on the acceleration of wound healing in diabetic neuropathic foot ulcers. A total of 60 patients with diabetic neuropathic foot ulcers were recruited for this prospective experimental study. Participants were randomly assigned to either a vibration therapy group (n=30) or a control group (n=30). The vibration therapy group received daily vibration therapy sessions for four weeks, while the control group did not receive any adjuvant intervention. Wound healing was measured using standardized wound assessment tools at the beginning of the study and at four-week intervals for a total of 12 weeks. The wound area, wound depth, and wound volume were measured, as well as the presence of exudate, granulation tissue, and epithelialization. In addition, patients were assessed for wound pain and quality of life using a visual analog scale (VAS) and the Short Form-36 (SF-36) questionnaire, respectively. The results of this study showed that vibration therapy significantly accelerated wound healing in diabetic neuropathic foot ulcers compared to the control group ($p<0.05$). The vibration therapy group had a significant decrease in wound area, wound depth, and wound volume, as well as an increase in granulation tissue and epithelialization, compared to the control group. Furthermore, patients in the vibration therapy group reported a significant decrease in wound pain and an improvement in quality of life compared to the control group. The findings of this study suggest that vibration therapy may be an effective adjunctive therapy for the treatment of diabetic neuropathic foot ulcers. However, it is important to note that the study's sample size is relatively small, and further research is needed to confirm these results."

B. Abstract 2

"Capacity to make one's own decisions is fundamental to the autonomy of the individual. Capacity is a functional assessment made by a clinician to determine if a patient is capable of making a specific decision. Competency is a global assessment and legal determination made by a judge in court. Capacity evaluation for a patient with dementia is used to determine whether the patient is capable of giving informed consent, participate in research, manage their finances, live independently, make a will, and have ability to drive. Patients with dementia cannot be assumed to have impaired capacity."

Even a patient with moderate or severe dementia, with obviously impaired capacity may still be able to indicate a choice and show some understanding. Four key components of decision-making in a capacity evaluation include understanding, communicating a choice, appreciation, and reasoning. Assessment of capacity requires a direct interview with the patient using open-ended questions and may include both informal and formal approaches depending on the situation and the context. A baseline cognitive evaluation with a simple test to assess executive function is often useful in capacity evaluation. All capacity evaluations are situation specific, relating to the particular decision under consideration, and are not global in scope. The clinician needs to spend adequate time with the patient and the family allaying their anxieties and also consider the sociocultural context. The area of capacity has considerable overlap with law and the clinician treating patients with dementia should understand the complexities of assessment and the implications of impaired capacity. It is also essential that the clinician be well informed and keep meticulous records. It is crucial to strike a balance between respecting the patient autonomy and acting in his/her best interest."

C. Abstract 3

"Before 2009 essentially all societies, guidelines, and statements required fasting before measuring a lipid profile for cardiovascular risk prediction. This was mainly due to the increase seen in triglycerides during a fat tolerance test. However, individuals eat much less fat during a normal day and nonfasting triglycerides have been shown to be superior to fasting in predicting cardiovascular risk. Lipids and lipoproteins only change minimally in response to normal food intake: in four large prospective studies, maximal mean changes were +0.3 mmol/L (26 mg/dL) for triglycerides, -0.2 mmol/L (8 mg/dL) for total cholesterol, -0.2 mmol/L (8 mg/dL) for LDL cholesterol, and -0.1 mmol/L (4 mg/dL) for HDL cholesterol. Further, in 108,602 individuals from the Copenhagen General Population Study in random nonfasting samples, the highest versus the lowest quartile of triglycerides, total cholesterol, LDL cholesterol, remnant cholesterol, non-HDL cholesterol, lipoprotein(a), and apolipoprotein B were all associated with higher risk of both ischaemic heart disease and myocardial infarction. Finally, lipid-lowering trials using nonfasting blood samples for assessment of lipid levels found that reducing levels of nonfasting lipids reduced the risk of cardiovascular disease.

To date there is no sound scientific evidence as to why fasting should be superior to nonfasting when evaluating a lipid profile for cardiovascular risk prediction. Indeed, nonfasting samples rather than fasting samples have many obvious advantages. First, it would simplify blood sampling in the laboratory. Second, it would benefit the patient, avoiding the inconvenience of fasting and therefore needing to have blood drawn early in

the day. Third, for individuals with diabetes, the risk of hypoglycaemia due to fasting would be minimised. Many countries are currently changing their guidelines towards a consensus on measuring a lipid profile for cardiovascular risk prediction in the nonfasting state, simplifying blood sampling for patients, laboratories, and clinicians worldwide."

D. Abstract 4

"This study aimed to examine the relationship between lipid profile and the risk of developing coronary heart disease (CHD) in 200 participants. Lipid profile, including total cholesterol, LDL cholesterol, HDL cholesterol, and triglycerides, were measured in all participants. The study found that total cholesterol levels were positively correlated with an increased risk of coronary heart disease, with an odds ratio of 1.5 (95% CI: 1.2-1.8, $p < 0.001$). HDL cholesterol levels were inversely correlated with the risk of coronary heart disease, with an odds ratio of 0.7 (95% CI: 0.5-0.9, $p < 0.05$). Specifically, the study found that for every 1 mmol/L increase in LDL cholesterol, the risk of CHD increased by 26%, and for every 1 mmol/L increase in triglycerides, the risk of CHD increased by 16%. Conversely, for every 1 mmol/L increase in HDL cholesterol, the risk of CHD decreased by 9%. These findings suggest that lipid profile may be a significant predictor of CHD risk, and that interventions aimed at improving lipid profile may be effective in reducing the risk of CHD."

Have you sure about your decisions? If so, then let us reveal the correct answers. If you think Abstract 2 and Abstract 3 are legit, and the other two are ChatGPT-generated, then you are accurate, and ChatGPT is probably not yet fully able to fool humans. Otherwise, ChatGPT, even in its current state, is already able to fool actual readers.

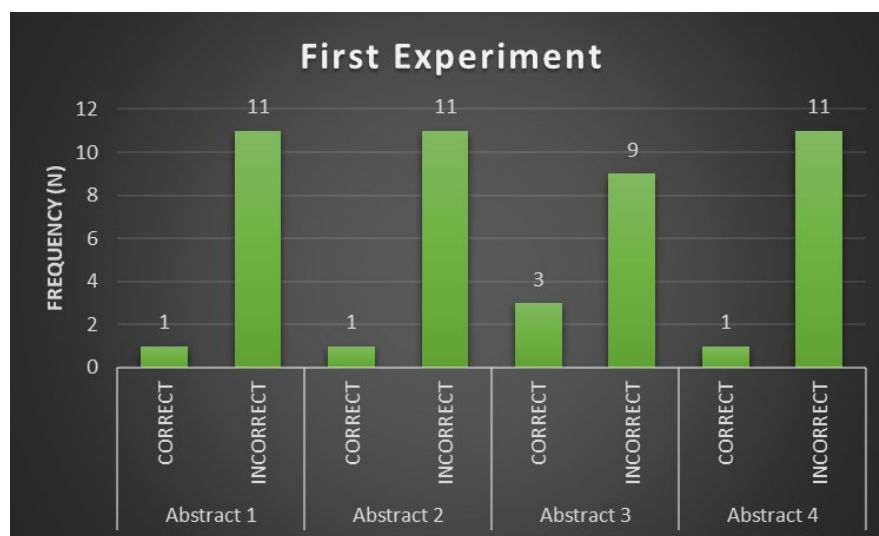


Fig. 1. The result of the first experiment.

Abstracts 2 and 3 are, respectively, obtained from the study [28] and [29], which appeared at the top of the Google Scholar list when we searched the keywords: "Capacity to Consent in Dementia" for Abstract 2 and "Lipid Profile and Cardiovascular Disease" for Abstract 3. We generated Abstract 1 by providing ChatGPT command: "Write me an abstract on an article titled *"The Effect of Vibration on the Acceleration of Wound Healing of Diabetic Neuropathic Foot Ulcer: A Prospective Experimental Study on Human Patients"*. This study involved 60 participants.". As for Abstract 4, we generated it by feeding the command: "Write me an abstract on an article titled *"the investigation of lipid profile and coronary heart disease" involving 200 participants. Please include numerical results."*

In Fig. 1, we present the result of the first experiment. As observed, among 12 participants, only one chose the correct answer for Abstract 1 (ChatGPT-generated), while the other 9 presumed that Abstract 1 was human-made from original research. On the contrary, only one believed that Abstract 2 was ChatGPT-generated while it was actually human-made. Some of the participants who incorrectly guessed stated that Abstract 1 seems realistic because it contains reasonable numerical results and logical findings. The conclusions generated by ChatGPT in Abstract 1 are also in line with previous studies. Previous research with the same title as in the command given to the ChatGPT: *"The Effect of Vibration on the Acceleration of Wound Healing of Diabetic Neuropathic Foot Ulcer: A Prospective Experimental Study on Human Patients [30]"*, suggests that vibration therapy has the potential to accelerate the wound healing. Abstract 1 became more convincing because it contains legit questionnaires such as the visual analog scale (VAS) and the Short Form-36 (SF-36) questionnaires. Abstract 1 also contains a statistical result (i.e., $p < 0.05$), making it became more difficult to identify. As for Abstract 2, it was believed that it was generated by ChatGPT since it was written using a general writing style without technical and scientific terms, without numerical numbers, and less specific. Therefore, the participants believed that such general abstracts could be easier generated by ChatGPT than the specific ones. However, this is not necessarily mean that Abstract 2 is taken from a narrative-descriptive paper, which abstract is often written using a narrative-descriptive approach.

Similar reasons can be used to explain the results for Abstracts 3 and 4. In Abstract 3, for example, only 3 out of 12 participants correctly guessed, and the other 9 provided an incorrect answer. Although Abstract 3 contains reasonable numerical results, one of those 9 participants stated that involving 108,602 individuals in such research was unusual. Thus, they thought that Abstract 3 was ChatGPT-generated. Lastly, only 1 out of 12 participants correctly guessed Abstract 4. Again, this is because Abstract 4

contains numerical results, making it convincing enough, although those numerical results are hard to be validated. Moreover, the findings, such as "*The study found that total cholesterol levels were positively correlated with an increased risk of coronary heart disease*" and "*HDL cholesterol levels were inversely correlated with the risk of coronary heart disease.*" were theoretically correct and in accordance with former studies.

In the second experiment, we provided two different abstracts, one of which was generated by the ChatGPT. Again, as the readers, you are encouraged to guess either one of the abstracts below, which is generated by OpenGPT.

E. Abstract 5

"This case report describes a patient with a rare combination of Kikuchi-Fujimoto's disease (KFD) and systemic lupus erythematosus (SLE). KFD, also known as histiocytic necrotizing lymphadenitis, is a benign, self-limiting condition characterized by cervical lymphadenopathy and fever. SLE is a chronic autoimmune disorder that affects multiple organs and systems. The patient, a 23-year-old female, presented with fever, cervical lymphadenopathy, and a rash. Laboratory tests revealed leukopenia, anemia, and thrombocytopenia, consistent with SLE. A biopsy of the lymph node showed features of KFD. Treatment with prednisolone and immunosuppressive therapy led to resolution of symptoms. This case highlights the importance of considering KFD in the differential diagnosis of patients with SLE, especially those with fever and lymphadenopathy."

F. Abstract 6

"Kikuchi-Fujimoto's disease (KFD) or histiocytic necrotising lymphadenitis is a benign and self-limited disease, of unknown aetiology, which affects mainly young women. It presents with localised lymphadenopathy, predominantly in the cervical region, accompanied by fever and leukopenia in up to 50% of the cases. KFD has been rarely described in association with systemic lupus erythematosus (SLE), and its diagnosis can precede, postdate or coincide with the diagnosis of SLE. We present a patient with the diagnosis of SLE characterised by arthritis, leukopenia, malar rash, photosensitivity and positive ANA, besides cervical lymphadenopathy whose biopsy was compatible with KFD, which improved after using prednisone. Although the presence of lymphadenopathy is not uncommon in SLE patients, particularly in the phases of disease activity, the concomitance with KFD has rarely been reported in the literature. Its recognition is necessary because one can avoid laborious investigation for infectious and lymphoproliferative diseases."

If you think that Abstract 5 is legit and Abstract 6 is generated by ChatGPT. Then you are fooled. Abstract 5 is generated through the command: write me an abstract on a case report article titled "Kikuchi-Fujimoto's disease associated with systemic lupus erythematosus", while Abstract 2 is cited from the case report with a similar title in [31]. In this experiment, 5 of our participants correctly guessed the abstract creator, while the other 7 provided incorrect guesses.

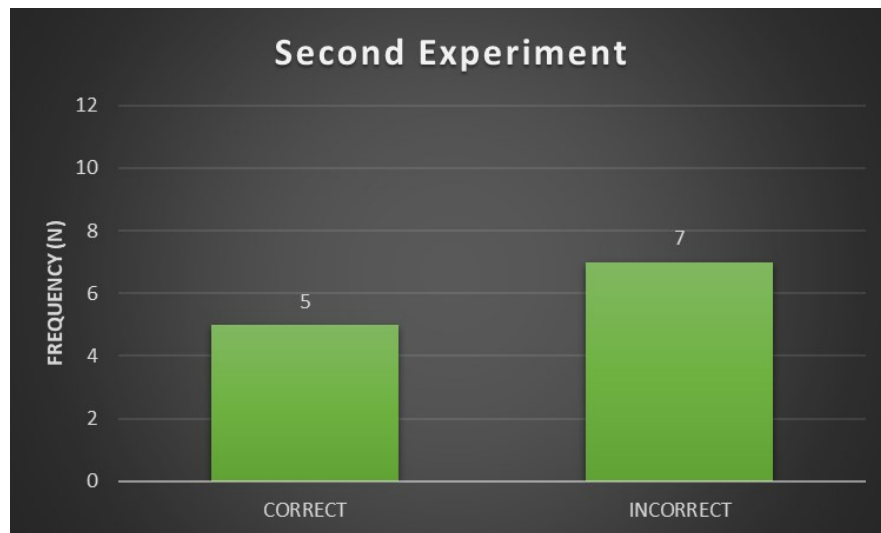


Fig. 2. The result of the second experiment.

In our experiments, we weigh the correct answer for each question as 1 point and the incorrect ones as 0 points. In total, there are 5 points for 5 items of the questionnaire. As depicted in Fig. 3., most of our participants correctly guessed only 1 out of 5 items (N=6) or did not correctly guess even a single item (N=4). The other two, however, were able to obtain 2 and 3 scores. Still, considering that the total questionnaire items were 5, being able to guess 2 or 3 out of 5 is not impressive, meaning that ChatGPT-generated abstracts are difficult enough to distinguish.



Fig. 3. The score distribution obtained from the first and second experiments.

First Discussion: ChatGPT Introduces Potential Abusive Usage in Academic Writing

Regardless of whether you are able to distinguish the ChatGPT-generated abstracts, our findings from both experiments showed that abstracts generated by ChatGPT could not easily be identified by our participants. Therefore, it is likely possible that along with its development, ChatGPT might be able to produce a completely generated paper without readers being able to distinguish. At this point, ChatGPT has potentially brought terrifying ethical disruption to the world of research.

In academic publications, artificial intelligence (AI) language model, such as ChatGPT, has the potential to be misused by irresponsible users for unethical applications [32, 33]. Following the fast development of AI technologies, it is not unrealistic that in the coming years, paper mill-generated manuscripts will become increasingly difficult to recognize and, eventually, will be impossible to detect. A well-known example of unethical conduct in academia is what we so-called papermill. An AI-based paper mill is an AI-assisted or even AI-fully automated article generator that can be used to fabricate random, illogical low-quality academic manuscripts. In its current state, such AI-generated articles can easily be spotted and differentiated if we pay enough attention to them.

Readers can easily identify problems in those kinds of AI-generated articles, such as low-quality and grammatically incorrect sentences, illogical claims, and even completely meaningless and irrelevant random sentences because the existing paper mills are far from perfect. Unfortunately, many of papermill's articles still continue to pass through the editorial systems and end up being published. Of course, the majority of the publishers that allow such papers to be published were currently, or previously, regarded as 'predatory publishers'. However, some of the world's well-known publishers, including Wiley [34], Elsevier [35], Springer Nature [36], IOP [37], Hindawi [38], IEEE [39], ACM [39], and MDPI [40], have also ever been fooled by such articles and published manuscript that paper mills most likely produced. Once identified, big publishers then retract such papers. There were an enormous number of such papers. For instance, in 2022 alone, ACM [39], Hindawi-Wiley [38], Elsevier [35, 36], and IOP [37] have retracted, respectively, >300, >500, >500, and 850 papers that a paper mill may have produced. Still, it is possible that tens to hundreds of thousands of published paper mill-generated manuscripts have yet to be identified.

ChatGPT might not be able to fully write a high-quality complicated paper with complex findings, at least in its current state. While it is feasible for ChatGPT to perform numerical calculations and model simulations, it is impossible for ChatGPT to conduct wet-lab scientific experiments, which is typical in medical research. However, as demonstrated in the above

experiment's findings, ChatGPT can generate 'random' yet reasonable numerical data related to medical research that usually can only be obtained through physical measurement. This is because similar data are already available on the internet. Due to this reason, ChatGPT might not be able to be used to create a breakthrough pseudo-wet-lab-based paper with findings that have never been available before. With that being said, however, ChatGPT still can be employed by irresponsible academia to fabricate papers (although not breakthrough ones) with fake yet reasonable numerical results and reasoning without the authors being required to conduct experiments. In addition, authors can easily utilize ChatGPT, even in its current state, to write a review or survey paper. With the rising of ChatGPT, it is possible that the already disrupted ethics in the publication world will further be broken. ChatGPT is indeed a powerful tool for irresponsible authors, or paper mills, to enhance their papers' quality so that the paper mills-generated paper eventually can no longer be detected.

ChatGPT to Answer Exam and Homework Questions: Is it Possible?

As the title of this section, we would like to know whether ChatGPT can correctly answer the typical exam and homework questions of the medical education standard. To do that, we obtained three questions that have been included in The Medical Profession Program Student Competency Test of Indonesia (*Uji Kompetensi Mahasiswa Program Profesi Dokter [UKMPPD]*) and asked ChatGPT to answer them.

A. Problem 1

Question (Translated into English):

A 40-year-old man came to the clinic with complaints of right flank pain radiating to the right lower leg. Pain accompanied by tingling to the inside of the ankle and toes of the right foot. The pain appears suddenly when the patient is working. The patient said that when pushing, the pain increases. On examination, TTV was within normal limits. The most appropriate neurological examination is...

- A. Arm drop test*
- B. Lasegue test*
- C. Wartenberg test*
- D. Finkelstein test*
- E. McMurray test*

ChatGPT's answer:

*The most appropriate neurological examination in this case is likely the **Finkelstein test**. This test is used to evaluate for de Quervain's tenosynovitis, which is a condition that causes pain and inflammation in the tendons on the thumb side of the wrist.*

This condition is often seen in people who perform repetitive motions, such as gripping or grasping, and is often accompanied by tingling in the fingers. The Finkelstein test specifically assesses the patient's ability to move the thumb in a specific way and can help identify the presence of this condition.

Remark:

The answer provided by ChatGPT is incorrect. The correct answer is Lasegue Test (Straight Leg Raising Test = SLRT). As for the available answers, the arms drop test is the examination for the rotator cuff tear, the Wartenberg test is for myasthenia gravis, and the Finkelstein test and McMurray test are, respectively, the examination for de Quervain syndrome and meniscus injury. For this matter, ChatGPT is able to generate the correct definition of the Finkelstein test and its usability, although not perfect. Finkelstein test does not only assesses the patient's ability to move the thumb. In the Finkelstein test, the patient is instructed to make a fist around a thumb and perform an ulnar deviation.

B. Problem 2

Question (Translated into English):

A 40-year-old man comes to the clinic with a history of gout. There was a history of previous pain, given anti-pain medication, then the complaints disappeared. History of recurrent pain complaints. On physical examination, vital signs were within normal limits, at MTP 1 dextra was found tofus, hyperemia (-). The uric acid laboratory test result is 9 mg/dL. What is the right therapy to lower uric acid levels?

- A. Kodein*
- B. Meloxicam*
- C. Paracetamol*
- D. Allupurinol*
- E. Kolkisin*

ChatGPT's answer:

*Based on the patient's symptoms, diagnosis of gout and laboratory results, the most appropriate therapy to lower uric acid levels would be **D. Allopurinol**. Allopurinol is a xanthine oxidase inhibitor that reduces the production of uric acid in the body. This helps to prevent the formation of uric acid crystals in the joints and can also help to dissolve existing crystals. It is a first-line therapy for gout.*

A.Codeine, B. Meloxicam, C. Paracetamol are not used to lower uric acid levels. They are primarily used to manage pain and inflammation caused by gout. E. Kolkisin is not a medication that is commonly used in the treatment of gout.

It is important to note that the treatment of gout requires long-term therapy, and may include a combination of lifestyle changes and medications. Regular follow-up and monitoring of uric acid levels is necessary to adjust the therapy if needed.

Remark:

The answer provided by ChatGPT is correct. Codeine (antitussive), Meloxicam (analgesic), and Paracetamol (analgesic) are not used to lower uric acid levels. Kolkisin (analgesic) is, however, commonly used in the inflammation treatment of gout in many countries.

C. Problem 3

Question (Translated into English):

A 20-year-old man came with complaints of fever for 7 days. complaints accompanied by difficulty defecating and decreased appetite. On physical examination, a dirty tongue, epigastric tenderness, and hepatosplenomegaly were found. What is the causative agent of the case?

A. Escherichia Coli

B. Salmonella Thypi

C. Vibrio cholera

D. Rotavirus

E. Helicobacter Pillory

ChatGPT's answer:

*Based on the symptoms of fever, difficulty defecating, decreased appetite, dirty tongue, epigastric tenderness and hepatosplenomegaly, it is likely that the causative agent of the case is a **bacterial infection**.*

Bacteria such as Escherichia coli, Salmonella typhi, and Vibrio cholerae can all cause similar symptoms. Helicobacter Pylori is a bacteria that can cause stomach ulcers and gastritis, can also cause similar symptoms such as abdominal pain, nausea, vomiting, loss of appetite and weight loss. It is necessary to perform further examination such as blood culture, stool culture, serological test and examination to determine the specific causative agent.

It is also important to note that D. Rotavirus is a viral infection that can cause diarrhea, vomiting, and fever but it is unlikely to be the causative agent based on the symptoms mentioned.

Remark:

The answer is correct, but not necessarily. While the further examination is required, among the available options, it is highly likely that *Salmonella Thyphi* is the causative agent of the above case.

Fortunately (or unfortunately), the ChatGPT can only correctly answer 1 question, indicating that ChatGPT is not yet knowledgeable enough to provide accurate answers and logical reasoning behind them. However, this does not mean that ChatGPT will forever not be able to resolve these kinds of questions. As the information consumed by ChatGPT grows, it is very likely that in the near future, ChatGPT will be able to provide correct answers for given exam and homework, which raises other questions: Does it mean exam and homework are now redundant? As an educator, do we still need them? How can we maintain the quality of education and make sure that the students will not cheat?

**Second Discussion: Exam and Homework in the Medical Education Standard.
Are They Still Needed?**

In our humble opinion, homework is no longer necessary in the modern medical education standard. This is mainly because homework is usually designed as a means to encourage the student to further learn the educational material in addition to the regular class. Previously, the difference between the outcomes of diligent and lazy students can be seen in their homework. Diligent students will try their best to solve homework problems, while the lazy ones will not or will do homework poorly. However, with the availability of AI tools such as OpenGPT, these differences will diminish. This is due to the probability that any medical education problems can be solved by AI tools in the coming years. In addition, it is possible that lazy students will obtain better homework scores, thanks to AI tools because the diligent ones do not bother to use such tools.

Indeed technology-assisted homework solvers already existed decades ago. However, so far, the abusive use of such technology can be easily detected. For instance, students can easily find the solutions to most problems in search engines (e.g., Google). However, they cannot just 'copy' and 'paste' such solutions since the answers can be easily detected by a plagiarism checker. As a result, they might want to paraphrase their answers. There are indeed paraphrasing tools. However, the results are often far from decent, making the students need to re-read them and fix them carefully. These re-read and fixing activities mean that the homework intention is successfully delivered (i.e., learning).

Another argument supporting the statement that homework is no longer required in the medical education standard is that medical doctors shall need practical skills, which cannot be obtained through theoretical homework. In addition, the fact that they will face patients in the future is often enough to encourage them to study, regardless of whether there is any homework or not. Further, to be a doctor, medical students must pass the national examination. Hence, not properly studying will bring consequences to them.

As for the exam, we believe that exam is still an appropriate tool to assess students' skills and knowledge, given that an appropriate exam design is carried out. Moreover, exams are proven to be an effective way to motivate students to learn educational materials. Even though online assignments and exams were solutions during the pandemic setting, we believe that online exams or take-home exams (or assignments, in this context) are no longer recommended since it will be difficult for the educator to maintain exam ethics and fairness. As an example, Australian universities returned to 'pen-and-paper' exams after students were caught using AI to write essays [41]. Additionally, online or take-home exams cannot be used to assess the laboratory and practical skills of the students, which are critical in medical doctor professions. Finally, an appropriate real-time offline exam with volunteers acting as patients is still preferred in medical education. In Indonesia, this exam is carried out through Objective Structured Clinical Examination (OSCE).

Conclusion

ChatGPT is a powerful language model developed by OpenAI that has the ability to generate human-like text. Its natural language processing capabilities make it a valuable tool for various applications such as customer service, content generation, and language translation. However, its capabilities also make it a double-edged sword, with the potential for both positive and negative impacts in various sectors, including education.

In the education sector, ChatGPT can be used as a powerful tool for creating educational content and assisting in language learning. It can generate essays, summaries, and even entire textbooks with little human input, making it a valuable asset for educators and students alike. Additionally, its ability to generate text in multiple languages can assist in language learning and translation.

However, ChatGPT's capabilities also open up the potential for unethical applications in education. One such example is the use of the model to generate plagiarized content. Students could use ChatGPT to generate essays and other written assignments, presenting them as their own work. This not only undermines the integrity of the education system but also does not help the students to learn.

Another ethical concern is the potential for ChatGPT to be used for test-taking in online education. With its ability to understand and respond to natural language, ChatGPT could be used to cheat on online exams, potentially invalidating the results and undermining the integrity of the education system.

Although, as shown in previous studies, as well as demonstrated in this work, the current ChatGPT is not yet fully capable of being constituted as a threat. However, with the rapid development of language models (e.g., ChatGPT), and along with the growth of the information munched by them, it is highly possible that ChatGPT, or any other large language models, can resolve their current weakness and become a substantial threat to the education ethics.

In this article, it is demonstrated that ChatGPT, although it has not been able to correctly answer standard medical education questions, its generated abstracts have already been able to fool the study's participants. However, it is worth noting that the standard medical education questions used in this study were randomly taken from the questions from Indonesia's medical profession program student competency test. It is also worth noting that the participants in this study were all Indonesian and worked in Indonesia with Bahasa Indonesia as their daily language. Therefore, the results of this study cannot necessarily be generalized for a different set of questions or broader populations. Lastly, since there were no subjects' identities, other than the latest academic degrees and occupations (medical doctor or university lecturer), gathered during the study, there is no way to verify the accuracy of the study. Future studies with a larger sample size with more heterogeneous backgrounds may be conducted. Still, the findings of this study provide a hint that ChatGPT's capabilities to generate abstracts are beyond distinguishable.

In conclusion, large language models such as ChatGPT are powerful tools with the potential to revolutionize the education sector. However, its capabilities also open up the potential for unethical applications, such as cheating and plagiarism. It is important to consider the potential ethical implications of using such tools in education and to develop guidelines and regulations to ensure their responsible use.

References




- [1] King, M. R. (2022). The future of AI in medicine: a perspective from a Chatbot. *Annals of Biomedical Engineering*, 1-5.
- [2] Kirmani, A. R. (2022). Artificial Intelligence-Enabled Science Poetry. *ACS Energy Letters*, 8, 574-576.
- [3] Zhai, X. (2022). ChatGPT user experience: Implications for education. *Available at SSRN 4312418*.
- [4] Aydın, Ö., & Karaarslan, E. (2022). OpenAI ChatGPT generated literature review: Digital twin in healthcare. *Available at SSRN 4308687*.
- [5] Cahan, P., & Treutlein, B. (2023). A conversation with ChatGPT on the role of computational systems biology in stem cell research. *Stem Cell Reports*, 18(1), 1-2.
- [6] Graham, F. (2022). Daily briefing: Will ChatGPT kill the essay assignment?. *Nature*.

- [7] Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity?. *arXiv preprint arXiv:2212.09292*.
- [8] Zhang, B., Ding, D., & Jing, L. (2022). How would Stance Detection Techniques Evolve after the Launch of ChatGPT?. *arXiv preprint arXiv:2212.14548*.
- [9] Frye, B. L. (2022). ChatGPT Talks About Animal Rights. *Available at SSRN*.
- [10] Zaremba, A., & Demir, E. (2023). ChatGPT: Unlocking the Future of NLP in Finance. *Available at SSRN 4323643*.
- [11] Oxford Analytica. (2023). Commercial use of ChatGPT requires caution. *Emerald Expert Briefings, (oxan-es)*.
- [12] Castelveccchi, D. (2022). Are ChatGPT and AlphaCode going to replace programmers?. *Nature*.
- [13] *Professor catches student cheating with ChatGPT: 'I feel abject terror'*. (2022). NY Post. <https://nypost.com/2022/12/26/students-using-ChatGPT-to-cheat-professor-warns> (accessed on 2 January 2023).
- [14] Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays-should academics worry?. *Nature*.
- [15] de Winter, J. C. F. (2023). Can ChatGPT Pass High School Exams on English Language Comprehension?.
- [16] *AI-generated answers temporarily banned on coding Q&A site Stack Overflow*. (2022). The Verge. <https://www.theverge.com/2022/12/5/23493932/ChatGPT-ai-generated-answers-temporarilybanned-stack-overflow-llms-dangers> (accessed on 2 January 2023).
- [17] Kim, N., Htut, P. M., Bowman, S. R., & Petty, J. (2022). \$^ 2\$: Question Answering with Questionable Assumptions. *arXiv preprint arXiv:2212.10003*.
- [18] *ChatGPT produces made-up nonexistent references*. (2022). Brian Lovin. <https://brianlovin.com/hn/33841672> (accessed on 3 January 2023).
- [19] *A computer can now write your college essay — Maybe better than you can*. (2022). Forbes. <https://www.forbes.com/sites/emmawhitford/2022/12/09/a-computer-can-now-write-your-collegeessay---maybe-better-than-you-can/?sh=35deca9ddd39> (accessed on 4 January 2023).
- [20] Kung, T. H., Cheatham, M., Medinilla, A., ChatGPT, Sillos, C., De Leon, L., ... & Tseng, V. (2022). Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. *medRxiv*, 2022-12.
- [21] O'Connor, S., & ChatGPT. (2023). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Education in Practice*, 66, 103537. <https://doi.org/10.1016/j.nepr.2022.103537>
- [22] Zhavoronkov, A. (2022). Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective. *Oncoscience*, 9, 82.
- [23] GPT Generative Pretrained Transformer, Almira Osmanovic Thunström, Steinn Steingrímsson. Can GPT-3 write an academic paper on itself, with minimal human input?. 2022. <hal-03701250>
- [24] Stokel-Walker C. (2023). ChatGPT listed as author on research papers: many scientists disapprove. *Nature*, 10.1038/d41586-023-00107-z. Advance online publication. <https://doi.org/10.1038/d41586-023-00107-z>
- [25] *Final Rule Human Subjects Research Exemptions- NIH Infographic*. (2019). National Institutes of Health (NIH). https://grants.nih.gov/sites/default/files/exemption_infographic_v8_508c_1-15-2020.pdf (accessed on 4 February 2023).
- [26] *Subpart A. Basic HHS Policy for Protection of Human Research Subjects*. (2018). Office for Human Research Protections. <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/common-rule-subpart-a-46104/index.html> (accessed on 4 February 2023).
- [27] Walch-Patterson, A. (2020). Exemptions and limited institutional review board review: A practical look at the 2018 common rule requirements for exempt research. *Ochsner Journal*, 20(1), 87-94.
- [28] Hegde, S., & Ellajosyula, R. (2016). Capacity issues and decision-making in dementia. *Annals of Indian Academy of Neurology*, 19(Suppl 1), S34.
- [29] Langsted, A., & Nordestgaard, B. G. (2019). Nonfasting versus fasting lipid profile for cardiovascular risk prediction. *Pathology*, 51(2), 131-141.
- [30] Syabariyah, S., Nurachmah, E., Widjojo, B. D., Prasetyo, S., Sanada, H., Nakagami, G., ... & Hisan, U. K. (2023, January). The Effect of Vibration on the Acceleration of Wound Healing of Diabetic Neuropathic Foot Ulcer: A Prospective Experimental Study on Human Patients. In *Healthcare* (Vol. 11, No. 2, p. 191). Multidisciplinary Digital Publishing Institute.
- [31] Santana, A., Lessa, B., Galvão, L., Lima, I., & Santiago, M. (2005). Kikuchi-Fujimoto's disease associated with systemic lupus erythematosus: case report and review of the literature. *Clinical Rheumatology*, 24(1), 60-63.
- [32] Hisan, U. K., & Amri, M. M. (2022). Artificial Intelligence for Human Life: A Critical Opinion from Medical Bioethics Perspective–Part I. *Journal of Public Health Sciences*, 1(02), 100-111.




- [33] Hisan, U. K., & Amri, M. M. (2022). Artificial Intelligence for Human Life: A Critical Opinion from Medical Bioethics Perspective–Part II. *Journal of Public Health Sciences*, 1(02), 112-130.
- [34] *Exclusive: Hindawi and Wiley to retract over 500 papers linked to peer review rings.* (2022). Retraction Watch. <https://retractionwatch.com/2022/09/28/exclusive-hindawi-and-wiley-to-retract-over-500-papers-linked-to-peer-review-rings/> (accessed on 17 December 2022).
- [35] *Exclusive: Elsevier retracting 500 papers for shoddy peer review.* (2022). Retraction Watch. <https://retractionwatch.com/2022/10/28/exclusive-elsevier-retracting-500-papers-for-shoddy-peer-review/> (accessed on 17 December 2022).
- [36] *A tale of 32 retraction notices: On publishers, paper mill products, and the sleuths that find them.* (2022). Retraction Watch. <https://retractionwatch.com/2022/08/15/a-tale-of-32-retraction-notices-on-publishers-paper-mill-products-and-the-sleuths-that-find-them/> (accessed on 17 December 2022).
- [37] *Physics publisher retracting nearly 500 likely paper mill papers.* (2022). Retraction Watch. <https://retractionwatch.com/2022/09/09/physics-publisher-retracting-nearly-500-likely-paper-mill-papers/> (accessed on 17 December 2022).
- [38] *Exclusive: Hindawi and Wiley to retract over 500 papers linked to peer review rings.* (2022). Retraction Watch. <https://retractionwatch.com/2022/09/28/exclusive-hindawi-and-wiley-to-retract-over-500-papers-linked-to-peer-review-rings/> (accessed on 17 December 2022).
- [39] *More than 300 at once: Publisher retracts entire conference proceedings.* (2022). Retraction Watch. <https://retractionwatch.com/2022/04/20/more-than-300-at-once-publisher-retracts-entire-conference-proceedings/> (accessed on 17 December 2022).
- [40] *'Our deepest apology': Journal retracts 30 likely paper mill articles after investigation published by Retraction Watch.* (2022). Retraction Watch. <https://retractionwatch.com/2022/07/05/our-deepest-apology-journal-retracts-30-likely-paper-mill-articles-after-investigation-published-by-retraction-watch/> (accessed on 17 December 2022).
- [41] *Australian universities to return to 'pen and paper' exams after students caught using AI to write essays.* (2023). The Guardian. <https://www.theguardian.com/australia-news/2023/jan/10/universities-to-return-to-pen-and-paper-exams-after-students-caught-using-ai-to-write-essays> (accessed on 20 January 2023).

Authors



Urfa Khairatun Hisan    is a lecturer at the Faculty of Medicine, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, and a graduate student of the Department of Bioethics, Universitas Gadjah Mada, Indonesia. She received her B.Med. and M.D. degrees from the Faculty of Medicine, Universitas Gadjah Mada, Indonesia, in 2017 and 2019, respectively. Her research interests include public health and bioethics in medicine. She can be contacted at email: urfa.hisan@med.uad.ac.id.



Muhammad Miftahul Amri    received his B.S. degree from the Department of Computer Science and Electronics, Universitas Gadjah Mada Indonesia in 2018, and the M.S. degree from the Department of Electrical and Computer Engineering, Sungkyunkwan University South Korea in 2021, where he is currently pursuing the Ph.D. degree. In 2022, he received the M.M. degree and professional engineer title, respectively, from Universitas Terbuka and Universitas Muhammadiyah Yogyakarta Indonesia. In 2021, he joined the faculty at Universitas Ahmad Dahlan, Indonesia, where he is currently a lecturer in the Department of Electrical Engineering. His research interests include wireless communications and reconfigurable intelligent surfaces. He can be contacted through: muhammad.amri@te.uad.ac.id.