BioData Mining

RESEARCH Open Access

Comparing new tools of artificial intelligence to the authentic intelligence of our global health students

Shilpa R. Thandla[1†], Grace Q. Armstrong[1†], Adil Menon[2], Aashna Shah[1], David L. Gueye[1], Clara Harb[1,3], Estefania Hernandez[1], Yasaswini Iyer[1], Abigail R. Hotchner[1], Riddhi Modi[1], Anusha Mudigonda[1], Maria A. Prokos[4], Tharun M. Rao[1], Olivia R. Thomas[1], Camilo A. Beltran[1], Taylor Guerrieri[5], Sydney LeBlanc[1], Skanda Moorthy[1], Sara G. Yacoub[1], Jacob E. Gardner[1,2], Benjamin M. Greenberg[5], Alyssa Hubal[6,7], Yuliana P. Lapina[5], Jacqueline Moran[5], Joseph P. O'Brien[1,2], Anna C. Winnicki[6,8], Christina Yoka[1,9], Junwei Zhang[1] and Peter A. Zimmerman[1,6,8*]

†Shilpa R. Thandla and Grace Q. Armstrong contributed equally to this work.

*Correspondence: paz@case.edu

[1] Master of Public Health Program, Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, USA
[2] School of Medicine, Case Western Reserve University, Cleveland, OH, USA
[3] Bioethics Department, Case Western Reserve University, Cleveland, OH, USA
[4] Nutrition Department, Case Western Reserve University, Cleveland, OH, USA
[5] Anthropology Department, Case Western Reserve University, Cleveland, OH, USA
[6] Pathology Department, Case Western Reserve University, Cleveland, OH, USA
[7] Division of Infectious Diseases and HIV Medicine, Cleveland, OH, USA
[8] Center for Global Health and Diseases, Cleveland, OH, USA
[9] Department of Public Health, Cleveland, OH, USA

Abstract

Introduction: The transformative feature of Artificial Intelligence (AI) is the massive capacity for interpreting and transforming unstructured data into a coherent and meaningful context. In general, the potential that AI will alter traditional approaches to student research and its evaluation appears to be significant. With regard to research in global health, it is important for students and research experts to assess strengths and limitations of GenAI within this space. Thus, the goal of our research was to evaluate the information literacy of GenAI compared to expectations that graduate students meet in writing research papers.

Methods: After completing the course, Fundamentals of Global Health (INTH 401) at Case Western Reserve University (CWRU), Graduate students who successfully completed their required research paper were recruited to compare their original papers with a paper they generated by ChatGPT-4o using the original assignment prompt. Students also completed a Google Forms survey to evaluate different sections of the AI-generated paper (e.g., Adherence to Introduction guidelines, Presentation of three perspectives, Conclusion) and their original papers and their overall satisfaction with the AI work. The original student to ChatGPT-4o comparison also enabled evaluation of narrative elements and references.

Results: Of the 54 students who completed the required research paper, 28 (51.8%) agreed to collaborate in the comparison project. A summary of the survey responses suggested that students evaluated the AI-generated paper as inferior or similar to their own paper (overall satisfaction average = 2.39 (1.61–3.17); Likert scale: 1 to 5 with lower scores indicating inferiority). Evaluating the average individual student responses for 5 Likert item queries showed that 17 scores were < 2.9; 7 scores were between 3.0 to 3.9; 4 scores were ≥ 4.0, consistent with inferiority of the AI-generated paper. Evaluation of reference selection by ChatGPT-4o ($n = 729$ total references) showed that 54%

($n = 396$) were authentic, 46% ($n = 333$) did not exist. Of the authentic references, 26.5% (105/396) were relevant to the paper narrative; 14.4% of the 729 total references.

Discussion: Our findings reveal strengths and limitations on the potential of AI tools to assist in understanding the complexities of global health topics. Strengths mentioned by students included the ability of ChatGPT-4o to produce content very quickly and to suggest topics that they had not considered in the 3-perspective sections of their papers. Consistently presenting up-to-date facts and references, as well as further examining or summarizing the complexities of global health topics, appears to be a current limitation of ChatGPT-4o. Because ChatGPT-4o generated references from highly credible biomedical research journals that did not exist, our findings conclude that ChatGPT-4o failed an important component in using information effectively. Moreover, misrepresenting trusted sources of public health information is highly concerning, particularly given recent experiences from the COVID-19 pandemic and more recently in reporting on the impact of, and response to natural disasters. This is a significant limitation of GenAI's ability to meet information literacy standards expected of graduate students.

Introduction

The concept of artificial intelligence (AI) began to emerge in the 1950s, with Alan Turing's Imitation Game and Isaac Asimov's science fiction novel, *I, Robot*, [1, 2]. The birth of formal research on AI has been traced to the 1956 Dartmouth Conference, where interests of John McCarthy, Marvin Minsky, Allen Newell and Herbert A. Simon first coalesced [3]. After decades primarily occupying the academic world, generative AI (GenAI) exploded into the mainstream in November 2022 when the large language model (LLM) ChatGPT was released for open public use by OpenAI [4]. As of July 2024, the top three AI tools now used include ChatGPT (3 billion monthly visits (53.5% market share), 180 million users) Canva (833 million monthly visits (14.86% MS), 170 million users) and Google Gemini (316 million monthly visits (5.65% MS), 100 million users) [5]. The transformative feature of this technology is the massive capacity for interpreting and transforming unstructured data into a coherent and meaningful context. Given this potential, there is significant concern that AI will be disruptive to traditional instruction/evaluation of academic research [6, 7] and professional communications [8].

As was evident during the COVID-19 pandemic, evaluating and interpreting global health information challenged everyone with interests in human health from the general public to academic (students and faculty), government leaders, and policy makers [9–13]. This is all the more challenging given that the global health landscape is continuously changing and dependent on simultaneously understanding hyper-local, national and international contexts. Improving global health awareness and more effective public health response, consequently requires more effective communication across this complex landscape [14–18].

Engaging in the challenge of global health education requires that students possess both broad skill sets and societal awareness [19–25]. Meaningful participation within the global health classroom environment requires that students have working fluency with technical disciplines of infectious diseases, nutrition, biotechnology, epidemiology, mathematical modeling, climate science and engineering. Well-developed communication

and computational skills are also essential. Furthermore, students must be able to engage appropriately on topics with inherent geographical, cultural, socio-economic, biomedical, and political complexities prioritizing the dignity of all people irrespective of disparities in access to health care. With these expectations, advanced undergraduate and graduate students are best suited for global health courses.

As GenAI tools are rapidly proliferating, with many free to access, its use is having a significant impact on what is written around the world. Across the biomedical arena, a PubMed search found that ChatGPT now appears in the title of 3,506 published manuscripts (1,626 in 2023; 1,880 through October 29, 2024). How GenAI impacts a critically important global health narrative opens a knowledge gap that, so far, has relatively few published manuscripts (PubMed search with ChatGPT and "global health" All Fields = 59 published manuscripts; 24 in 2023; 35 through October 29, 2024). Therefore, this study evaluating how GenAI-generated content would compare with graduate student research papers sought to investigate how ChagGPT-4o would perform relative to a number of unique, current global health concerns. For this comparison, we were mindful of both punitive (guidelines regarding "Definitions of Violations" that customarily focus on plagiarism, misrepresentation of submitted work and/or obstruction of other students' scholarly work) as well as more constructive perspectives for our evaluations.

From the constructive position, as global health scholars, we focused on information literacy (IL – concept stewardship falls largely within the realm of library and information science (LIS) profession) [26]. The evolution of IL deserves further literature review as it influences standards of competency across many fields [27] and these standards underlie methods for teaching students how to write research papers. In coining the term, Paul Zurkowski (President of the Information Industry Association, 1974) contrasted literacy, the ability to read and write, with IL as further molding of information [26]. Basic principles of IL include, recognizing when information is needed and having the ability to locate, evaluate and use the needed information effectively [28, 29]. Effective use of information can be evaluated from corporate (e.g., product development and protection, patenting) and academic (e.g., all phases of which emphasize accurately acknowledging (citing) sources of information) worlds [30]. Over time, IL has required specific expansion related to grade-level and subject matter, and to keep pace with the evolution of new technologies (e.g., computer, digital, internet, media, news) [31, 32]. Information literacy takes on added importance in the setting of global health due to differences among international states. To emphasize the importance of an information literate society on the global level, the United Nations Educational, Scientific and Cultural Organization (UNESCO) partnered with the International Federation of Library Associations and Institutions (IFLA) to convene "expert meetings" that culminated in the Prague Declaration (2003) [33] and the Alexandria Proclamation (2005) [34]. These documents emphasize that IL is a basic human right "assisting individuals and their institutions to meet technological, economic and social challenges, to redress disadvantage and to advance the well-being of all [34]." Given these priorities, the latest advances in information technology giving rise to GenAI tools must also be evaluate in the context of IL.

Methods

Fundamentals of global health course

Course

Fundamentals of Global Health (INTH 301/401) is offered to third and fourth-year undergraduate students and graduate students across the University and is a requirement for Department-specific Certificates in Global Health and for students in the Master of Public Health, Global Health Concentration. The course integrates multiple perspectives in global health by investigating how the disciplines of Biology, Bioethics, Epidemiology, GIS, Molecular Diagnostics and Bioinformatics, Mathematics, Anthropology, Nursing, Social Work, Environmental Science, Engineering, Medicine, Bioethics and Public Health analyze and approach intersecting international health problems. This is achieved through course modules organized by faculty subject matter experts that incorporates background materials from different peer-reviewed sources (e.g., journals, monographs, podcasts, blogs). In this interdisciplinary context, students are encouraged to develop a shared vocabulary to understand these multiple perspectives from within (and outside) their own discipline. The course emphasizes issues related to international health agencies, health consequences of development projects, emergency response to health care crises, climate change and diseases of historical, present, and emerging global health importance. Current and emerging narratives integrate course presentations and assessment activities.

Research paper

Graduate students taking INTH 401 are required to write a research paper.

In brief, the general instructions for the research paper are as follows. The paper should be 10 – 15 pages of text (2,500 to 3,750 words) exclusive of references, tables or figures and should discuss an infectious disease of global health importance in a specific place (a country is acceptable, region in a country, a specific city or village preferred) from 3 perspectives (e.g., (1) the biology of the disease, (2) spatial epidemiology and (3) the ethics of vaccine distribution). Each perspective should then be presented in a separate section with its own separate heading. The final section should synthesize how these perspectives are related to each other and thereby state how they are integrated. All research papers must include a minimum of 25 references from the peer-reviewed literature. Wikipedia and websites are not acceptable references. From 2021 to 2024 specific statements were made to discourage any use of artificial intelligence. Research papers were screened for plagiarism using Turnitin software provided through the CWRU Canvas Wizard.

Comparative research paper assignment

Pilot project

During the 2024 Spring semester offering of INTH 301/401, the course instructor first followed the INTH 401 research paper prompt to generate a paper using ChatGPT 3.5 and Perplexity AI. During this same time frame, after student papers were written following traditional guidelines to avoid plagiarism (no AI assistance other than conventional spelling and grammar checking available through word processing and

Google Docs software), two students were asked to use an AI virtual assistant of their choosing to write their same paper research papers using the assignment guidelines as the prompt. Results were shared with the course instructor and teaching assistants from the 2023 and 2024 course offerings and used to evaluate the feasibility of the intended project and to optimize the study design; this work was not graded.

Comparative AI paper

In the pilot project AI chatbots ChatGPT-3.5, Perplexity AI, and EssayGenius, were used. We ultimately selected ChatGPT-4o for the next phase of the project. Instructions for generating the AI-generated paper, closely followed those provided below (Fig. 1, Supplemental Method 1). Based on experience from the pilot project, additional prompts were suggested to emphasize the page/word limit and reference inclusion. This level of guidance is known as zero-shot prompting [35–37], wherein the AI tool receives a task description in a prompt that lacks labeled data for training on specific input–output mappings.

The student's original research paper (following Research Paper methods, above) and the comparative AI paper (following Comparative AI Paper methods, above and in Methods Supplement 2) were similar in that they used the same prompt to guide/develop/compose their two papers. The papers were different in that the students' original paper (written following traditional scholarly methods) was written during the time they were enrolled in the course, and they received a grade for their original work. The Comparative AI Paper was generated using ChatGPT-4o after they had completed the course, and the AI paper was not graded.

Recruitment and incentivizing students

Students from 2021 – 2024 were recruited ($n = 54$) using their student email addresses provided through the CWRU Canvas Wizard. Inactivation of student email addresses was evident for seven students based on a mail delivery system bounce-back message. The recruitment email to students (Supplemental Method 2) provided a brief overview regarding the motivation from the project and the goal of (1) producing a poster for presentation at the March 2025 annual meeting of the Association for Prevention Teaching and Research, and (2) writing a paper to submit to a relevant journal for peer review. As with the Pilot Project, student author participation was not a part of the assigned course work and was not graded.

Evaluation of paper narrative and student surveys

Before the students submitted their ChatGPT4o-generated papers, they were asked to read and then evaluate the paper narrative and then compare it using the queries in a Google Forms survey (Supplemental Method 3) and submit their ChatGPT4o-generated papers (in full) to the course instructor and TAs by email or as a Google Doc.

The Google Forms Survey included queries to evaluate—Adherence to Introduction Guidelines (Yes/No), Evaluation of Three Perspectives, Summary and Overall Quality (5 Likert scale items: 1 = Significantly inferior; 2 = Inferior; 3 = Similar; 4 = Somewhat better; 5 = Significantly better)). The survey also included the following open-ended queries.

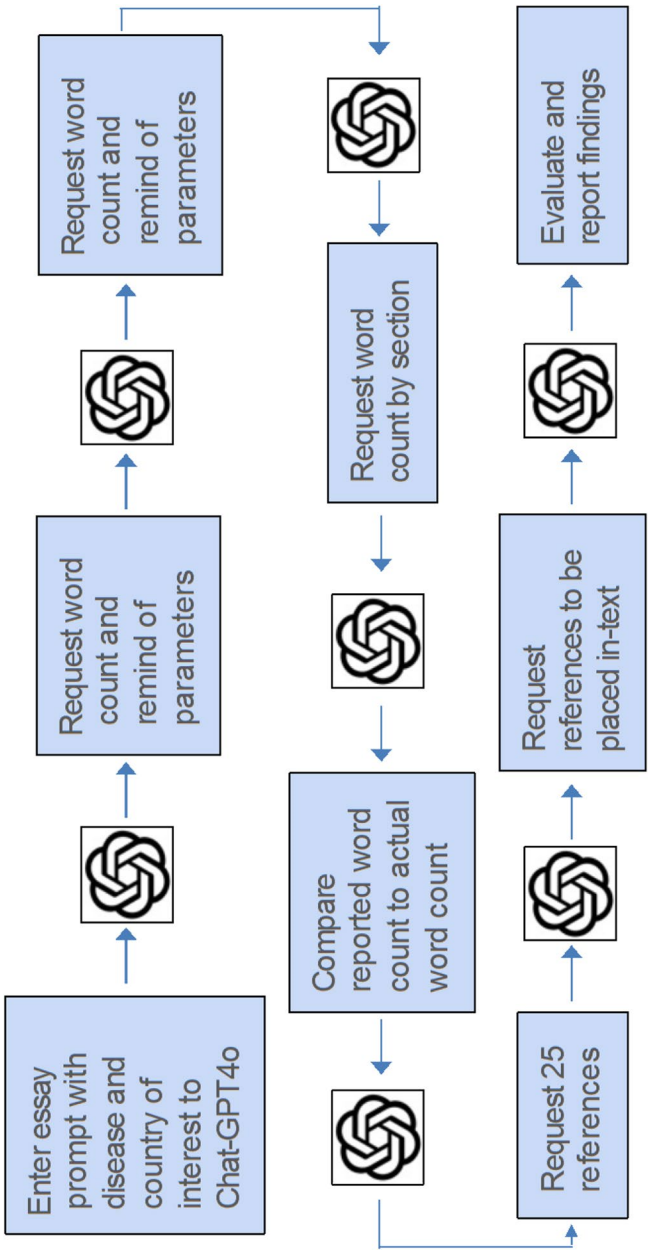

Fig. 1 Zero-shot prompt used by students to guide ChatGPT-4o through generation of a global health research paper conforming to the assignment guidelines

- How familiar are you with AI tools (which tools do you use and if so, in what context)? How do you envision AI tools helping you in your own future research?
- Please reflect on this project in terms of your expectations and comparing your own work to the ChatGPT-4o-generated output.

Evaluation of student and ChatGPT-4o-selected references

The following steps were followed to evaluate the authenticity of cited references. If a website or doi number was provided, this content was entered into the Google search window and the search was initiated. If information was provided to identify a specific journal article, the Journal name/volume/issue/pages information was entered into the Google search window and the search was initiated. If this search strategy did not reveal the anticipated article, additional searches were attempted using Google Scholar and PubMed. When difficulties were encountered using information provided by students or ChatGPT-4o, Journal name/volume/issue/pages, information was used at the Journal website to attempt to authenticate the content appearing in the journal at the suggested destination by searching through the journal's archived material (if available).

As a first round to evaluate the authenticity of references cited by students and ChatGPT-4o, journals and online materials were queried via Google. Journals were further evaluated for current impact factor (IF) data using the online journal IF search engine, Bioxbio. All IF metrics were compared to Journal landing page information for concordance. Data for all students and ChatGPT-4o references were entered into a spreadsheet to facilitate further assessments and comparisons.

Finally, as with grading of student papers, references were examined in the sentence/paragraph structure where they appeared in the narrative. Therefore, both authenticity and relevance factored into the overall assessment of both papers.

Data analysis

Information pertaining to all references was entered into a Microsoft Excel spreadsheet to enable preliminary comparisons and descriptive analyses. Statistical tests were performed by R.

Institutional guidance from Case Western Reserve University (CWRU) on the use of artificial intelligence in the classroom

At the time this study was conducted the guidance to faculty on integrating AI (artificial intelligence) into the CWRU classroom is summarized briefly, below; a link to further guidance is provided here.

1. Faculty members have complete discretion regarding the extent to which they will allow AI tools to be used by their students.
2. Faculty should clearly communicate expectations regarding the use of AI tools to students in all course syllabi.
3. Faculty must enforce high ethical standards for students' academic conduct. The university's various Academic Integrity policies prohibit academic dishonesty, including misrepresentation of a student's own work.

4. Faculty may take advantage of software tools that can detect the use of AI by students in their work.

Based on discussions with faculty colleagues who presented module content and the Senior Associate Dean of Graduate Studies, this study did not qualify as human subjects research and therefore did not require Institutional Review Board approval.

All students agreed with the project protocol and consented to participate (Supplemental Method 1).

Results and Discussion

Overview of student research papers

From 2021 to 2024, 117 students (60 undergraduate, 57 graduate) have completed Fundamentals of Global Health. Writing the research paper that is the focus of this project is required of graduate students only. Of 54 completed research papers, 28 (51.8%) students accepted the invitation to collaborate as authors (Supplemental Table 1). Infectious diseases of global health consequence and the regional concentrations of the research papers are summarized in Fig. 2. This included 20 viral infections, 5 bacterial infections and 3 parasitic infections. High representation of papers focused on COVID-19 stemmed from student interests in the pandemic during the 2021 and 2022 course offerings.

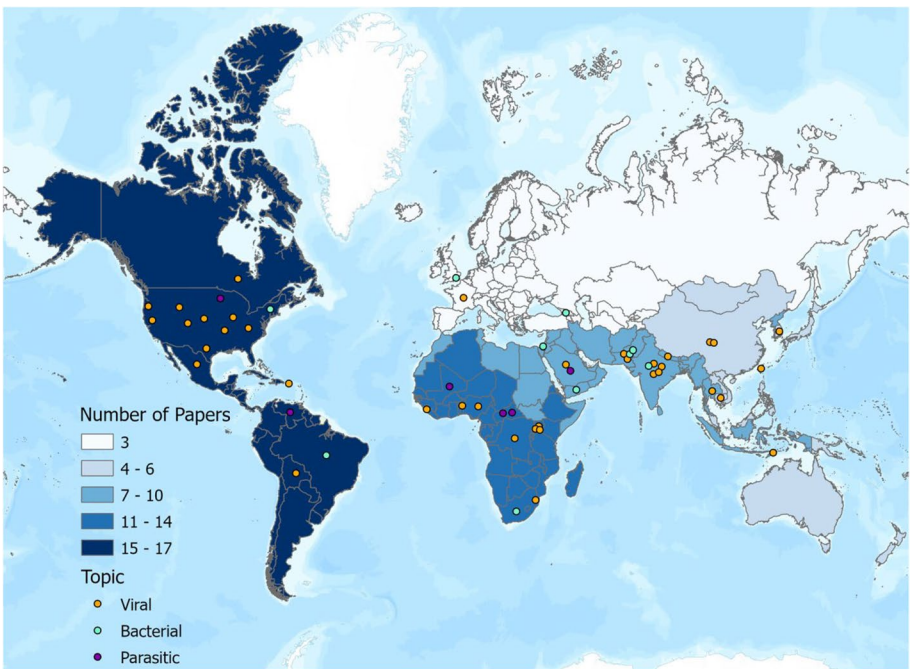

Fig. 2 Map illustrating the geographic distribution of 54 student papers according to World Health Organization regional groupings (African, 12; Americas, 17; Eastern Mediterranean, 8; European, 3; Southeast Asian, 8; Western Pacific, 4). Additional data shows students' interests in infectious disease of global health importance (20 viral (COVID, 19; HIV/AIDS, 2; Polio, 2; Ebola, 2; HPV, 1; Influenza, 1), 5 bacterial (Tuberculosis,4; Cholera, 2; Leptospirosis, 1; Lyme Disease, 2; Typhoid 1) and 3 parasite (Malaria, 2; Leishmaniasis, 1; Babesiosis, 1)

The further perspectives students took to assess the complexity of the infectious diseases of global health importance, based on weekly modules presented during the course, are summarized in Table 1. Taken together, data from Fig. 2 and Table 1, illustrate that the content of the research papers was unique to the individual students.

Assessment of ChatGPT-4o global health research papers

ChatGPT-4o Narrative – The 28 ChatGPT-4o-generated papers (Supplemental Table 1) were first evaluated based on the word-length targets provided in the prompt guidelines Results in Fig. 3 show that the content generated by ChatGPT-4o was significantly shorter (red crosses) than targets (blue dashed lines/boxes) suggested for the Introduction P1, P2, P3 sections and total paper; word-length of the Summary section met the suggested target. A cross comparison also demonstrated the content generated by ChatGPT was significantly ($p < 0.001$) shorter than student written papers.

All students indicated that the ChatGPT-4o paper included an Introduction that followed prompt guidelines. Results summarizing student responses to the 5 Likert scale items (1 = Significantly inferior; 2 = Inferior; 3 = Similar; 4 = Somewhat better; 5 = Significantly better) are presented in Fig. 4. An overall evaluation of the student surveys suggested that the majority of the students evaluated the AI-generated paper as inferior or similar to their own paper (overall satisfaction average = 2.39 (1.61–3.17); 2 = Significantly inferior; 16 = Inferior; 7 = Similar; 3 = Somewhat better; 0 = Significantly better).

Table 1 Modular perspectives for assessing global health challenge

	Perspective	Count (Gen)	Count (Spe)
1 History/Policy	History	10	1
	Policy		6
	ID in Conflict		1
	Colonialism		1
	Economical		1
2 ID	Bio ID	29	27
	GBoDisease		2
3 Nutrition	Nutrition	8	8
4 Epidemiology	Epi ID	18	18
5 Spatial Epi	Spatial Epi	15	15
6/7 Biotechnology	Biotechnology	8	7
	Bio Vaccines		1
8 Modeling	Modeling	4	4
9 Env Sci/Climate	Climate Change	15	9
	Env Health		3
	Ecology		3
10/11 Health Delivery	HC Delivery	19	4
	LTH Facilities		1
	SDOH		14
12 Anthropology	Anthropology	19	18
	Gender Inequality		1
13 Bioethics	Bioethics	17	8
	Vac Dist Ethics		9

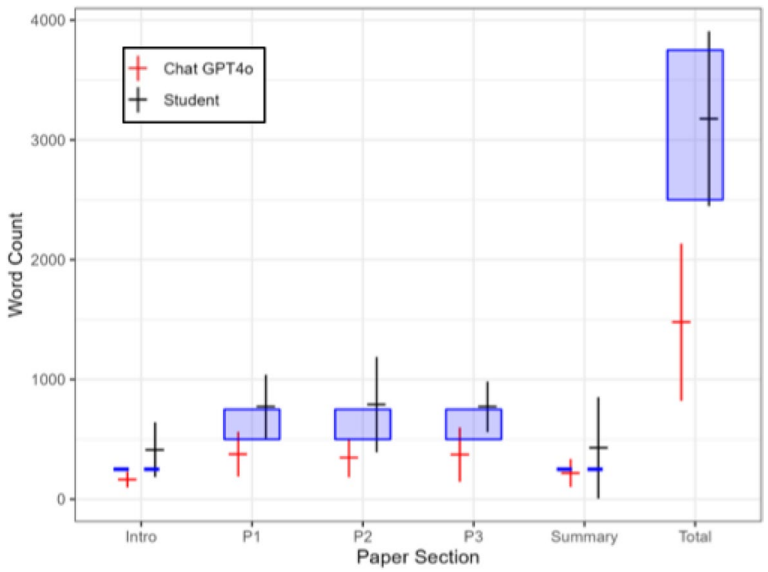

Fig. 3 Summary of ChatGPT-4o-generated Word Count and Paper Requirements. Introduction (250 words); Perspective 1 (P1, 500–750 words); Perspective 2 (P2, 500–750 words); Perspective 3 (P3, 500–750 words); Summary (Sum, 250 words); Total (2,500–3,750). Blue dashed lines and boxes identify these targets

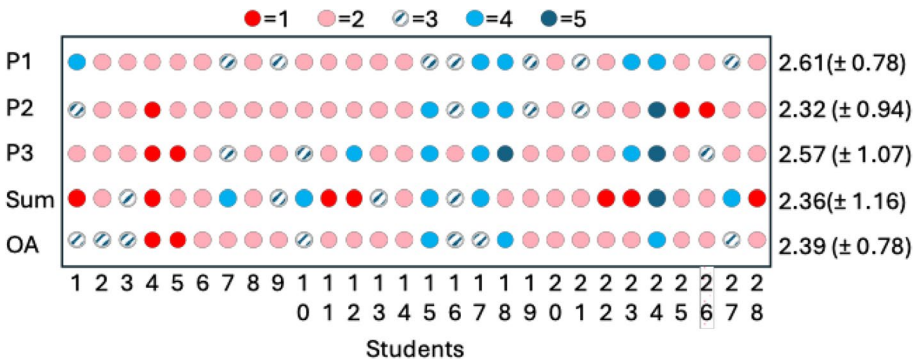

Fig. 4 Assessment of Student Survey Responses. 28 students participated in the project and completed the Google Form survey. Responses to each of the 5 Likert scale questions pertaining to five evaluations of the student research paper (Left of grid – P1 = Perspective 1; P2 = Perspective 2; P3 = Perspective 3; Sum = Summary/ Integration; OA = Overall Assessment) are shown as coded circles (see legend above the grid). The average and standard deviation of responses to each of the 5 survey items are shown at the right. There are a total of 140 individual responses: 14 responses of 1; 76 responses of 2; 26 responses of 3; 20 responses of 4; 4 responses of 5

Individual student responses showed both within and between variation. Only four of the 28 students selected the same score for all 5 survey items (students 6, 8, 14 and 20), responding with a score of 2 to each item. All other students entered different responses among the 5 items, suggesting that the students evaluated each component of the research paper independently. Consistent with the inferior assessment of the AI-generated paper, the average individual student responses for the 5 queries showed that 17 scores (60.7%) were < 2.9; 7 scores (25%) were between 3.0 to 3.9; 4 scores (60,7%) were ≥ 4.0. Additionally, for the 5 Likert scale items, we found that the upper bound of the standard deviation was less than 2.9 for 17 students (60.7%),

3–3.9 for 7 students (14.3%) and 4–4.9 for 4 students (14.3%) (Supplemental Table 2). By this individual student assessment, the majority of the students felt that the AI-generated paper was inferior to the paper that they had written; 12% indicated that the AI-generated paper was superior.

In response to the open-ended query regarding experience with AI tools, 6 students indicated first-time use in the context of this project, 9 responded a little/non-academic use, 10 responded that they had moderate experience (in the context of studying, improving grammar and spelling, organizing daily plans), 3 indicated frequent experience in writing and research. The students who had used AI tools prior to this project had used earlier versions of ChatGPT, Microsoft CoPilot and Perplexity AI. In response to their overall expectations for ChatGPT-4o in this project, the students provided both positive and negative assessments of the AI-generated paper and excerpts from a sample of these responses are provided in the following Table 2 (Minor editing to delete "…" or [modify] text were made by the instructor (e.g., from

Table 2 Student assessments of ChatGPT-4o-generated paper

1. …favorite things about public health is its relationship to medicine, math, history, and politics... It seemed that ChatGPT struggled with interweaving these topics
2. I was surprised to see how much of the essay was spent [on] defining the terms given in the prompt…I was astounded by the ability of AI to generate a paper that I would have pored over for hours to ponder each sentence and connection in a short period of time. I cannot believe that this is possible
3. ChatGPT did a great job of summarizing the existing work [conclusion], which was honestly better than mine. The introduction also did a pretty nice job of setting the scene. Yet, the ChatGPT did not really explain any claims it made throughout the body of the paper, which simply will not work in a [graduate level research paper]
4. I believe AI tools are best [for aggregating] resources… A great deal of nuance is missing from the ChatGPT version that I included in my paper because I had the historical, cultural, and global context of previous research…
5. … the synthesis, which was just regurgitated information from other sections and lacked deeper analysis
6. The modeling section was surprising, as ChatGPT did an analysis similar to mine, and created an SEIR model of COVID-19 and influenza dual-endemicity
7. I expected a higher quality output... I think part of the problem was in the quality of the inputs. I didn't include nuanced topics that I researched...
8. Perhaps due to the niche nature of the subject I was surprised that the chat bot did not provide more recent examples compared to my paper written nearly 3 years ago (original student paper Spring 2021)
9. In my own paper, I attempted to present more background information than I found in the generated paper, but the generated paper presented the biological/technical information better than I did, I think. I was both surprised and frightened by the quality of ChatGPT's output and annoyed when it presented the information better than I did
10. It is…very simple…to use and very quick. I am sure some of the grammar…in the ChatGPT version is better than my own. … very complicated to use a tool like this to completely synthesize different perspectives, specifically …as broad as Global Health
11. …reinforced my prior assumptions about ChatGPT…they can write at a high-school, 10th grade level but have repetitive sentence structure, poor transitions, and overall stilted writing style…ChatGPT made up facts about the pathophysiology of malaria
12. ChatGPT-4o output did not meet the word requirements of the paper but provided a lot of relevant information to the three perspectives. While it is helpful…, it still needs a lot of work in order to fully create a paper that meets all of the requirements
13. I think some more critical analysis of sources and synthesis of the material was needed to bring this up to par with the work of a graduate student, but it provided a useful overview of the perspectives in the prompt
14. I think I have it deeply ingrained in me that I can do it better myself, and so I'm unwilling to give up the steering wheel, so to speak
15. The ChatGPT provided similar information to my original paper. However, it was very surface level information that lacked depth and details of the topic at hand. It was a very general paper with few references that supported each point

"will not work in a research paper, let alone graduate level research" to "will not work in a [graduate level research paper]")).

These responses consistently reflected the quantitative assessment of the overall paper (Fig. 4) and provided further insights into the students' assessments of their own papers and the ChatGPT-4o product. General summaries of their comments found that the ChatGPT-4o was generated very quickly, was grammatically correct and might serve as a very useful outline or organized guide to a complete paper. However, the students found that the content was noticeably superficial, repetitive of the prompt, and nuanced details of their complex global health problem were not developed. Additionally, students commented on a number of ChatGPT-4o shortcomings with identification and integration of references. Of note, ChatGPT seldom used more than one reference to support its content. Additionally, although prompted to integrate references into the narrative (in some cases more than once), ChatGPT-4o did not complete this step in 11 of 28 papers.

Effective use of citations is an indication of acquired skills in academic writing as emphasized in principles of IL. Importantly, appropriate use of citations provides attribution to those who have previously published, content, concepts, methods, made discoveries and developed theories [38]. Written narrative on complex topics that is supported by well-selected references builds credibility and authority of the author. It shows that the writer has read the appropriate background to understand many aspects of the topic being presented, provides validation to critical facts underpinning the topic, and calls attention to the history of contributions that have led to the present. Ultimately, a well-referenced work enables readers to feel as though the authors have helped them understand the essentials linked to the topic. Understanding present-day global health requires that students read from multiple sources and use references to (1) avoid bias or over-simplification and (2) accurately represent the history, culture, geographic, biomedical and public health perspectives contributing to challenges on which they are writing. Inadequate referencing leads to the appearance of an incomplete or confused presentation of important topics. Failure to cite previous work is viewed as plagiarism [39] and stiff punishments are often meted out (failure of an assignment or expulsion from school) to those who violate this basic principle of academic integrity.

Therefore, products generated by the students and ChatGPT-4o were finally evaluated with the importance of referencing in mind. Two criteria were used to evaluate references – accuracy and impact. Accuracy, or "hit-rate" (n cited/n exist) was determined by whether the reference cited could be found to exist using the approach outlined in the method. Impact was determined by assigning the Journal impact factor (IF) to an individual reference and averaging the Journal impact factors in the paper.

For the 790 references cited by the students (28.21 references per paper), the hit-rate was 100%; average IF across the 28 student papers was 10.43. For the 729 references cited by ChatGPT-4o (26.03 references per paper), the hit-rate was 54.3% (396/729); average IF across the 28 ChatGPT-4o papers was 14.14. Factors contributing to the ChatGPT-4o 46.7 miss rate (333/729) are summarized in Table 3. Of the ChatGPT-4o references deemed to be accurate,

26.5% (105/396; 105/729 = 14.4% of total references) were then determined to be relevant to the paper narrative where they were cited. The most common reason for Failed Relevance resulted from ChatGPT-4o not integrating citations into the text (259/396).

Table 3 Citation authenticity assessment

n=396	Exact match
n=14	Near miss (e.g., moved from a preprint server to published manuscript)
n=46	Not a direct hit, most often due to updated or nonfunctioning ULR
n=111	A different manuscript at the designated volume/issue/pages
n=96	Journal site could not find manuscript
n=5	Manuscript (suggested author and title) found in a different journal or book
n=111	Manuscript duplicate in "References cited"
n=50	Google cound not find citation

Table 4 Specific examples of reference failures for accuracy and relevance

Failed Accuracy Examples

1. Cited manuscript did not appear in *Nature Medicine*, 26 pp. 1641–1645, 2020; it did appear in *Nature Medicine* 27 pp. 94–105, 2021

2. Cited manuscript did not appear in *PLoS Negl Trop Dis*. 2016;10(1); it did appear in *PLoS Negl Trop Dis*. 2011 Jan 25;5(1):e1003

3. Cited manuscript did not appear in *Int. J. of STD & AIDS*. 2005, 16(3):217–223; it did appear in *Int. J. of STD & AIDS*. 2005, 16(4):217–223

4. Cited manuscript did not appear in *Nature Microbiology*. 2014; 2,14012; it did appear in *Nature Microbiology*. 2014; 4(9):1508–1515, with an expanded author group

5. Ferguson, N. M., et al. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. *Nature*, 585(7807), 257–261

It was published as an internal document @ Imperial College of London on March 16,2020

6. *Lancet* 396(10249) did not contain pages 1138-40. These pages were found in *Lancet* 36(10258). No matching manuscript was found

Failed Relevance Examples

1. Sentence reads - In addition to vaccines, antiviral medications such as oseltamivir and zanamivir can reduce the severity and duration of influenza symptoms if administered early in the course of illness [REF]. The reference (Gostic KM, et al. (2020) Practical considerations for measuring the effective reproductive number, Rt. *PLoS Comput Biol* 16(12): e1008409) makes no mention of treatments for severity and duration of influenza symptoms by the indicated drugs

2. Sentence reads - France, leveraging its economic strength, swiftly mobilized funds for healthcare, research, and economic stimulus to mitigate the pandemic's impact [REF]. The reference (Emanuel EJ, et al.(2020) Fair Allocation of Scarce Medical Resources in the Time of Covid-19. *New England Journal of Medicine* 21;382(21):2049–2055) is focused on rationing of medical equipment and interventions in the United States. There was no mention of France in the article

3. Sentence reads - Understanding the genetic and environmental factors that influence the progression of COVID-19 and its variants is essential for developing effective public health policies [REF]. The reference (Bastard P, et al. (2020) Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* 370(6515):eabd4585. doi: 10.1126/science.abd4585) was not focused on COVID-19 variants or public health

4. Sentence reads - A study by Evans et al. [REF] reported that the timely establishment of Ebola Treatment Centers (ETCs) in Sierra Leone was associated with a reduction in case fatality rates (CFRs) from 70% to 40%, highlighting the importance of accessible and effective treatment facilities. The reference (Evans, D. K., Goldstein, M., & Popova, A. (2015). Health-care worker mortality and the legacy of the Ebola epidemic. *The Lancet Global Health*, 3(8), e439-e440) modeled how the loss of health-care workers - defined here as doctors, nurses, and midwives - to Ebola might affect maternal, infant, and under-5 mortality. There was no mention of ETCs and no specific mention of reduced CFRs from 70% to 40%

5. Sentence reads - A study by Tiffany et al. (REF) indicated that community engagement efforts led to increased compliance with public health measures and a greater willingness to report suspected cases. The reference (Tiffany, A., et al. (2017). Estimating the number of secondary Ebola cases resulting from an unsafe burial and risk factors for transmission during the West Africa Ebola epidemic. *PLoS Neglected Tropical Diseases,* 11(6), e0005491.) focused on safe dignified burial practices and not on willingness to report suspected cases

More specific examples of reference failures for accuracy and relevance are provided in Table 4.

Since November 2022, AI technology has emerged and users are testing its virtual powers by the hundreds of billions (e.g., planning workouts [40], dinner parties [41], vacations [42], seeking career advice [43], optimizing computer code [44]). There is also clear indication that a large majority of students have accessed AI tools for assistance in completing their academic assignments [6]. Therefore, this project was motivated by the wide appeal for using these tools, and curiosity regarding how they might impact and/or contribute to our understanding of complexities in global health.

This global health research paper project provided a range of opportunities to observe ChatGPT-4o performance on zero-shot prompting (GenAI tool receives a task description in a prompt that lacks labeled data for training on specific input–output mappings [35–37]) across a wide range of landscapes. Following the path from basic guideline requirements (e.g., word length, reference numbers, style, integration into the narrative) provided easily scorable outcomes and this is a major thrust of the work we have performed so far. Assessing performance of ChatGPT-4o to "hit the targets" is a surface level test. We have extended our assessment of the basic countable items by evaluating whether references and narrative make sense. In this analysis we found that 46% of references identified by ChatGPT-4o did not exist. Further assessment of narrative and references found that only 73.5% of authentic references were not relevant to the narrative with which they were integrated. Our results reflect the recent findings by Aljamaan et al., where their reference hallucination score [45] summarizes inconsistencies within references and between references and text as we have reported (Table 2). These findings indicate that ChatGPT-4o failed to meet information literacy expectations of a graduate student research paper. Emergence of newly released OpenAIo1 (aka Strawberry) has introduced reasoning into its repertoire of approaches to information assembly. Because, as described, this new tool “thinks” before it presents its information, it will be important to determine if higher rates of continuity between sentences and references are observed by users.

Limitations of this study

Settling upon the zero-shot prompting approach was intentional, so that we could observe if/how ChatGPT-4o would engage with open-ended global health outbreak scenarios, because these situations will not always have well-developed training materials to serve as guides, or because new episodes of a somewhat familiar outbreak may include important nuances that differ from earlier outbreaks. The ChatGPT-4o-generated responses we observed were consistent with more limited findings from other researchers; that generative AI responses are often very generalized, stressed preparedness and were often repetitive of content in the prompt. As stated earlier, there are relatively few papers at the intersection of ChatGPT and global health, and so it is important to know if it is possible to improve upon the outcomes we observed. At this point we do not know if content generated by zero-shot prompting will improve as GenAI tools are upgraded. Based on experience while designing the current study, we do have some insight into

whether the same outcomes would be observed if the study were repeated by the same cohort. For example, when we asked ChatGPT-4o to revise content to reach word count requirements, we observe rewriting but very similar content in the GenAI product. We also observe the same flaws in the references selected by ChatGPT-4o that have been elaborated in Tables 2 and 4. Future studies will determine if new versions of GenAI tools will produce content that more closely reaches graduate student standards of IL.

Zero-shot prompting is different from approaches where a GenAI tool is provided with substantial content as raw material (e.g., Few-shot, Chain-of-thought [35–37]). These more highly guided strategies are also being used by biomedical researchers to test LLMs abilities to assist with writing basic research manuscripts and authoritative review articles [45–48]. Would our results have been different if we had employed these alternative strategies? As these studies have used very similar GenAI tools (ChatGPT-3.5 or ChatGPT-4o) we are not surprised to see many of the same shortcomings observed in our study (e.g., reference and content hallucinations [45]). We also wondered whether using one or more published manuscripts as AI query guides would overly influence the GenAI product by the circumstances of past events? Again, future studies using a wide assortment of prompting strategies [35–37] should be performed by subject-matter experts to investigate how information relevant to global health research is generated and assess its quality.

Finally, beyond the results of this technical comparison between student and ChatGPT-4o-generated research papers, there is considerable important work to perform at the intersection of GenAI and global public health to fill a substantial knowledge gap. As demonstrated by the student papers, numerous ongoing outbreaks of globally important infectious diseases provide content for testing and training LLMs beyond the basic recognition of words and phrases in written content describing these outbreaks. Developing new tools to predict how various scenarios might amplify transmission of infectious diseases beyond local, regional, and national public health agency capacity to control and constrain disease outbreaks will be important to assess. Testing how sensitive and specific GenAI would be in detecting conditions for new infectious disease outbreaks would be a major global public health development. For insights on this, one of our students who wrote their paper on a complex respiratory virus model provided the following favorable assessment of was ChatGPT-4o. "ChatGPT created an SEIR (Susceptible-Exposed-Immune-Recovered) model of COVID-19 and influenza dual-endemicity with varying scenarios of intervention. This model was calibrated on public CDC 2019–2023 data, which was not available [at the time the student paper was written in 2022]." Other published studies provided varying assessments. Three studies conducted "natural conversations (NC)" with ChatGPT finding that the GenAI tool had capacity to refine and debug code and develop a model that could fit 10 days of prevalence data to estimate a basic reproductive number and final epidemic size [49]. However, two additional studies using similar NC strategies found either that ChatGPT stated clearly that it was not capable of predicting disease-specific trends [47] or was positioned, at best, to play a supportive role to human expertise for early prediction, prevention, and management of future pandemics [50]. Additional studies have found that ChatGPT demonstrates

significant capacity for compiling free text data to improve the accuracy of symptoms, monitoring social media trends, as well as detecting and dismissing conspiracy beliefs as non-credible and lacking scientific evidence [51, 52]. With the observations form the present study and limited numbers of published studies at the GenAI / global public health intersection / knowledge gap, substantial subject-matter expert research will be required to enable AI-generated content to make positive contributions to authoritative sources (CDCs or WHO) on emerging public health threats.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-024-00408-7.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Supplementary Material 4.

Acknowledgements
CWRU faculty who have contributed to Fundamentals of Global Health (INTH 301/401) from 2020 to 2024 include, Charles King, Daniel Tisch, Andrew Curtis, James Swain, Ernest R. Chan, Jürgen Bosch, Karen Abbott, Maria Diaz-Insua, Karen Mulloy, Kurt Rhoads, Janet McGrath, David Miller, Joachim Voss, James Leslie and Nicole Deming. Special thanks to Ronald Blanton (Chair, Department of Tropical Medicine, Tulane School of Public Health and Tropical Medicine), the founding director of INTH 301/401. We are grateful to guidance and constructive criticisms in developing this manuscript from Christine Arcari (Senior Associate Dean for Academic Affairs, Tulane School of Public Health and Tropical Medicine), Katherine Elkins (Director, Kenyon College Digital Humanities Collaboration), Jon Chun (Kenyon College, Visiting Instructor of Humanities) and Rajiv Thandla (Microsoft).

Authors' contributions
SRT, GQA and PAZ optimized the study design. SRT, GQA, AdM, AS, DLG, CH, EH, YI, RM, AnM, MAP, TMR, ORT, ARH, CAB, TG, SL, SM, SGY, JEG, BMG, AH, YPL, JM, JPO, ACW, CY and JZ wrote original student research papers, prompted ChatGPT4o to write comparison papers, compared their original vs their ChatGPT4o papers via a survey developed in Google Forms. SRT and GQA performed data analysis. SRT, GQA and PAZ produced figures, tables and supplemental materials. PAZ designed the study and wrote the manuscript. All authors read and approved the final manuscript.

Funding
CWRU Master of Public Health Program Student Activity Fund.

Data availability
Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate
Based on discussions with faculty colleagues who presented module content and the Senior Associate Dean of Graduate Studies, this study did not qualify as human subjects research and therefore did not require Institutional Review Board approval. All students agreed with the project protocol and consented to participate.

Consent for publication
All authors have read and agreed to the published version of the manuscript.

Competing interests
The authors declare no competing interests.

Received: 19 September 2024 Accepted: 19 November 2024
Published online: 18 December 2024

References

1. Turing AM: Computing Machinery and Intelligence. Mind. 1950;LIX(236):433–460.
2. Asimov I: I, Robot. New York, NY: Gnome Press; 1950.
3. Hayes PJ, Morgenstern L. On John McCarthy's 80th Birthday, In Honor of His Contributions. AI Mag. 2007;28(4):93–102.
4. Barbaro MH: The Daily. In: Did Artificial Intelligence Just Get Too Smart? Edited by Barbaro M: The New York Times Company; 2022.

5. Cardillo A: 20 Most Popular AI Tools Ranked (August 2024). In: Exploding Topics. Edited by Dean B, vol. 2024. San Francisco, California; 2024.
6. Chun J, Elkins K. The crisis of artificial Intelligence: A new digital humanities curriculum for human-centered AI. International Journal of Humanities and Arts Computing. 2023;17(2):147–67.
7. Foster A, Elkins K, Chun J: How Well Can GPT-4 Really Write a College Essay? Combining Text Prompt Engineering And Empirical Metrics. In. https://digital.kenyon.edu/dh_iphs_ss/24: Kenyon College; 2023.
8. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Karf AK, Baabdullah AM, Koohang A, Raghavan V, Ahuja M, et al. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage. 2023;71(C):102642.
9. Kupferschmidt K: Outbreak of virus from China declared global emergency. Science 2020.
10. Zimmerman PA, King CL, Ghannoum M, Bonomo RA, Procop GW. Molecular Diagnosis of SARS-CoV-2: Assessing and Interpreting Nucleic Acid and Antigen Tests. Pathog Immun. 2021;6(1):135–56.
11. Scales D, Gorman J, Jamieson KH. The Covid-19 Infodemic - Applying the Epidemiologic Model to Counter Misinformation. N Engl J Med. 2021;385(8):678–81.
12. SteelFisher GK, Blendon RJ, Caporello H. An Uncertain Public - Encouraging Acceptance of Covid-19 Vaccines. N Engl J Med. 2021;384(16):1483–7.
13. Baker M, Brabaro M: The Daily. In: The Public Health Officials Under Siege. Edited by Barbaro M: The New York Times Company; 2021.
14. Burkle FM Jr. Global Health Security Demands a Strong International Health Regulations Treaty and Leadership From a Highly Resourced World Health Organization. Disaster Med Public Health Prep. 2015;9(5):568–80.
15. Cohen J. Africa steps up battle against mpox outbreaks. Science. 2024;384(6694):373–4.
16. Fidler DP, Gostin LO. The new International Health Regulations: an historic development for international law and public health. J Law Med Ethics. 2006;34(1):85–94, 84.
17. Gostin LO, Meier BM, Abdool Karim S, Bueno de Mesquita J, Burci GL, Chirwa D, Finch A, Friedman EA, Habibi R, Halabi S, et al. The World Health Organization was born as a normative agency: seventy-five years of global health law under WHO governance. PLOS Glob Public Health. 2024;4(4):e0002928.
18. Lenharo M. Hope for global pandemic treaty rises - despite missed deadline. Nature. 2024;630(8016):282.
19. Adams LV, Wagner CM, Nutt CT, Binagwaho A. The future of global health education: training for equity in global health. BMC Med Educ. 2016;16(1):296.
20. Battat R, Seidman G, Chadi N, Chanda MY, Nehme J, Hulme J, Li A, Faridi N, Brewer TF. Global health competencies and approaches in medical education: a literature review. BMC Med Educ. 2010;10:94.
21. Calhoun JG, McElligott JE, Weist EM, Raczynski JM. Core competencies for doctoral education in public health. Am J Public Health. 2012;102(1):22–9.
22. Calhoun JG, Spencer HC, Buekens P. Competencies for global heath graduate education. Infect Dis Clin North Am. 2011;25(3):575–92 viii.
23. Clark M, Raffray M, Hendricks K, Gagnon AJ. Global and public health core competencies for nursing education: A systematic review of essential competencies. Nurse Educ Today. 2016;40:173–80.
24. Koplan JP, Bond TC, Merson MH, Reddy KS, Rodriguez MH, Sewankambo NK, Wasserheit JN. Consortium of Universities for Global Health Executive B: Towards a common definition of global health. Lancet. 2009;373(9679):1993–5.
25. Peluso MJ, Encandela J, Hafler JP, Margolis CZ. Guiding principles for the development of global health education curricula in undergraduate medical education. Med Teach. 2012;34(8):653–8.
26. Zurkowski PG: The information service enviroment relationships and priorities. Related Paper No. 5. Washington D.C.; 1974.
27. Leachman C: Introduction to Standards. In: Teaching and Collecting Technical Standards: A Handbook for Librarians and Educators (2023) Purdue Information Literacy Handbooks 5. edn. Edited by Leachman C, Rowley EM, Phillips M, Solomon D. Purdue University: Purdue e-Pubs; 2023: 3–13.
28. Stern C, Kaur T. Developing theory-based, practical information literacy training for adults. International Information & Library Review. 2010;42(2):69–74.
29. Association AL: Presidential Committee on Information Literacy. Final Report In.; 1989.
30. Technology AASTFoILfSa: Information Literacy Standards for Science and Engineering/Technology. In.; 2006.
31. UNESCO: Background Document of the Expert Meeting: Towards Media and Information Literacy Indicators. Prepared by Susan Moeller, Ammu Joseph, Jesús Lau, Toni Carbo. In. Paris: UNESCO; 2010: 53.
32. Li Y, Chen Y, Wang Q. Evolution and diffusion of information literacy topics. Scientometrics. 2021;126(5):4195–224.
33. UNESCO: The Prague Declaration: Towards an information literate society. In.: Literacy Meeting of Experts,; 2003: 1.
34. UNESCO: Beacons of the Information Society: The Alexandria Proclamation on Information Literacy and Lifelong Learning. In.: International Federation of Library Associations and Institutions (IFLA); 2005: 2.
35. Brown TB, B. M, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A *et al*: Language Models are Few-Shot Learners. *arXiv* 2020.
36. Sahoo P, Singh AK, Saha S, Jain V, Mondal S, Chadha A: A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv* 2024.
37. Schulhoff S, Ilie M, Balepur N, Kahadze K, Liu A, Chenglei S, Li Y, Gupta A, Han HJ, Schulhoff S *et al*: The Prompt Report: A Systematic Survey of Prompting Techniques. *arXiv* 2024.
38. Bahadoran Z, Mirmiran P, Kashfi K, Ghasemi A. The Principles of Biomedical Scientific Writing: Citation. Int J Endocrinol Metab. 2020;18(2): e102622.
39. Kumar PM, Priya NS, Musalaiah S, Nagasree M. Knowing and avoiding plagiarism during scientific writing. Ann Med Health Sci Res. 2014;4(Suppl 3):S193–198.
40. Blackmore E: Why you shouldn't rely on ChatGPT for exercise suggestions just yet. In: The Washington Post. Washington, DC: William Lewis; 2024.
41. Yohannes A: Welcome to my AI-hosted dinner party. In. Edited by Chambers L, vol. 2024. Mozilla; 2024.

42. Williams R: How to use AI to plan your next vacation. In: Artificial Intelligence. Edited by Heaven WD, July 8, 2024 edn. Massachusetts Institute of Technology; 2024: https://www.technologyreview.com/2024/2007/2008/1094733/how-to-use-ai-to-plan-your-next-vacation/.
43. Abril D: AI career coaches are here. Should you trust them? In: The Washington Post. Washington, DC: William Lewis; 2024.
44. Wrótniak K: 10 Best AI Coding Assistant Tools in 2024– Guide for Developers. In., vol. 2024. Droids on Roids; 2024.
45. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, Mesallam TA, Farahat M, Malki KH. Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study. JMIR Med Inform. 2024;12: e54345.
46. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health. 2023;13:01003.
47. Cheng K, He Y, Li C, Xie R, Lu Y, Gu S, Wu H. Talk with ChatGPT About the Outbreak of Mpox in 2022: Reflections and Suggestions from AI Dimensions. Ann Biomed Eng. 2023;51(5):870–4.
48. Wang ZP, Bhandary P, Wang Y, Moore JH. Using GPT-4 to write a scientific review article: a pilot evaluation study. BioData Min. 2024;17(1):16.
49. Kwok KO, Huynh T, Wei WI, Wong SYS, Riley S, Tang A. Utilizing large language models in infectious disease transmission modelling for public health preparedness. Comput Struct Biotechnol J. 2024;23:3254–7.
50. Jana PK, Majumdar A, Dutta S. Predicting Future Pandemics and Formulating Prevention Strategies: The Role of ChatGPT. Cureus. 2023;15(9): e44825.
51. Sallam M, Salim NA, Al-Tammemi AB, Barakat M, Fayyad D, Hallit S, Harapan H, Hallit R, Mahafzah A. ChatGPT Output Regarding Compulsory Vaccination and COVID-19 Vaccine Conspiracy: A Descriptive Study at the Outset of a Paradigm Shift in Online Search for Information. Cureus. 2023;15(2): e35029.
52. Wei WI, Leung CLK, Tang A, McNeil EB, Wong SYS, Kwok KO. Extracting symptoms from free-text responses using ChatGPT among COVID-19 cases in Hong Kong. Clin Microbiol Infect. 2024;30(1):142 e141–142 e143.

Publisher's Note