

RESEARCH ARTICLE

Open Access



Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences

Andrew Williams^{1*}

*Correspondence:
andrew.williams@ucl.ac.uk

¹ Department of Education,
Division of Medicine, University
College London, London WC1E
6BT, UK

Abstract

The value of generative AI tools in higher education has received considerable attention. Although there are many proponents of its value as a learning tool, many are concerned with the issues regarding academic integrity and its use by students to compose written assessments. This study evaluates and compares the output of three commonly used generative AI tools, ChatGPT, Bing and Bard. Each AI tool was prompted with an essay question from undergraduate (UG) level 4 (year 1), level 5 (year 2), level 6 (year 3) and postgraduate (PG) level 7 biomedical sciences courses. Anonymised AI generated output was then evaluated by four independent markers, according to specified marking criteria and matched to the Frameworks for Higher Education Qualifications (FHEQ) of UK level descriptors. Percentage scores and ordinal grades were given for each marking criteria across AI generated papers, inter-rater reliability was calculated using Kendall's coefficient of concordance and generative AI performance ranked. Across all UG and PG levels, ChatGPT performed better than Bing or Bard in areas of scientific accuracy, scientific detail and context. All AI tools performed consistently well at PG level compared to UG level, although only ChatGPT consistently met levels of high attainment at all UG levels. ChatGPT and Bing did not provide adequate references, while Bing falsified references. In conclusion, generative AI tools are useful for providing scientific information consistent with the academic standards required of students in written assignments. These findings have broad implications for the design, implementation and grading of written assessments in higher education.

Keywords: Assessment, Artificial intelligence, Higher education, Academic writing, ChatGPT, Essay, Biomedical science, Medicine

Introduction

Emerging digital technologies have a long history in university education practice (Doroudi, 2022). These include virtual learning environments (VLEs), augmented reality, intelligent tutoring systems, automatic marking and grading systems, and now generative artificial intelligence (AI) (Woolf 2010). The adoption of technology into education (EdTech) and the recent incorporation of AI, known as artificial intelligence in education (AIEd), has been received with scepticism and optimism in equal

measure (Rudolph et al., 2023). Generative AI tools use a type of machine learning algorithm, known as a large language model (LLM), that is capable of producing novel output in response to a text prompt. Familiar examples include ChatGPT (developed by OpenAI), Google Bard and Microsoft 365 Copilot (which leverages the Open AI GPT-4 LLM). A Generative Pretrained Transformer (GPT) is a sophisticated LLM that applies deep learning technology. They are 'generative' due to their ability produce novel text in response to a user-provided input and they are "pretrained" on terabytes of data from the corpus of existing internet information. They are called 'transformers' because they employ a neural network to transform input text to produce output text that closely resembles human language. Many consider generative AI as a transformative technology that will revolutionise industry, academia and society (Larsen, 2023).

Many in the higher education (HE) sector are concerned that students will use generative AI to produce written assignments and therefore as a tool for plagiarism (Perkins, 2023). The ease of access to generative AI applications and the simplicity of generating written text, may have far-reaching consequences for how students approach their education and their written assessments (Nazari et al., 2021). The rapid development and progressive improvements in LLMs mean they are becoming more effective at producing human-like outputs (Kasneci et al., 2023; Team, 2022). The implications on both academic integrity and the development of student's academic skills are therefore considerable (Cassidy, 2023). However, there is a lack of research on evaluating the effectiveness of LLMs across a broad range of assessment types and across different disciplines within higher education.

The importance of the student essay, as a form of assessment, is still relevant to the biomedical sciences, while academic writing is considered a key skill for trainee scientists and many other disciplines (Behzadi & Gajdács, 2021). Although the adoption of AI may address some of the challenges faced by teachers in higher education, for example the acceleration of marking large numbers of student scripts, educators must still ensure academic integrity is maintained (Cotton et al. 2023) as well as assessing the ability or performance of the student. In the biomedical sciences, an in-depth knowledge and critical evaluation of the scientific field is often a requisite of written assignments (Puig et al., 2019) and an important skill to acquire for professional development. The performance of students in key marking criteria, such as mechanistic detail and scientific accuracy, are also valued by teachers to allow an informed academic judgement of student's submissions.

Previous analysis of generative AI on university student assessment has been mixed in terms of comparative performance to student's own work. ChatGPT performed very well on a typical MBA course, achieving high grades on multiple questions (Terwiesch, 2023). ChatGPT also performed well on the National Board of Medical Examiners (NBME) exam, equivalent to a third-year medical student and demonstrated reasoned answers to complex medical questions (Gilson et al., 2022). A larger study of 32 university-level courses from eight disciplines revealed that the performance of ChatGPT was comparable to university students, although it performed better at factual recall questions (Ibrahim et al., 2023). However, the text output in response to questions was limited to two paragraphs and therefore not equivalent to a longer, essay-style answer.

Doubts remain as to the depth and accuracy of ChatGPT responses (and other generative AI tools) to more complex and mechanistic questions related to biomedical subject material.

This study aims to evaluate how different generative AI tools perform in writing essays in undergraduate and postgraduate biomedical sciences courses. In response to example essay questions, AI generated answers were anonymously evaluated with reference to the level descriptions from the Frameworks for Higher Education Qualifications in the UK (QAA, 2014), providing evaluations of generative AI output across four levels of undergraduate and postgraduate education. The performance of three commonly used AI tools were investigated, ChatGPT 3.5, Google Bard and Microsoft Bing. Scientific accuracy, mechanistic detail, deviation, context and coherence were evaluated by four independent markers, inter-rater reliability calculated and written instructor comments provided on each output. The findings provide insights into the effectiveness of LLMs in academic writing assignments, with broad implications for the potential use of these tools in student assessments in the biomedical sciences.

Methods

Written assessment questions

Essay questions from four courses at University College London (UK), within the biomedical sciences, were used to evaluate the performance of generative AI. An essay question from each of the undergraduate and taught postgraduate years of study was assessed, equivalent to undergraduate level 4 (first year), level 5 (second year), level 6 (third year) and postgraduate taught level 7 (MSc), according to the level descriptors provided by the Frameworks for Higher Education Qualifications (FHEQ) in the UK (QAA, 2014). Escalating levels of student attainment are expected as the level descriptions increase, providing an opportunity to evaluate the performance of generative AI at each level. The essay subject material was derived from the undergraduate level 4 course 'Kidneys, Hormones and Fluid Balance'; the level 5 course 'Molecular Basis of Disease'; the level 6 course 'Bioinformatics'; and the level 7 course 'Principles of Immunology', and were representative of the types of essay question that would be used in these modules.

The following essay questions were used to evaluate the responses generated by AI:

Level 4: Describe how the glomerular filtration rate is regulated in the kidney.

Level 5: Describe the pathomechanisms of chronic obstructive pulmonary disease.

Level 6: Describe how bioinformatics is used in biomedical research. Please provide examples of how bioinformatics has helped in our understanding of human diseases.

Level 7: Describe the major histocompatibility complex II antigen processing pathway. Please provide details of the key molecular interactions at each step.

Generative AI tools

Three generative AI tools were evaluated, selected on the basis of student access and their interface with the two most frequently used internet search engines. The free version of ChatGPT 3.5 (built on a variant of the OpenAI transformer model), Google Bard (built on the Google Language Model for Dialogue Applications—LaMDA) and Microsoft Bing (leveraging the OpenAI GPT-4 transformer model) were evaluated (Aydin & Karaarslan, 2023). Questions were submitted to each LLM application between 11th

August 2023 and 9th September 2023. Each question was entered into ChatGPT, Bard or Bing without modifications as the following prompt: '{Question}. Please write your response as a 1000 word university level essay.' No additional prompts or subsequent text requests were made in relation to the generation of the essay answer. For the level 7 essay question only, a follow-up prompt: 'Could you list 4 references to justify your answer?' was applied. This was asked only for the level 7 question in order to determine the overall ability of each AI model to generate references. No alterations in the basic settings of any of the generative AI applications were made in order to mimic the intrinsic functionality of each model.

FHEQ level descriptors

The FHEQ level descriptors (McGhee, 2003) are widely recognised across the higher education sector as the benchmark standard for student attainment at each level of study and therefore provide a useful scaffold to evaluate the performance of generative AI. Particular reference to the knowledge and understanding component in relation to Annex D: Outcome classification descriptions for FHEQ Level 6 were made, as this criterion was most relevant to assessing responses to the essay questions (QAA, 2014). In brief the following statements were used as benchmark statements to evaluate generative AI output in response to essay questions.

Level 4: "knowledge of the underlying concepts and principles associated with their area(s) of study, and an ability to evaluate and interpret these within the context of that area of study".

Level 5: "knowledge and critical understanding of the well-established principles of their area(s) of study, and of the way in which those principles have developed".

Level 6: "a systematic understanding of key aspects of their field of study, including acquisition of coherent and detailed knowledge, at least some of which is at, or informed by, the forefront of defined aspects of a discipline".

Level 7: "a systematic understanding of knowledge, and a critical awareness of current problems and/or new insights, much of which is at, or informed by, the forefront of their academic discipline, field of study or area of professional practice".

Generative AI performance evaluation

Responses generated by ChatGPT, Google Bard and Microsoft Bing were assessed according to the FHEQ framework and marking criteria described (Table 1). The following terms were used to evaluate text generated in response to each essay question: 'Scientific Accuracy'; 'Mechanistic Detail'; 'Deviation'; 'Context'; and 'Coherence', each marking criteria therefore mapping to the FHEQ descriptors. Markers were provided with the marking rubric and a percentage value and an ordinal grade of low, medium or high was attributed to each marking criteria, for each of the generative AI models evaluated and for each essay question across all four level descriptors. The higher the level descriptor, the greater the expectation for each marking criteria. Illustrative examples from generative AI responses were selected, to highlight areas of good and poor academic writing related to the prescribed marking criteria and FHEQ descriptors. In addition, assessor feedback comments on relevant aspects of the responses were provided.

Table 1 Evaluation criteria descriptions

Level	Scientific accuracy
Low (< 50%)	Inadequate or inaccurate descriptions of scientific terminology or topics. Sufficient only to address key terminology or subject area
Medium (50–70%)	Demonstrates a sound breadth and depth of scientific terminology or topics. Some evidence of critical insight and subject-specific principles and concepts
High (> 70%)	Exceptional knowledge and understanding of subject material, principles and concepts, with evidence of clear critical insight beyond material that was taught
Level	Mechanistic detail
Low (< 50%)	Inadequate or inaccurate descriptions of mechanisms or scientific processes. No clear critical insight
Medium (50–70%)	Demonstrates a clear understanding of key mechanism or scientific processes, with evidence of some critical insight
High (> 70%)	Exceptional understanding of mechanisms or scientific process, with clear critical insight and evidence of research beyond the taught material
Level	Deviation
Low (< 50%)	A high degree of deviation from the subject material
Medium (50–70%)	Only minor instances of deviation from the subject material
High (> 70%)	No deviation from subject material, with evidence of clear subject-specific focus
Level	Context
Low (< 50%)	Little or no evidence of contextualising scientific concepts or mechanisms
Medium (50–70%)	Some evidence of contextualising scientific concepts or mechanisms
High (> 70%)	Strong evidence of contextualising scientific concepts or mechanisms, with evidence of critical evaluation across a range of scientific principles
Level	Coherence
Low (< 50%)	Lacks a coherent narrative and poor logical progression
Medium (50–70%)	A clear and coherent narrative and logical progression
High (> 70%)	Exceptionally clear and coherent narrative and logical progression, with evidence of directing strong scientific arguments

Each essay question was anonymised prior to marker evaluation, to avoid marker bias. Four independent markers were used to evaluate each generated output across all AI tools and all FHEQ levels. Marking criteria and model answers (not shown, supplementary material) were provided to all independent markers prior to evaluation. The inter-rater reliability (or inter-rater agreement) was calculated using Kendall's coefficient of concordance with SPSS Statistics (IBM) software. The Kendall's W score and p -value (sig.) were calculated for each AI generated output for each marking criteria, and for the cumulative criteria scores for each essay paper. The overall grade (total score) was calculated as the average percentage score of each marking criteria score.

Results

Firstly, an evaluation of the ability of generative AI to answer university-level essay questions, and an evaluation of the performance of different AI tools, was undertaken. The generative AI models evaluated were ChatGPT 3.5, Google Bard and Microsoft Bing. Essays were chosen from four courses within the biomedical sciences. Secondly, to evaluate the performance of generative AI against increasing levels of expected student attainment, essays were chosen from each level of UK undergraduate (UG) and post-graduate study. This equates to essays from level 4, level 5, level 6 and level 7 courses,

which are equivalent to the three years of undergraduate study (levels 4–6) and one of postgraduate (level 7). The performance of generative AI was assessed by mapping responses to the UK FHEQ descriptors (QAA, 2014), which set out intended attainment levels for each year of university study. Scripts were anonymised prior to marking.

Specified marking criteria (Table 1) were used to assess the performance of generative AI, based on scientific accuracy, mechanistic detail, deviation and coherence of the essays produced. Scientific accuracy reflected the appropriateness of responses in terms of factual correctness and correct use of scientific terminology, while mechanistic detail assessed the depth to which answers explored processes and systems (including providing relevant examples) expected in essay answers from students at each FHEQ level. Deviation assessed whether responses were explicit and did not deviate from the original question, while cohesion assessed whether responses provided an intelligible narrative that would be expected of an answer written by a human. Each generative AI model was assigned an indicator of how well it performed according to the marking criteria, as well as a percentage for each criteria. The total essay score was calculated as an average of the marking criteria. Inter-rater reliability was calculated using Kendall's coefficient of concordance. This was applied to essay answers at all descriptor levels. For each AI response, instructor feedback in the form of comments was also provided, based on the specific marking criteria. Illustrative examples of AI generated output (in italics) followed by instructor comments are provided in the context of each essay answer (ChatGPT, Google Bard, Microsoft Bing, personal communication, 2023).

Level 4

For each generative AI model, the following prompt was used: Describe how the glomerular filtration rate is regulated in the kidney. Please write your response as a 1000 word university level essay. Essays were evaluated in relation to the FHEQ level 4 descriptor and the prescribed marking criteria (Table 2). Four independent markers evaluated anonymised scripts. The inter-rater reliability and agreement between assessors were measured by Kendall's coefficient of concordance for each paper, for each criteria and across the cumulative marks for each paper.

ChatGPT

For level 4, ChatGPT generated an answer with a high degree of scientific accuracy with only minor factual errors. The level of mechanistic detail was appropriate although the

Table 2 Output summary in response to level 4 essay question

LLM	Scientific accuracy	Mechanistic detail	Deviation	Context	Coherence	FHEQ level 4 total
ChatGPT	70.25% High	69.5% Medium	70.5% High	69.5% Medium	73.75% High	70.7% High
Bard	67.75% Medium	66.5% Medium	69.25% Medium	62.5% Medium	63.25% Medium	65.85% Medium
Bing	54.75% Medium	53.5% Medium	67% Medium	52.75% Medium	50.75% Medium	55.75% Medium
Kendall's W	0.817	1.0	0.481	1.0	1.0	0.708
Sig	0.038	0.018	0.146	0.018	0.18	< 0.001

response fell below that expected of a first-class answer. There was little deviation away from the essay question or subject matter and the essay was written in a clear, precise and coherent manner. More context could have been provided in regard to the physiological importance of the subject. Of additional note was the use of American English, rather than UK English, although the summary text was considered to be of a high standard.

Examples:

ChatGPT response (accuracy): "This leads to a decreased sodium chloride concentration at the macula densa."

Instructor comments: "This leads to an increased sodium chloride concentration detected by the macula densa."

ChatGPT response (detail): "When systemic blood pressure increases, the afferent arterioles constrict, limiting the entry of blood into the glomerulus."

Instructor comment: "I would expect more detail here on the mechanism of myogenic autoregulation. For example, stretch receptors in myocytes lining the afferent arterioles, opening of nonspecific cation channels, depolarisation, calcium release, cell vessel contraction."

ChatGPT response (detail): "This mechanism provides negative feedback to prevent excessive filtration and maintain GFR within an appropriate range."

Instructors comment: "Would expect (more detail) here that autoregulation only works within a specific range of GFR and can be overridden by various hormones, e.g., angiotensin II."

Bard

Although the scientific accuracy was appropriate for a level 4 essay, the extent of scientific detail and mechanistic insight was lower in comparison to ChatGPT. Some descriptions of scientific mechanisms were vague or the information provided was repetitive. There was a lack of context with regard to the fundamental physiological importance of the subject material.

Examples:

Bard responses (detail): "ADH increases the GFR by causing the afferent arterioles to constrict. This constriction reduces the blood flow to the glomerulus, which in turn increases the GFR."

"Aldosterone increases the GFR by causing the efferent arterioles to constrict. This constriction increases the blood pressure in the glomerulus, which in turn increases the GFR."

"Atrial natriuretic peptide (ANP): ANP decreases the GFR by causing the afferent arterioles to dilate. This dilation increases the blood flow to the glomerulus, which in turn decreases the GFR."

Instructor comments: "Repetitive and lacks mechanistic detail."

Bard response (context): "It is important to monitor the GFR regularly, especially in people with risk factors for kidney disease."

Instructor comments: "This sentence lacks context, for example, in what clinical setting should GFR be monitored?"

Bing

The output generated by Bing in response to this question contained a lower level of scientific accuracy and mechanistic detail compared to ChatGPT and Bard. Furthermore, the response lacked context and there was little cohesion or logical narration to the output and the response lacked a clear narrative. Of the three LLMs evaluated, Bing generated an answer that received the lowest overall grade.

Examples:

Bing response (detail): "The distal convoluted tubule (DCT), which is located after the loop of Henle, passes close to the glomerulus and forms a structure called the juxtaglomerular apparatus (JGA)."

Instructor comments: "Good level of detail provided."

Bing response (detail and context): "A low GFR indicates impaired kidney function, which can be caused by various factors, such as diabetes, hypertension, infection, inflammation, or kidney damage."

Instructor comments: "Does not indicate the range of abnormal or low GFR. Good examples of disease that affect GFR, although these lack specific context and mechanism."

Bing response (context): "Renin converts angiotensinogen, a plasma protein, to angiotensin I, which is then converted to angiotensin II by angiotensin-converting enzyme (ACE) in the lungs."

Instructor comments: "Reasonable level of detail. However, ACE expression is also high in the kidney proximal tubules and contributes to angiotensin conversion."

Bing response (context): "ANP also inhibits the secretion of renin, aldosterone, and ADH and increases the excretion of sodium and water. This decreases the blood volume and blood pressure and increases the urine volume and urine concentration."

Instructor comments: "No connection made between cause and effect. The two sentences are disjointed."

Level 5

For each generative AI model, the following prompt was used: Describe the pathomechanisms of chronic obstructive pulmonary disease. Please write your response as a 1000 word university level essay. Essays were evaluated in relation to the FHEQ level 5 descriptor and the prescribed marking criteria (Table 3). Four independent markers evaluated anonymised scripts. The inter-rater reliability and agreement between

Table 3 Output summary in response to level 5 essay question

LLM	Scientific accuracy	Mechanistic detail	Deviation	Context	Coherence	FHEQ level 5 total
ChatGPT	67.65% Medium	66.5% Medium	75% High	70.75% High	72% High	70.4% High
Bard	63% Medium	61.75% Medium	68.75% Medium	62% Medium	66.25% Medium	64.35% Medium
Bing	49.5% Low	46.25% Low	62.25% Medium	51.25% Medium	53.74% Medium	52.8% Medium
Kendall's W Sig.	0.95 0.022	0.95 0.022	0.477 0.148	0.929 0.024	0.95 0.022	0.822 < 0.001

assessors were measured by Kendall's coefficient of concordance for each paper, for each criteria and across the cumulative marks for each paper.

ChatGPT

ChatGPT produced a well-structured essay, which contained a rational narrative with elements of contextualisation regarding disease and mechanistic context. ChatGPT outperformed Bard and Bing on this task, generating a reasonably high level of detail, with no major scientific inaccuracies. However, the essay lacked specific or appropriate examples and lacked a broader scientific insight that would be expected from students at level 5. In addition, although ChatGPT generated a convincing introduction and clear narrative, some of the language used was profuse and the output contained several generalisations. However, ChatGPT generated a convincing piece of academic writing that scored highly on context and coherence.

Examples:

ChatGPT response (context): "This essay delves into the multifaceted nature of COPD, exploring the interconnected roles of chronic inflammation, oxidative stress, protease-antiprotease imbalance, structural alterations, impaired mucociliary clearance, genetic susceptibility, and epigenetic modifications in the development and progression of this complex respiratory disorder."

Instructor comments: "Good introduction to essay topic and provides some context. Language is slightly exuberant."

ChatGPT response (detail): "These mediators amplify the inflammatory process, attracting further immune cells and initiating a self-perpetuating cycle of inflammation and tissue damage."

Instructor comments: "Well written and scientifically accurate. However, lacks details on which immune cells and which mediators are involved in this process."

ChatGPT response (detail): "The resulting protease-antiprotease imbalance contributes to airflow obstruction, alveolar damage, and impaired gas exchange."

Instructor comments: "This is a generalised mechanism. No discussion in the specific association of these proteins and COPD pathogenesis."

Bard

Compared to ChatGPT the level of detail was lower and overall the essay lacked mechanistic insight and failed to demonstrate the depth of knowledge expected of a level 5 student. Bard performed well in providing a clear narrative and the standard of written text was high and plausible. The generated output was also repetitive throughout and lacked a high-level of context.

Examples:

Bard response (context): "It is the third leading cause of death worldwide, after heart disease and stroke. COPD is caused by a combination of genetic and environmental factors, including cigarette smoking, air pollution, and occupational dust exposure."

Instructor comments: "Good. Provides epidemiological context but provided no definitive numbers etc."

Bard response (detail): "However, the current understanding is that the disease is

caused by a combination of genetic and environmental factors, including cigarette smoking, air pollution, and occupational dust exposure.”

Instructor’s comments: “Repetitive—repeat of the intro.”

Bard response (detail): “COPD patients are more likely to get respiratory infections, which can damage the lungs and worsen the symptoms of COPD.”

Instructor comments: “No examples of respiratory infections provided, e.g adenovirus, influenza, RSV etc.”

Bing

The response generated by Bing scored lower in all categories compared to that generated by ChatGPT and Bard. The answer lacked mechanistic detail, context and key examples. Therefore, Bing did not demonstrate a deep understanding of the topic and provided generic content, even though the written text was plausible, it lacked accuracy and coherence.

Examples:

Bing response (detail): “These cells release inflammatory mediators, such as cytokines and chemokines, that recruit more inflammatory cells and fibroblasts (cells that produce collagen) to the site of injury. The fibroblasts produce collagen and other extracellular matrix components that thicken and stiffen the bronchiolar wall.”

Instructor comments: “Correct but lacks specific detail on which cytokines and chemokines are involved in COPD pathogenesis.”

Bing response (detail): “The inflammation is caused by chronic exposure to irritants, such as cigarette smoke, dust, or fumes, that stimulate the production of mucus and inflammatory mediators by the bronchial epithelium (lining of the airways).”

Instructor’s comments: “Well written but generic and repetitive.”

Level 6

For each generative AI model, the following prompt was used: Describe how bioinformatics is used in biomedical research. Please provide examples of how bioinformatics has helped in our understanding of human diseases. Please write your response as a 1000 word university level essay. Essays were evaluated in relation to the FHEQ level 6 descriptor and the prescribed marking criteria (Table 4). Four independent markers evaluated anonymised scripts. The inter-rater reliability and agreement between

Table 4 Output summary in response to level 6 essay question

LLM	Scientific accuracy	Mechanistic detail	Deviation	Context	Coherence	FHEQ level 6 total
ChatGPT	70.75% High	66.75% Medium	77% High	71.25% High	73.5% High	71.85% High
Bard	62.5% Medium	56.25% Medium	69.5% Medium	61.75% Medium	60.5% Medium	62.05% Medium
Bing	53% Medium	51.75% Medium	61% Medium	58.5% Medium	52% Medium	55.25% Medium
Kendall's W Sig.	1.0 0.018	0.813 0.039	1.0 0.018	0.817 0.038	1.0 0.018	0.877 < 0.001

assessors were measured by Kendall's coefficient of concordance for each paper, for each criteria and across the cumulative marks for each paper.

ChatGPT

The level of scientific accuracy and use of scientific terminology was of a high standard, and while some mechanistic detail was provided in places, the essay lacked specific examples expected from a level 6 student. ChatGPT produced a well written essay, with a clear narrative that could have plausibly been written by a human, scoring highly on context and cohesion. There was some revealing language usage (for instance the first example below), where colloquial or verbose text was generated, although in general placed the essay into appropriate context (second example), both in the introduction and the summary.

Examples:

ChatGPT response (context): "In the realm of modern biomedical research, the integration of biology, computer science, and data analysis has given birth to the burgeoning field of bioinformatics."

Instructor comments: "Slightly colloquial in language used."

ChatGPT response (context): "This essay explores the multifaceted applications of bioinformatics in biomedical research, highlighting how it has contributed to our understanding of human diseases through a variety of illustrative examples."

Instructor comments: "Well written—Puts essay into context."

ChatGPT response (detail): "Bioinformatics can be succinctly defined as the interdisciplinary domain that employs computational methods, statistical analyses, and information technologies to acquire, manage, and interpret biological data."

Instructor's comments: "Good definition."

ChatGPT response (detail): "To make sense of these voluminous datasets, bioinformatics employs algorithms, databases, and visualization tools to derive meaningful insights, ultimately enhancing our comprehension of fundamental biological processes and their perturbations in disease states."

Instructor comments: "Generalised statement, e.g., which datasets are you referring to?"

ChatGPT response (detail): "Genome-wide association studies (GWAS) leverage large-scale genotyping data to identify genetic loci linked to disease susceptibility."

Instructor comments: "Correct—but you fail to mention the specifics of how bioinformatics achieves this. Need to provide relevant examples of bioinformatic tools/applications, e.g. GWAS Interrogator."

Bard

Google Bard struggled with scientific accuracy and correct use of scientific terminology. The output contained several generalisation or superficial descriptions of key elements related to the question and performed relatively poorly on mechanistic detail and context compared to ChatGPT. The generated output also lacked a narrative and therefore also performed poorly compared to ChatGPT in this criteria. However, there was little deviation away from the topic of the question.

Examples:

Bard response (context): "Bioinformatics is a powerful tool that is having a major impact on biomedical research."

Instructor's comments: "Generic statement—no context or indication of what this impact might be."

Bard response (accuracy): "Bioinformatics can be used to design new vaccines by identifying the genes that are needed for a pathogen to cause disease."

Instructor's comments: "Vaccine targets are not necessarily the same molecules that are responsible for disease."

Bard response (accuracy): "Metabolomics is the study of metabolites, which are the small molecules that cells use to generate energy, build structures, and communicate with each other."

Instructor's comments: "They are also the by-products of biochemical reactions."

Bard response (detail): "Bioinformatics can be used to track the spread of diseases by analyzing genetic data from patients. This information can be used to identify high-risk populations and to develop interventions to prevent the spread of diseases."

Instructor's comments: "How? Be more specific in your descriptions. Lacks examples."

Bard response (detail): "This information can be used to improve the effectiveness of drugs and to reduce side effects."

Instructor's comments: "How—you need to provide more details and specific examples."

Bing

Bing generated a response that scored lower compared to ChatGPT and Bard, with several generic statements and an overall lack of scientific accuracy and mechanistic detail, with little evidence of specific examples. For a level 6 answer, students would be expected to provide key examples in the context of human disease, biological mechanisms and connections to the scientific literature. In addition, there was also some deviation away from the topic and the answer lacked coherence and a logical narrative, scoring poorly in these areas. Overall, Bing performed relatively poorly compared to ChatGPT and Bard.

Examples:

Bing response (context): "Bioinformatics is an interdisciplinary field that combines biology, computer science, mathematics, statistics, and engineering."

Instructor comments: "Correct. You could have expanded on the application of bioinformatics in biomedical research."

Bing response (detail): "Bioinformatics can help analyze the structure and function of proteins, which are the building blocks of life."

Instructor comments: "Generic statement. I would expect more detail in an essay at this level."

Bing response (detail): "Proteins perform various tasks in the cell, such as catalyzing reactions, transporting molecules, and signaling pathways. Bioinformatics can help predict protein structure from sequence data, identify protein-protein interactions, and design drugs that target specific proteins."

Instructor comments: “This is correct but this is basic textbook understanding. How does this relate to the different types of protein bioinformatic analysis tools available?”

Bing response (detail): “Bioinformatics has helped to elucidate the genetic and environmental factors that contribute to neurodegenerative diseases such as Alzheimer’s, Parkinson’s, and Huntington’s.”

Instructor comments: “Good disease examples. Could you provide further details of what specific biomarkers have been identified, for example.”

Level 7

For each generative AI model, the following prompt was used: Describe the major histocompatibility complex II antigen processing pathway. Please provide details of the key molecular interactions at each step. Please write your response as a 1000 word university level essay. Essays were evaluated in relation to the FHEQ level 7 descriptor and the prescribed marking criteria (Table 5). Four independent markers evaluated anonymised scripts. The inter-rater reliability and agreement between assessors were measured by Kendall’s coefficient of concordance for each paper, for each criteria and across the cumulative marks for each paper.

ChatGPT

ChatGPT performed comparatively better than either Bard or Bing in response to the level 7 essay question, attaining an average grade of 69.75% (borderline first class answer). Once again ChatGPT demonstrated a high degree of scientific accuracy and correct use of scientific terminology, although specific points relating to key mechanisms lacked the insight of a level 7 student in places. Several mechanisms relating to the question were covered adequately and the depth and detail were of a standard expected of a MSc student. The essay was written well, with human quality grammar, syntax and cohesion. Some context was provided, comparable to that written by a human, although elements were misplaced or out of context to fully demonstrate logical narrative.

Examples:

ChatGPT response (context): “This essay delves into the various steps of the MHC II antigen processing pathway, elucidating the key molecular interactions at each stage.”

Instructor comments: “Good introduction—put the essay into context well.”

Table 5 Output summary in response to level 7 essay question

LLM	Scientific accuracy	mechanistic detail	Deviation	Context	Coherence	FHEQ level 7 total
ChatGPT	70.5% High	68% Medium	71.25% High	68.25% Medium	70.75% High	69.75% Medium–High
Bard	69% Medium	65% Medium	71.5% High	65.75% Medium	63.25% Medium	66.9% Medium
Bing	64.5% Medium	64.5% Medium	66.75% Medium	62.5% Medium	61.75% Medium	64% Medium
Kendall’s W Sig.	0.217 0.420	0.286 0.319	0.250 0.368	0.350 0.247	0.661 0.071	0.464 0.022

ChatGPT response (accuracy): "Within the endosomal compartments, the Ii undergoes proteolytic degradation, resulting in the removal of a fragment called the Class II-associated invariant chain peptide (CLIP)."

Instructor comments: "Actually, CLIP remains associated with the MHC II binding groove, until it is replaced by HLA-DM."

ChatGPT response (accuracy): "HLA-DM binds to the MHC II molecule and stabilizes it in a peptide-receptive state, enabling the efficient loading of antigenic peptides."

Instructor comments: "Not quite. MHC II molecules are unstable without a peptide (or CLIP) in place. HLA-DM facilitates the removal of CLIP and the association with kinetically stable antigenic peptide, resulting in loading of peptide with the highest stability."

ChatGPT response (detail): "The molecular interactions between trafficking proteins, such as SNAREs (soluble N-ethylmaleimide-sensitive factor attachment protein receptors) and Rab GTPases, ensure the precise targeting and fusion of vesicles containing MHC II-peptide complexes to the cell membrane."

Instructor comments: "Good, well written—but you could have expanded the series of molecular events that orchestrate this process."

Bard

Compared to ChatGPT, the essay lacked mechanistic insight and provided limited context and few key examples related to the biological process discussed. In comparison to ChatGPT this answer was slightly lower in quality. Although the sentence construction was plausible, the answer lacked cohesion or narrative and the output produced less meaningful context to the broader scientific understanding of the topic compared to ChatGPT.

Examples:

Bard response (context): "Macropinocytosis is a process of engulfing large amounts of extracellular fluid, along with any antigens that are present. APCs express receptors on their surface that bind to the extracellular matrix."

Instructor comments: "Sentence are disconnected. How is this interaction associated with macropinocytosis?"

Bard response (accuracy): "The endosomal compartment is a network of vesicles that are interconnected by tunnels."

Instructor comments: "What do you mean by 'tunnels'? What is the correct scientific terminology?"

Bard response (detail): "Once the particles are bound, they are engulfed by the APC and taken into a vesicle."

Instructor comments: "What are these specialised vesicles called, e.g. endosome, lysosome, phagolysosome?"

Bard response (detail): "This allows MHC class II molecules to bind to peptides that are generated by the degradation of exogenous antigens."

Instructor comments: "Lacks mechanistic details here. What about the role of CLIP, HLA-DM and HLA-DO?"

Bing

Some context was provided in the intro section. However, the essay did not fully capture the complexities of the biological processes asked in the essay question compared to ChatGPT or Bard. There were also some scientific inaccuracies (for example comment two below) and was therefore below the standard expected of a high attaining student at this level, although Bing did generate reasonably robust scientific output overall.

Examples:

Bing response (context): "The process of antigen processing and presentation by MHC II involves several steps and molecular interactions that are described below"

Instructor comments: "Good intro section—puts the essay into context."

Bing response (accuracy): "APCs can capture antigens from the extracellular environment by various mechanisms, such as phagocytosis, macropinocytosis, receptor-mediated endocytosis or cross-presentation."

Instructor comments: "Cross-presentation is associated with the MHC class I antigen presentation pathway, rather than the MHC II pathway."

Bing response (detail): "CLIP is then exchanged for an antigenic peptide by the action of HLA-DM, a molecule that catalyzes the release of CLIP and stabilizes the peptide-MHC II complex."

Instructor comments: "CLIP is then exchanged for an antigenic peptide by the action of HLA-DM, a molecule that catalyzes the release of CLIP and stabilizes the peptide-MHC II complex?"

Bing response (detail): "The Ii also contains a sorting signal that directs the MHC II-Ii complex to the endolysosomal pathway."

Instructor comments: "Good—but what is this sorting sequence? Please provide details"

ChatGPT, Bard and Bing comparison

Across all four FHEQ levels, ChatGPT scored the highest compared to the other LLMs, followed by Bard and then Bing, although no discernible performance difference was observed as the difficulty in the FHEQ level increased (Fig. 1). In particular, ChatGPT and Bard performed consistently high in all five marking criteria, while Bing tended to perform relatively poorly in scientific accuracy mechanistic detail, coherence and context, although performed slightly better in response to the level 7 essay question. All four LLMs showed a high level of conformity, deviating little from the subject material stated in the essay question. None of the AI models performed well at delivering key examples from the scientific literature or relating examples to the biological processes being discussed. In terms of essay structure and providing a clear narrative, ChatGPT performed better than Bard or Bing, with the latter models often producing lists of factual information rather than clear narrative and logical paragraph progression. In general, the marker agreement was high for all three levels of undergraduate essay questions, particularly for scientific accuracy, mechanistic detail, coherence and context, with some disagreement regarding the deviation criteria. There was lower marker agreement for the level 7 essay responses, although agreement was high when all five marketing criteria were evaluated simultaneously (total score).

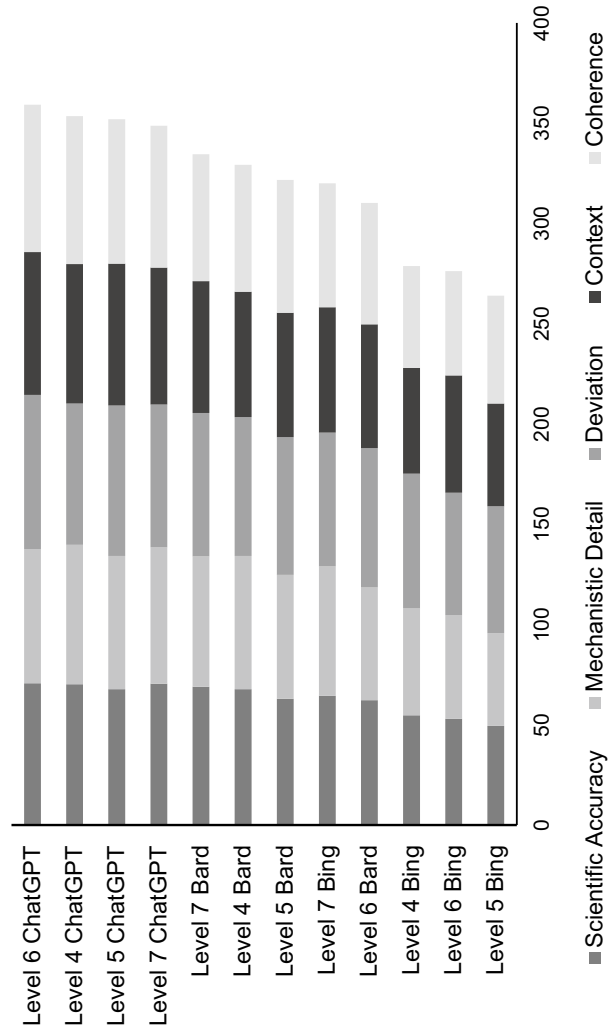


Fig. 1 Cumulative marking criteria score for ChatGPT, Bard and Bing across FHEQ levels

The average marker score was calculated for each marking criteria. Data is presented in ascending order of cumulative score for each generative AI tool and for each output generated in response to essay questions at FHEQ level 4, level 5, level 6 and level 7.

The text prompt for each AI included an instruction to provide a 1000-word output. However, none of the AI models consistently provided outputs of that length (Table 6), with ChatGPT generating an average essay length of 837, Google Bard of 619 and Microsoft Bing of 617, across all four FHEQ levels. Google Bard also falsified the final word count, thereby generating a discrepancy between the prompt and the output.

When asked a follow-up prompt to generate references for the essay text (level 7), none of the AI models were able to generate appropriate references from the scientific literature, although ChatGPT provided an explanatory paragraph. Google Bard provided genuine websites with relevant scientific information. Microsoft Bing generated references that were entirely fictitious, despite providing an additional, short description of the relevance of each falsified reference in the style of an annotated bibliography. Each falsified reference that was generated appeared consistent with standard scientific referencing styles, including author, date of publication, title, journal name, volume and page number. In response to the prompt: "Could you list 4 references to justify your answer?" the following responses were generated:

ChatGPT

"I apologize for any confusion, but as of my last knowledge update in September 2021, I do not have direct access to external sources such as specific references. Therefore, I am unable to provide you with a list of references. The essay I provided is based on information available up to that point and represents a synthesis of the knowledge and understanding of the MHC II antigen processing pathway as of that time. If you require up-to-date references, I recommend consulting scientific literature databases, textbooks, or reputable online sources related to immunology and antigen processing."

Bard

https://en.wikipedia.org/wiki/MHC_Class_II
<https://openstax.org/books/microbiology/pages/18-2-major-histocompatibility-complexes-and-antigen-presenting-cells>
https://en.wikipedia.org/wiki/Antigen_processing
<https://rupress.org/jem/article/191/9/1513/7984/Cytoplasmic-Processing-Is-a-Prerequisite-for>

Table 6 Summary of generated word counts

LLM	Level 4	Level 5	Level 6	Level 7	average
ChatGPT	979	743	935	691	837
Bard	1046	519	503	410	619.5
Bing	576	556	767	570	617.25

<https://microbenotes.com/mhc-molecules/>

Bing

Janeway, C. A., Travers, P., Walport, M., & Shlomchik, M. J. (2018). *Immunobiology: The immune system in health and disease* (10th ed.). Garland Science. {This textbook provides a comprehensive overview of the immune system, including the MHC class II antigen processing pathway.}

Wucherpfennig, K. W., & Strominger, J. L. (2000). The molecular basis of MHC class II-restricted antigen presentation. *Nature*, 408(6814), 797-804. {This review article provides a detailed molecular explanation of the MHC class II antigen processing pathway.}

Germain, R. N. (2009). MHC class II antigen presentation pathway. *Annual review of immunology*, 27, 489-517. {This review article provides an overview of the MHC class II antigen processing pathway, with a focus on the role of MHC class II molecules in the immune response.}

Zhu, J., & Cresswell, P. (2018). The MHC class II antigen processing pathway. *Current opinion in immunology*, 50, 1-8. {This review article provides a recent update on the MHC class II antigen processing pathway, with a focus on the role of HLA-DM and HLA-DO in peptide loading.}

Discussion

The study evaluates the performance of generative AI tools in writing essays for undergraduate and postgraduate biomedical sciences courses, according to well-defined marking criteria and using the FHEQ level descriptors as a benchmark standard for student attainment at each level of study (QAA, 2014). Three commonly used, and easily accessible, generative AI tools were evaluated, namely ChatGPT 3.5, Google Bard, and Microsoft Bing. AI generated output was anonymously evaluated by four independent markers, based on scientific accuracy, mechanistic detail, deviation, context, and coherence. The study selected example essay questions from courses at different undergraduate and postgraduate levels—undergraduate levels 4, 5, and 6, and postgraduate taught level 7, to assess the performance of generative AI tools across multiple levels. The higher the level descriptor, the greater the expectation for attaining the performance associated with each marking criteria. Illustrative examples from the AI responses were used to highlight areas of good and poor academic writing with respect to the marking criteria and FHEQ descriptors.

Implications

Previous studies on ChatGPT's performance in university assessments showed mixed results, with ChatGPT demonstrating equivalent performance on a typical MBA course and on the NBME 3rd year medical exam, although it previously performed better at factual recall questions rather than long-answer formats (Gilson et al., 2022; Ibrahim et al., 2023; Terwiesch, 2023). There is also some debate regarding the accuracy and reliability of AI chatbots in writing academically (Suaverdez & Suaverdez, 2023). Evaluating ChatGPT's performance, this study found that for all FHEQ level essays, there was

a high degree of scientific accuracy, with only minor factual errors, and high levels of context and coherence in the generated output. Overall, the evaluation indicated that while ChatGPT demonstrated high scientific accuracy and coherence, improvements in mechanistic detail are necessary to meet the standards expected of higher order essays. This was the case for undergraduate and postgraduate biomedical sciences courses. In evaluating ChatGPT's performance in writing essays for higher level 6 and level 7 biomedical sciences courses, the study found that the level of scientific accuracy and use of scientific terminology was high. However, the essays lacked specific examples expected from a student at these levels., although they were well written and featured a clear narrative that could have been plausibly written by a human. There were instances where revealing language usage, such as colloquial or verbose text, was generated. Despite this, the generated content generally placed itself into appropriate context. Overall, ChatGPT's output demonstrated a high level of scientific accuracy and terminology usage and outperformed Bard and Bing. These findings have implications for the potential use of ChatGPT in student assessments within the biomedical sciences, suggesting a need for further student training on the strengths and limitations of LLMs for different educational or subject-specific contexts. These findings underscore the potential limitations of AI tools in providing scientifically accurate and detailed examples from the literature, particularly in addressing nuanced concepts within the biomedical sciences curriculum.

Specific responses generated by the AI tool, Bard, to prompts related to the level 4 and level 5 essay questions, scored lower compared to ChatGPT, with differences in the scientific accuracy and depth of the information provided by Bard. In comparison, Bing, produced a good level of scientific detail at level 7 but performed consistently poorly compared to ChatGPT or Bard at UG level and lacked specific context or failed to connect cause and effect. In the context of undergraduate and postgraduate biomedical sciences courses, such deficiency may limit the educational value of the AI-generated content, particularly considering the scientific accuracy and mechanistic detail required in order to reach high attainment levels. The expectation at postgraduate level would be that these students also engage with critical thinking and evaluation, an area that generative AI is thought to perform particularly poorly at (Kasneci et al. 2023). The findings suggest that the three generative AI tools evaluated performed well at FHEQ level 7, which may reflect the mechanistic subject material of the essay question. In general, marker agreement was high across all undergraduate levels, for all marking criteria, except for deviation, which markers agreed less consistently on. Marker disagreement at level 7 was more evident and this may be due to individual marker idiosyncrasy and differences in academic judgement, which may be more pronounced for postgraduate level assessments.

Despite some limitations, the AI tools were able to generate essays that generally met the scientific accuracy criteria for both undergraduate and postgraduate levels. However, they also generated variation in the level of mechanistic detail, deviation, and coherence of the essays, with ChatGPT performing better than Bard or Bing, particularly in response to the undergraduate essay questions. Generative AI tools are also error prone, or suffer from hallucinations (Ahmad et al., 2023; Alkaissi & McFarlane, 2023), examples of which were reflected in this study. Teachers may therefore wish to supervise or provide guidance to students when utilizing these tools for educational purposes to ensure

scientific accuracy and sound academic writing, especially considering the value placed on such skills in biomedical and medical assessments, and professional practice (Ge & Lai, 2023). ChatGPT 3.5 is trained on an online data set that excludes scientific databases and therefore cannot access the scientific literature. This was further reflected in the poor provision of scientific references, although this aspect of certain generative AI models has been well documented (Fuchs, 2023). Training students in the appropriate use of generative AI technologies should also be a high priority for programmes and institutions. Such guidance should also be placed into the context of institutional academic integrity policies and the ethical use of AI more broadly.

Limitations

The current study was limited to comparing generative AI output between different AI tools and did not provide further comparisons with student generated text at equivalent levels. Although papers were anonymised prior to marking in the current study, blinded evaluations that directly compare AI and human responses would further elucidate the strengths and limitations of generative AI in this context. Further investigations on the impact of generative AI tools on academic writing and plagiarism in higher education, would be valuable in evaluating the potential consequences and challenges in maintaining academic integrity in student assessments. Similar research on the suitability of generative AI for diverse types of assessments, such as other long-form answers, short-answer questions and multiple-choice questions, would provide a better understanding of the limitations and possibilities of AI-generated content in meeting assessment requirements. Furthermore, the potential for integrating generative AI tools in the design and grading of written assessments, exploring the role of AI in providing efficient, fair, and accurate feedback to students, including within the context of peer review, and using AI to develop academic skills and critical thinking are areas of potential research.

Conclusion

Large language models have wide-ranging utilities in educational settings and could assist both teachers and students in a variety of learning tasks. This study evaluated the output of three commonly used generative AI tool across all levels of undergraduate and taught postgraduate biomedical science course essay assessments. ChatGPT performed better than Bard and Bing at all FHEQ levels, and across all marking criteria. Although all three generative AI tools generated output that was coherent and written to human standards, deficiencies were particularly evident in scientific accuracy, mechanistic detail and scientific exemplars from the literature. The findings suggest that generative AI tools can deliver the depth and accuracy expected in higher education assessments, highlighting their capabilities as academic writing tools. LLMs are continuously being updated, and new versions of AI technologies and applications are released regularly, which are likely to demonstrate improvements in the quality of generated output. This has implications for the way students use AI in their education and may influence how higher education institutions implement their policies on the use of generative AI. Finally, these findings may have broad implications for higher education teachers regarding the design of written assessments and maintaining academic integrity in the context of the rapid evolution of generative AI technologies.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s41239-024-00485-y>.

Supplementary Material 1.

Acknowledgements

I wish to thank faculty members at UCL for assisting with the independent marking, in particular Dr Johanna Donovan (Division of Medicine, UCL), Dr Sarah Koushyar (UCL Cancer Institute) and Dr John Logan (Division of Medicine, UCL) for supporting the study.

Author contributions

AW produced and analysed the data used in this study and was responsible for writing and editing the manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The author declares that they have no competing interests.

Received: 6 February 2024 Accepted: 15 August 2024

Published online: 13 September 2024

References

- Ahmad, Z., Kaiser, W., & Rahim, S. (2023). Hallucinations in ChatGPT: An unreliable tool for learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 15(4), 12.
- Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2), e35179.
- Aydin, Ö., & Karaarslan, E. (2023). Is ChatGPT leading generative AI? What is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems*, 11(3), 118–134.
- Behzadi, P., & Gajdác, M. (2021). Writing a strong scientific paper in medicine and the biomedical sciences: A checklist and recommendations for early career researchers. *Biologia Futura*, 72(4), 395–407. <https://doi.org/10.1007/s42977-021-00095-z>
- Cassidy, C. (2023). Australian universities to return to 'pen and paper' exams after students caught using AI to write essays. *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 Apr 2024.
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*. <https://doi.org/10.1080/14703297.2023.2190148>
- Doroudi, S. (2022). The intertwined histories of artificial intelligence and education. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-022-00313-2>
- Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse? *Frontiers in Education*. <https://doi.org/10.3389/feduc.2023.1166682>
- Ge, J., & Lai, J. C. (2023). Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. *Hepatology Communications*, 7(4), e0097.
- Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2022). How well does ChatGPT do when taking the medical licensing exams? The implications of large language models for medical education and knowledge assessment. *medRxiv*, 2022.2012.2022283901.
- Ibrahim, H., Liu, F., Asim, R., Battu, B., Benabderrahmane, S., Alhafni, B., Adnan, W., Alhanai, T., AlShebli, B., & Baghdadi, R. (2023). Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Scientific Reports*, 13(1), 12187.
- Kasneji, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hül-lermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneji, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Larsen, B. (2023). *Generative AI: A game-changer society needs to be ready for*. <https://www.weforum.org/agenda/2023/01/davos23-generative-ai-a-game-changer-industries-and-society-code-developers/>. Accessed Apr 2024.
- McGhee, P. (2003). *The academic quality handbook : assuring and enhancing learning in higher education*. Kogan Page Ltd.
- Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of artificial intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, 7(5), e07014. <https://doi.org/10.1016/j.heliyon.2021.e07014>
- Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*. <https://doi.org/10.53761/1.20.02.07>

- Puig, B., Blanco-Anaya, P., Bargiela, I. M., & Crujeiras-Pérez, B. (2019). A systematic review on critical thinking intervention studies in higher education across professional fields. *Studies in Higher Education*, 44(5), 860–869. <https://doi.org/10.1080/03075079.2019.1586333>
- QAA. (2014). *The frameworks for HE qualifications of UK degree-awarding bodies*. https://www.qaa.ac.uk/docs/qaa/quality-code/qualifications-frameworks.pdf?sfvrsn=170af781_18. Accessed Apr 2024.
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1), 342–63.
- Suaverdez, J., & Suaverdez, U. (2023). Chatbots impact on academic writing. *Global journal of Business and Integral Security*, (2).
- Team, O. (2022). *ChatGPT: Optimizing language models for dialogue*. OpenAI. <https://openai.com/blog/chatgpt>. Accessed Apr 2024.
- Terwiesch, C. (2023). *Would chat GPT3 get a Wharton MBA: a prediction based on its performance in the operations management course*. Mack Institute for Innovation Management/University of Pennsylvania/School Wharton.
- Woolf, BP. (2010). Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning. Morgan Kaufmann

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Andrew Williams is an Associate Professor of teaching at UCL. He is Deputy Director for BSc Applied Medical Sciences within the Division of Medicine and Programme Director and Academic Lead for UCL's portfolio of intercalated BSc (iBSc) programmes within UCL Medical School. He teaches on several undergraduate and postgraduate modules within the Faculty of Medical Sciences.