Article



Volume 25 (2026), pp. 279-304
*American Journal of STEM Education: Issues and Perspectives*
eISSN 30.3-1190| Print ISSN: 3069-0072
https://doi.org/10.32674/tfptev94

# Evaluating the Performance of 3 Large Language Models in Higher Education Essay-like Assessments in 2024 and 2026

David Hunt
*University of Worcester, United Kingdom*

Mathieu Di Miceli
*University of Leeds, United Kingdom*
*https://orcid.org/0000-0003-3713-0370*

## ABSTRACT

*Recent advances in artificial intelligence, especially generative large language models (LLMs), have transformed the higher education sector, raising concerns with academic integrity. The current literature lacks direct comparative analyses between LLMs. In the current study, we evaluated and compared the performance of ChatGPT, Gemini and Copilot (free versions) in 2024 and 2026, following prompts related to coursework essay assessments in computer science education or biomedical science. Our results indicate that LLMs struggle to abide by the word count beyond 1000 words, with Gemini presenting greater deviations. Copilot presented the lowest frequency of reference hallucinations. Overall performance of the 3 LLMs did not reveal any statistically significant differences. Quality assessments of the outputs revealed issues with content and criticality for all the LLMs. Similar performances were observed in 2024 and 2026.*

**Keywords:** Artificial intelligence; assessments; essays; large language models.

## INTRODUCTION

The first use of artificial intelligence (AI) was developed in the 20th century by Alan Turing (1912-1954), with Minsky (1927-2016) originally coining the term. In 2017, the transformer architecture (Vaswani et al., 2017) allowed language models to be created, which were stepping stones to the generative language processes. In fact, the earliest language processors, based on transformers, were not able to generate content. These tools were named "pre-trained transformers", since they were machine learning processes which were trained through large datasets and could be subsequently fine-tuned. Generative pre-trained tools were available a year later, in 2018, with the first generative pre-trained transformer (GPT) system developed by OpenAI (Radford & Narasimhan, 2018). This system included question answering, language inference, semantic classification and sentence similarity. The specific type of AI used to create text is known as generative "large language models", abbreviated as LLMs. Apart from OpenAI, Google (Devlin et al., 2019) and Microsoft (Peng et al., 2023) also developed LLMs. Many other LLMs exist, with the potential to transform learning and teaching methods (Xing et al., 2024).

## LITERATURE REVIEW

Many ethical questions have arisen as regards the use of AI in education (Lund et al., 2026; Strzelecki, 2024). In fact, both the secondary and higher education sectors are increasingly concerned about such usage, especially with regards to coursework assessments. As recently summarized, generative text can pose a threat in terms of academic integrity (Eke, 2023), as students may receive a grade for work that they did not produce. Furthermore, detecting auto-generated text remains a challenge (Farrelly & Baker, 2023), due to the question of the reliability of software detecting AI-generated text. Furthermore, the evidence suggests that LLM detectors place non-native English speakers at a disadvantage (Liang et al., 2023). Rightly or wrongly accusing students of using generative text tools can also have drastic consequences, for both institutions and individuals (Cotton et al., 2023; Stone, 2022). In a study conducted on 399 university students in Hong Kong (Chan & Hu, 2023), more than 60% of surveyed individuals reported using generative technologies, but not specifically for academic purposes.

Furthermore, the students were acutely aware of potential limitations and biases, demonstrating a very good overall understanding of the flaws of LLMs in higher education (Chan & Hu, 2023). In this study, the majority of the students nevertheless envisioned using these tools to support or enhance their learning. However, a lower score was obtained when students were asked whether they would agree with the statement on how AI use could undermine university education (Chan & Hu, 2023), highlighting the need for improvement in students' perception.

A systematic review of 41 studies demonstrates that generative AI and LLMs present both significant pedagogical opportunities and substantial risks to academic integrity in higher education (Bittle & El-Gayar, 2025). While these technologies can enhance accessibility and learning support, the literature consistently highlights concerns regarding plagiarism (Khalil & Er, 2023; Susnjak & McIntosh, 2024), ghost-writing, overreliance, and the limited reliability of current detection tools (Deep et al., 2025; Hadra et al., 2026). Consequently, the higher education sector currently emphasises the need for clear institutional policies, ethical AI literacy, expert assessment, and redesigned curricula that prioritise higher-order cognitive tasks to preserve academic integrity. Previous research has established potential drawbacks to using LLMs in academia. These include fabrication of references (Aljamaan et al., 2024; Bhattacharyya et al., 2023; Gravel et al., 2023; Walters & Wilder, 2023) and the generation of poor quality papers and figures (Haider et al., 2024), which can lead to articles being withdrawn.

Integrating LLMs as pedagogical tools in STEM education appears crucial (Bewersdorff et al., 2025), as it can allow students to better understand complex scientific concepts (El Fathi et al., 2025), highlighting how these tools can be used as supporting resources. For example, generative AI can be used to improve problem-solving skills, increase engagement, and help students with content creation (Redmond-Sanogo et al., 2026), whilst also increasing perceived productivity in students (Poudel et al., 2026). In computer science, students regularly use LLMs (Erez & Hazzan, 2025), and these models can enhance computational thinking skills (Huang & Qiao, 2024; Tian, 2024). However, several studies have highlighted that LLM use by students can have serious implications for academic integrity (Lund et al., 2026; Petingola et al., 2025; Strzelecki, 2024), especially when assessments are at stake (Mao et al., 2024). Furthermore, STEM educators also highlight concerns with ethics surrounding AI use (Coen & Cuddapah, 2026).

In this study, we aim to assess the performance of three LLMs in an educational setting. Using both quantitative and qualitative measurements, we directly compared the outputs generated by three different freely available LLMs (ChatGPT, Copilot, and Gemini) following prompts based on typical coursework essay instructions. In addition, we have measured the quality and appropriateness

of the reference list generated in these outputs. These aims will allow us to suggest improvements in coursework assessment designs while informing policies within the STEM higher education sector, which have proven to be inconsistent between higher education providers (Azevedo et al., 2024).

## RESEARCH METHOD

### Pilot Study

A pilot study was conducted in March 2024. These experiments were performed in ChatGPT version 3.5 (OpenAI, 2022) and aimed to establish the appropriate terminology to be used when prompting ChatGPT. The human prompts and ChatGPT-generated outputs are presented in the Supplementary Materials (section 1). During this phase, ChatGPT was given prompts related to a specific in-text citation format (Harvard) to check for the accuracy of the outputs. Next, the prompts were designed to assess whether ChatGPT could detect fabricated references. The final stage of the pilot study was designed to assess the reproducibility of the outputs produced by ChatGPT using the same prompt. This step allowed experimenters to design a reproducible prompt which could be used for all LLMs.

### Coursework Experiments

Following the pilot study, a blank scoring sheet was designed to quantify the quality of the outputs produced by three freely available LLMs: ChatGPT (OpenAI, 2022), Gemini (Google AI, 2024), and Copilot (Microsoft Corporation, 2024a). Each prompt was designed to specify word count in the output, minimum number of citations, the field of the topic (computer science education or biomedical sciences), and a typical coursework essay instruction. Experimenters designed the prompts according to their expertise, as detailed in Supplementary Materials (section 2). First, we quantified how many words were generated (excluding references) after asking for a specific word count (Supplementary Materials (section 3)) using triplicate prompts. Experimenters used the same prompt for each LLM, and outputs were scored. In total, 6 outputs were generated by each LLM. To avoid experimental bias, the same identical prompt was given to each LLM on the same day. The experimenters marked the outputs on a scale ranging from 0 to 100%, with a calibration at 40% for the minimum threshold for a pass mark, 50% for a 2:2 performance (lower second classification), 60% for a 2:1 performance (upper second classification), and 70% for a top performance (1st class attainment). In addition, the experimenters were also encouraged to provide a summary text justifying the scoring, whilst also being able to provide additional comments on any aspect of the output produced by the 3 LLMs. These qualitative

comments were then grouped according to themes: content, criticality, format, referencing, dates, quantitative or qualitative arguments.

There were 2 scorers for each output. Each experimenter scored LLM outputs in a blind manner to limit bias. The blind process ensured that no scoring bias could be introduced by disclosing previous scoring results. After experimenters finished scoring all outputs, the results were un-blinded for compiling (Supplementary Data). Finally, the reference lists generated were checked for accuracy and appropriateness. Each reference given by the LLMs was scrutinized for its Harvard style format (12th edition) by colour-coding the entire reference: names of authors (initials and surnames) and order, publication date, article title, journal title, issues and volumes, and pages. These analyses allowed the experimenters to assess the number of entirely fabricated references, the number of references with incorrect titles, incorrect names, incorrect author positions, incorrect dates, incorrect journal names, or incorrect volume/page numbers, as detailed in Supplementary Materials (section 4). In addition, markers assessed the appropriateness of each reference by determining whether the sentence(s) containing the corresponding in-text citation(s) accurately reflected the content of the cited source, when the cited reference could be retrieved. Experiments were performed in April 2024 and February 2026 using the online interfaces of the 3 LLMs and on the free versions.

**Data Analysis**

The data were analysed using RStudio version 4.0.3 (R Core Team, 2021). The required packages were fmsb and ltm, both available at the Comprehensive R Archive Network[1] (CRAN). Data normality was assessed using Shapiro-Wilk tests (Shapiro & Wilk, 1965). To compare 2 groups, normally-distributed data were analysed with Welch's t-tests (Welch, 1951), whilst Mann-Whitney U tests were used when data was not normally-distributed (Mann & Whitney, 1947).

To directly compare the performance of LLMs, one-, two- or three-way ANOVAs were used, as appropriate, followed by Šidák's (Šidák, 1967) or Tukey's (Tukey, 1949) post-hoc tests, as appropriate. To measure inter-rater scoring reliability between experimenters, intra-class correlation coefficients were computed, as the scores are measured on a 0-100% scale (equivalent to Cohen's kappa but for continuous data). For all statistical tests, p was set at 0.05 (a significant result when $p<0.05$).

All the data generated during the current study are available in the Supplementary Data (provided as a spreadsheet *via* an online repository[2]) or in the Supplementary Materials (also provided *via* an online repository[3]).

[1] https://cran.r-project.org/
[2] https://doi.org/10.6084/m9.figshare.27930330
[3] https://doi.org/10.6084/m9.figshare.32513574

## RESULTS

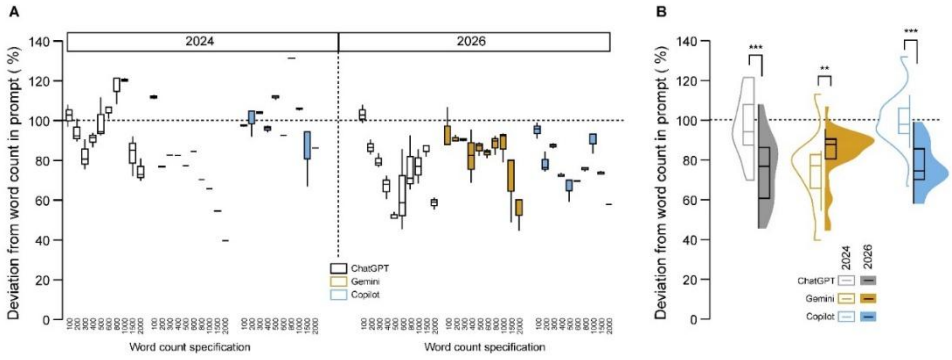
### Inter-Rater Reliability in Scoring

Two experimenters scored each LLM output following different essay-like coursework prompts (on the free versions). Reliability between scorers was assessed by computing intra-class correlation coefficients, which yielded 0.76 for scores in 2024 and 0.76 for 2026, indicating that internal reliability between markers was good.

### Accuracy of the Word Count in the Outputs

To assess the capacity of LLMs to produce outputs matching specific word count instructions from the prompts, we then generated triplicate outputs in each of the 3 LLMs using a fixed prompt approach (Supplementary Materials, section 3). The word count prompt specified from 100 to 2,000 words needed in the output, whilst keeping the rest of the instructions unchanged. Our results revealed that LLM output struggled to keep to the word count specifications in 2024 and 2026 (Figure 1A). In addition, we observed great variation in the length of the outputs produced by LLMs, with Gemini presenting greater deviations in 2024 than ChatGPT and Copilot, while results in 2026 show that all three LLMs struggle to produce longer essays (Figure 1B). When comparing the performance in 2024 versus 2026 (Figure 1B), ChatGPT produced significantly shorter outputs ($p<0.0001$) in 2026 (75.0 ± 16.5%) compared to 2024 (97.0% ± 15.5). Similarly, Copilot produced significantly shorter outputs ($p<0.0001$, Figure 1B) in 2026 (76.8 ± 11.4) than in 2024 (101.3 ± 13.9%). However, Gemini produced significantly longer ($p=0.04$) outputs in 2026 (93.0 ± 13.7%) compared to 2024 (74.6 ± 18.6%). Deviation from the word count specification was the greatest for Gemini (Figure 1B), achieving 39.7% of the required word count in 2024. When directly comparing the performance of the three LLMs in 2024 and 2026, we observed a significant impact of year ($F_{(1,174)}=31.99$, $p<0.0001$) and LLM ($F_{(2,174)}=7.26$, $p<0.0001$), with a significant interaction between LLM and year ($F_{(2,174)}=22.02$, $p<0.0001$). Altogether, these results indicate that freely-available LLMs struggle to generate outputs beyond 500 words.

**Figure 1**
*LLMs Struggle to Keep to the Word Count Specifications.*




## Reference Hallucination

Next, we investigated reference hallucination (Supplementary Materials, section 4). Out of the references displayed in the reference list for all outputs, several issues were identified, both in 2024 and 2026, which are summarized in Table 1. Common referencing mistakes for all three LLMs were reference fabrication and/or issues with retrieving the correct publication date, list of authors, journals, and volume/page numbers (Table 1). In 2024, reference retrieval by Gemini was poor for outputs related to biomedical sciences, whilst Copilot had almost faultless references. Reference hallucination was the highest for Gemini. In 2026, Copilot was the most accurate for reference transcription across both computer science education and biomedical sciences, whilst Gemini presented high proportions of reference fabrication. Overall, reference accuracy was the best for Copilot in both 2024 and 2026 (Figure 2), followed by ChatGPT and Gemini (both performing equally). These results suggest that reference retrieval by freely-available LLMs remains a significant issue, at least in the two fields investigated in the current study.

**Figure 2**
*Reference Hallucination by LLMs.*

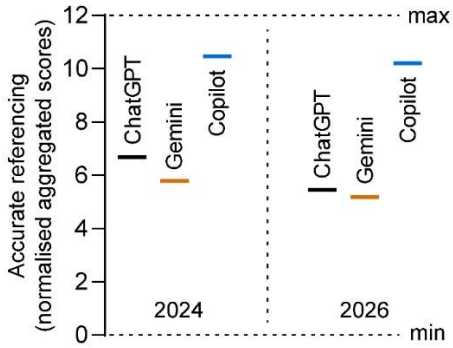



**Table 1**
*Reference Hallucination Details.*

| Items | April 2024 | | | | | |
|---|---|---|---|---|---|---|
| Number of … | Computer Science | | | Biomedical Science | | |
| | ChatGPT | Gemini | Copilot | ChatGPT | Gemini | Copilot |
| generated references | 10 | 9 | 8 | 17 | 25 | 11 |
| of which with errors | 4 (40) | 5 (56) | 4 (50) | 7 (41) | 24 (96) | 4 (36) |
| in titles | 4 (40) | 3 (33) | 3 (38) | 4 (24) | 23 (92) | 0 (0) |
| in journal names | 0 (0) | 2 (22) | 1 (13) | 1 (6) | 0 (0) | 0 (0) |
| in dates | 2 (20) | 2 (22) | 2 (25) | 4 (24) | 6 (24) | 4 (36) |
| in volumes/ pages | 4 (40) | 3 (33) | 0 (0) | 6 (36) | 20 (80) | 0 (0) |
| of which likely fabricated | 4 (40) | 3 (33) | 3 (37) | 3 (18) | 22 (88) | 0 (0) |
| Items | February 2026 | | | | | |
| Number of … | Computer Science | | | Biomedical Science | | |
| | ChatGPT | Gemini | Copilot | ChatGPT | Gemini | Copilot |
| generated references | 13 | 10 | 10 | 27 | 20 | 23 |
| of which with errors | 6 (46) | 3 (30) | 0 (0) | 17 (63) | 20 (100) | 10 (43) |
| in titles | 4 (31) | 3 (30) | 0 (0) | 10 (37) | 12 (60) | 9 (13) |
| in journal names | 0 (0) | 1 (10) | 0 (0) | 7 (26) | 6 (30) | 2 (9) |

| | | | | | | |
|---|---|---|---|---|---|---|
| in dates | 2 (15) | 2 (20) | 0 (0) | 2 (7) | 17 (85) | 1 (4) |
| in volumes/ pages | 5 (38) | 0 (0) | 0 (0) | 9 (33) | 15 (75) | 6 (26) |
| of which likely fabricated | 2 (15) | 3 (30) | 0 (0) | 7 (26) | 13 (65) | 5 (22) |

**Reference Appropriateness**

Out of all references generated by the three LLMs, not all could be retrieved (Table 1). However, amongst those which could be retrieved, major issues with accurate reflection of the content of the reference were detected (Table 2). Indeed, some references placed in the reference list did not have any associated in-text citation, others did not match the content of the source cited. The results for reference appropriateness exhibited substantial variability across LLMs, year of study and discipline areas.

**Table 2**

*Reference appropriateness.*

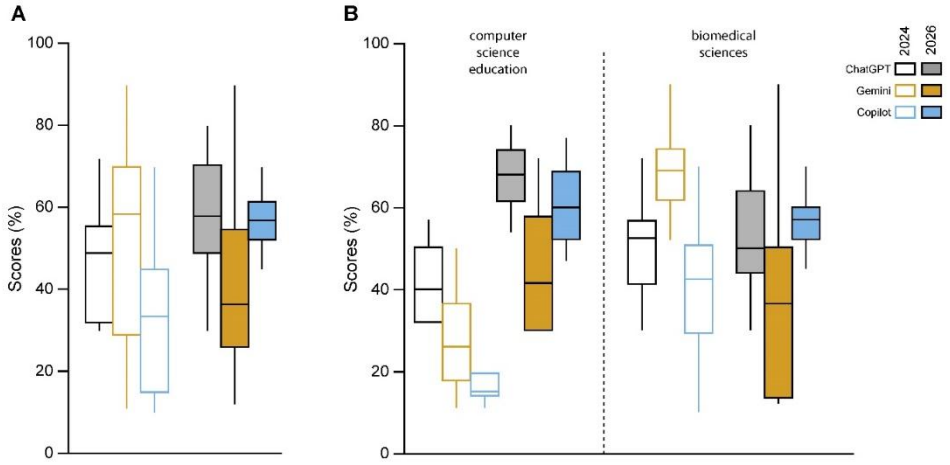
| Items | April 2024 | | | | | |
|---|---|---|---|---|---|---|
| Number of … | Computer Science | | | Biomedical Science | | |
| | ChatGPT | Gemini | Copilot | ChatGPT | Gemini | Copilot |
| retrievable references | 6 | 6 | 5 | 14 | 3 | 11 |
| of which with errors | 2 | 4 | 1 | 4 | 0 | 10 |
| no in-text citation | 1 | 0 | 1 | 0 | 0 | 0 |
| inappropriate content | 1 | 4 | 0 | 4 | 0 | 10 |
| Items | February 2026 | | | | | |
| Number of … | Computer Science | | | Biomedical Science | | |
| | ChatGPT | Gemini | Copilot | ChatGPT | Gemini | Copilot |
| retrievable references | 11 | 7 | 10 | 20 | 7 | 17 |
| which with errors | 5 | 2 | 1 | 6 | 7 | 2 |
| no in-text citation | 0 | 1 | 0 | 3 | 7[#] | 0 |
| inappropriate content | 5 | 1 | 1 | 3 | n/d | 2 |

**Performance of LLMs in Essay-Type Prompts**

ChatGPT, Gemini and Copilot (free versions) all performed similarly when scores given by markers were aggregated together (Figure 3A), with no differences between the performance of the LLMs ($F_{(2, 66)}$=0.86, p=0.43) and no differences between 2024 and 2026 ($F_{(1, 66)}$=2.85, p=0.10). However, a significant interaction between LLM and year was observed ($F_{(2, 66)}$=5.27, p=0.008), likely reflecting the wide range of performance across the LLMs. When differentiating output in computer science education to those in biomedical sciences (Figure 3B), we did not find differences in performance (non-significant three-way ANOVA). However, when taking outputs from computer science education only (Figure 3B), significantly greater scores were achieved in 2026 compared to 2024 ($F_{(1, 18)}$=23.60, p=0.0001), indicating that LLMs have significantly improved from 2024 to 2026. This improvement in the performance of the three LLMs was not observed for outputs in biomedical sciences ($F_{(1,42)}$=0.21, p=0.64).

**Figure 3**

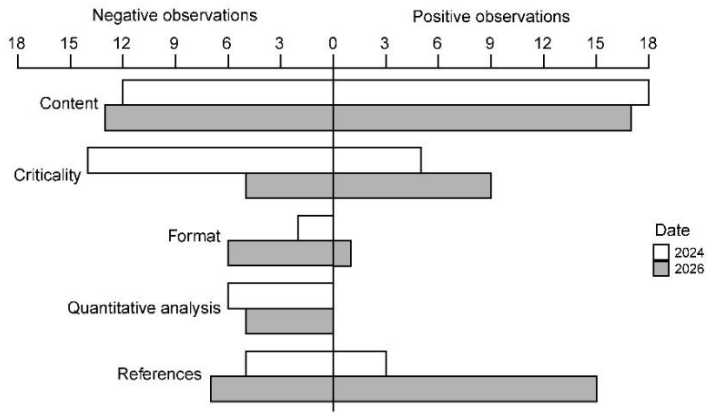
*Scores Achieved for Outputs by LLM in 2024 and 2026. Overall (A) and Subject-specific (B) Outputs Were Scored.*




Our qualitative analysis (Supplementary Materials, section 5) of the outputs produced by the three LLMs revealed 5 common themes: content, criticality, format, quantitative analysis and references (Figure 4). Whilst the quality of the content was generally praised, it could also be identified as a major pitfall of LLMs in both 2024 and 2026 (Figure 4). Furthermore, the absence of criticality was a major issue in 2024, but seem to have improved in 2026. For example, “lacking criticality”, “lacks specificity”, “vague/superficial” and “limited

analysis/explanations" were often cited by the markers as justifications for low scores given to outputs. These results suggest that the three LLMs tested in the present study can produce outputs with good elements, which are also accompanied by substantial issues, such as a lack of quantitative and critical analyses. We thus recommend markers to focus on critical analysis of evidence and the accuracy of the reference list when marking coursework essays.

**Figure 4**
*Qualitative Assessments for LLM Outputs.*




Altogether, the results of the present study demonstrate that ChatGPT, Gemini and Copilot present variable compliance with word count constraints, reflecting differences in instruction-following precision and output control. These disparities extend to output quality, where some outputs consistently contain good content, while others lack analytical depth. Notably, improved performance in meeting formal assessment instructions is often accompanied by a trade-off in reference reliability, as models generating more sophisticated text tend to exhibit increased citation fabrication.

## DISCUSSION

The performance of LLMs in addressing prompts reported here are in line with previous studies in politics (Dilling & Owen, 2024), dentistry (Giannakopoulos et al., 2023), medicine (Gilson et al., 2023; Huang et al., 2023; Tarabanis et al., 2024), mathematics (Zhang, Da, et al., 2024) and biomedical sciences (Feng et al., 2024). Accuracy in the outputs can sometimes be low, as reported for mathematics in higher education (Meissner et al., 2024). However, in medicine, LLMs appear to have good clinical knowledge (Truhn et al., 2023), but

human clinicians were able to perform significantly better (Hager et al., 2024; Singhal et al., 2023). It was also found that LLMs have superior divergent thinking capabilities (Hubert et al., 2024). Interestingly, critical thinking by the LLMs in the current study was weak, but another study in the field of tourism has found the opposite (Ülkü, 2023), although this study only used qualitative measurements.

We noted that LLMs could not adhere to the prompts when lengthier outputs were specified, which is similar to previous findings (Bai et al., 2024; Walters & Wilder, 2023), whilst also having decreased output quality when the word count requirement is increased (Choi et al., 2022). This aspect might be essential when designing assessment instructions in higher education, as LLMs struggle to comply with the word count. This is supported by another study in which the total word count increased with higher cognitive level demanded, with longer output produced by prompts requiring higher levels of explanations, although never going beyond 600 words, even for college-grade answers (Amin et al., 2024). The reason why LLMs cannot keep track of word counts in their output (McCoy et al., 2024) is simple: they operate on language prediction and probabilities. Generative LLMs, as their names suggest, generate new text based on prior text. The similarities, or differences, between LLMs and humans for word prediction have been observed in previous publications (Caucheteux & King, 2022; Contreras Kallens et al., 2023; Goldstein et al., 2022; Mitchell et al., 2008; Schrimpf et al., 2021).

Another major issue with LLMs is reference hallucination and appropriateness. This was detected in many previous studies (Athaluri et al., 2023; Chelli et al., 2024; Mugaanyi et al., 2024; Shen et al., 2023; Walters & Wilder, 2023). We report identical findings in the current study, with very frequent bibliographical errors in author names, article titles, journal volumes and/or page numbers, digital object identifiers and publication dates, in line with a previous report (Walters & Wilder, 2023). Perhaps the accuracy and suitability of bibliographical details could be used as frameworks for detecting LLM-generated outputs. We thus recommend that educators scrutinize the accuracy of the reference list in coursework submissions. Whilst this can be a time-consuming process, such a step is crucial to confirm human-generated text, as the current versions of the three LLMs tested in this study all present significant issues with reference hallucinations. In addition to reference hallucination, result hallucination has also been reported in outputs following oncological prompts (Huang et al., 2023), thus raising concerns (Farquhar et al., 2024), especially in the medical field (Clusmann et al., 2023; Giuffrè et al., 2024), and, more widely, in scientific subjects. This problem was also recently underlined by others (Azamfirei et al., 2023; Kobak et al., 2024).

The use of LLM as a support for learning might be the way forward, with the democratisation of such tools (Trenker et al., 2023). A longitudinal study in the Netherlands observed a significant decrease of ChatGPT use by university students (Polyportis, 2023). In fact, a study previously reported that using LLMs during

preparation for assessments could enhance student performance, or change the way students approach assignments (Bernabei et al., 2023). In computer science, LLMs could be used to create, annotate, proofread, edit and assess code (Hellas et al., 2023; Nam et al., 2024). This is also applicable to research in the field of ecology and evolution, although the quality of the code generated needs to be checked (Cooper et al., 2024). Some authors also suggested that LLMs could not only be used for code generation (Jiang et al., 2024), but also for data analysis (Nejjar et al., 2024). Thus, it was suggested to include LLMs as support tools in computer science education (Campbell et al., 2024), with some researchers even suggesting incorporating such educational tools in academic curricula (Mammides & Papadopoulos, 2024). Others also suggested a framework for designing good prompts in ChatGPT, especially applicable to computational biology (Lubiana et al., 2023).

We also report good marker homogeneity. Marking consistency in higher education essays is a challenge that directly impacts fairness, reliability, and the validity of summative grades. Previous research has shown that subjective scoring can lead to substantial variability among markers, with inter-marker reliability often compromised when criteria are ambiguous or when markers rely on general impressions rather than structured frameworks. Indeed, differences between experts and non-experts were reported in marking first year undergraduate biology reports (Bird & Yucel, 2013) and teacher candidates (Lyness et al., 2021). Timetabling specific sessions for students to engage with the marking criteria was also recommended, as this will enhance student engagement for upcoming assessments (Graham et al., 2022) whilst also decreasing anxiety (Taylor et al., 2024). The use of detailed marking criteria (Brookhart, 2018), also called rubrics, has been shown to improve consistency and agreement between markers, enhancing the reliability of grades and reducing bias (Jonsson & Svingby, 2007), though the degree of reliability achieved varies with rubric quality (Chakraborty et al., 2021) and marker training or experience (Benton, 2019). Consequently, developing clear assessment criteria and incorporating moderation processes are crucial for equitable and consistent essay marking practices in higher education.

## CONCLUSIONS & IMPLICATIONS

Our study reports on the performance of LLMs in producing essay-type responses across 2 different disciplines. The results have revealed notable differences between subjects and years. Despite these differences, scoring consistency between the two markers was high. In addition, LLMs demonstrated variability in adhering to specified word counts, struggling with longer answers. Performance analysis across LLMs showed overall similar capabilities. Issues with reference hallucination were less pronounced in Copilot compared to the other two models. Our qualitative assessments further underlined the challenges in content

accuracy, criticality, and reference integrity, pointing to areas for future improvement in LLMs. This study emphasises the potential and limitations of LLMs in educational settings, highlighting the need for continuous policy refinement in academia.

Based on the current and previous studies, coursework design should prioritise longer-form outputs, expert marking and rigorous evaluation of the quality and authenticity of the literature cited. We thus list 4 recommendations for colleagues to consider (Table 3), designed surrounding results obtained in the current study. Staff training should be designed on assessment design methodologies whilst marker training could focus on scrutinising reference hallucination. Coursework instructions with knowledge integration and data analysis might limit the capabilities of LLMs, although future updates may improve current performance. On an institutional level, the policies on generative AI should be made transparent to staff and students, whilst also allowing the said policies to evolve in accordance with the rapidly-changing technological landscape (Tsao, 2025).

Educators should thus focus on the following points when giving essay-style questions to students: A) are students allowed to use LLM, and are they aware of potential limitations to such use? B) Prioritise longer essay instructions in an attempt to limit LLM capabilities. C) Scrutinise the accuracy and suitability of the reference list. D) Evaluate critical analysis in the marking scheme.

**Table 3**

*Summary of Recommendations for Educators when Designing Coursework Assessments.*

| Item | Finding | Evidence | Implications | Design |
|---|---|---|---|---|
| 1 | Significant deviations from word count prompt. | Figure 2: LLMs struggle beyond 1,000 words. | Longer essays may limit LLM capabilities. | Ensure a minimum word count (or number of pages) is included in the instructions. |
| 2 | Some heterogeneity in scores depending on markers' expertise. | Figure 1 & 2B-C: differences between scores given by expert *versus* non-expert markers. | Staff expertise is linked to the scores given by markers. | Ensure marker training, standardisation and moderation processes are in place before/after marking. |
| 3 | Significant reference hallucinations | Figure 3D-E & Tables 1-2: reference | Reference hallucinations is an | Markers are encouraged to scrutinise the |

| | | | | |
|---|---|---|---|---|
| | and unsuitable content. | hallucinations vary between LLMs, with content mismatch. | indicator of LLM use. | accuracy and suitability of the reference lists. |
| 4 | Limited critical analysis. | Figure 4: several outputs lack critical analysis. | Critical analysis can be used as an indicator to human writing. | The marking criteria should include assessing students' critical analysis capacity. |

## LIMITATIONS

In the current study, only two scorers scrutinised the outputs produced by LLMs, which is one limitation of the current study, especially when compared to other studies in which the number of assessors was far greater (Feng et al., 2024). A few other studies have a similar number of assessors to the current study, all experts in their fields (Gilson et al., 2023; Hager et al., 2024; Meissner et al., 2024; Singhal et al., 2023; Truhn et al., 2023; Walters & Wilder, 2023). Interestingly, one study did not score outputs in a blind manner, which could have introduced a slight bias (Gilson et al., 2023). Another study only used one expert medical board-certified physician assessor who was blind to which answers were generated by humans or LLMs (Tarabanis et al., 2024). We also noted that another study did not report the number of assessors (Zhang, Da, et al., 2024). Our results were obtained using prompts related to computer science education and biomedical sciences. Extrapolations of these results beyond these two disciplines need to be carried out with caution, as the training of these LLMs might have induced discipline-related bias. As noted above, generative LLMs may exhibit substantial performance differences between subjects. In addition, the present study only focused on coursework-related prompts, the current results are thus not applicable to other forms of coursework such as data analysis, graphic display, short answer questions or reflective pieces.

The current study did not set specific temperatures for the LLMs on purpose, which are used to modulate the randomness of produced outputs. Our methodology was designed to reflect what students would do. Indeed, it is highly improbable that students would be aware that such a function is encoded in LLMs. In fact, it has been shown that ChatGPT produces hallucinations because its internal temperature is set between 0.7 (Beutel et al., 2023) and 1, the latter being the default mode of the Application Programming Interface (OpenAI, 2024). For Gemini, the default temperature value is also 1 (Google LLC, 2024a), whilst Copilot's default temperature is 0 (Microsoft Corporation, 2024b). Since temperature directly influences the degree of creativity in the LLMs, one can wonder how replicable experiments really are. Similarly to not setting a specific

temperature, we did not clear the prompt history, which was mitigated by not providing feedback to LLMs on all outputs generated. Again, this would align more closely with what a typical user would do.

Finally, one major limitation to the work presented here is how fast LLMs are evolving. Indeed, the experiments included in the current study were performed in April 2024 and February 2026. During that period, ChatGPT was upgraded to ChatGPT-4o, which was connected to the internet and had expanded capabilities (OpenAI, 2024). Similarly, Gemini and Copilot have also been upgraded (Google LLC, 2024b; Spataro, 2024). Thus, the results obtained in the current study might not remain fully applicable to LLMs in the future, due to their fast evolution (Tao et al., 2024), although an article emphasised that the technological architecture of LLMs remains the same (Wolfe, 2024).

**Use of Generative AI**

The authors did not use any AI tools in the drafting, writing, editing or refining of this manuscript. All content was generated, reviewed and refined solely by the authors.

**Declaration of Conflicts of Interest**

The authors declare that there are no conflicts of interest associated with the publication of this manuscript.

**Acknowledgements**

Authors wish to thank Dr. Peter Gossman for his initial scrutiny of the project aims and objectives, as well as Dr. Sian Evans and Pr. Ian Maddock for their support with administrative tasks.

**Funding**

This research was funded by the Learning, Teaching and Student Experience Project Fund (University of Worcester, to MDM). Authors express their gratitude to the University of Worcester for financial support (to MDM).

## REFERENCES

Aljamaan, F., Temsah, M. H., Altamimi, I., Al-Eyadhy, A., Jamal, A., Alhasan, K., Mesallam, T. A., Farahat, M., & Malki, K. H. (2024). Reference hallucination score for medical artificial intelligence chatbots: Development and usability study. *JMIR Medical Informatics, 12*(1), e54345. https://doi.org/10.2196/54345

Amin, K. S., Mayes, L. C., Khosla, P., & Doshi, R. H. (2024). Assessing the Efficacy of Large Language Models in Health Literacy: A Comprehensive

Cross-Sectional Study. *The Yale Journal of Biology and Medicine, 97*(1), 17–27. https://doi.org/10.59249/ZTOZ1966

Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, 15(4), e37432. https://doi.org/10.7759/cureus.37432

Azamfirei, R., Kudchadkar, S. R., & Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care, 27*(120). https://doi.org/10.1186/s13054-023-04393-x

Azevedo, L., Mallinson, D. J., Wang, J., Robles, P., & Best, E. (2024). AI policies, equity, and morality and the implications for faculty in higher education. *Public Integrity, 28*(2), 186-201. https://doi.org/10.1080/10999922.2024.2414957

Bai, Y., Zhang, J., Lv, X., Zheng, L., Zhu, S., Hou, L., Dong, Y., Tang, J., & Li, J. (2024). LongWriter: unleashing 10,000+ word generation from long context LLMs (arXiv:2408.07055). arXiv. https://doi.org/10.48550/arXiv.2408.07055

Benton, T. (2019). Which is better: One experienced marker or many inexperienced markers? *Research Matters: A Cambridge Assessment Publication, 28*. https://doi.org/10.17863/CAM.100393

Bernabei, M., Colabianchi, S., Falegnami, A., & Costantino, F. (2023). Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances. *Computers and Education: Artificial Intelligence, 5,* 100172. https://doi.org/10.1016/j.caeai.2023.100172

Beutel, G., Geerits, E., & Kielstein, J. T. (2023). Artificial hallucination: GPT on LSD? *Critical Care, 27*, 148. https://doi.org/10.1186/s13054-023-04425-6

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., & Nerdel, C. (2025). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learning and Individual Differences, 118*, 102601. https://doi.org/10.1016/j.lindif.2024.102601

Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus, 15*(5), e39238. https://doi.org/10.7759/cureus.39238

Bird, F. L., & Yucel, R. (2013). Improving marking reliability of scientific writing with the Developing Understanding of Assessment for Learning programme. *Assessment & Evaluation in Higher Education, 38*(5), 536–553. https://doi.org/10.1080/02602938.2012.658155

Bittle, K., & El-Gayar, O. (2025). Generative AI and Academic Integrity in Higher Education: A Systematic Review and Research Agenda. *Information, 16*(4), 296. https://doi.org/10.3390/info16040296

Brookhart, S. M. (2018). Appropriate Criteria: Key to Effective Rubrics. *Frontiers in Education, 3*, 22. https://doi.org/10.3389/feduc.2018.00022

Campbell, H., Bluck, T., Curry, E., Harris, D., Pike, B., & Wright, B. (2024). Should we still teach or learn coding? A postgraduate student perspective on the use of large language models for coding in ecology and evolution. *Methods in Ecology and Evolution, 15*(10), 1767–1770. https://doi.org/10.1111/2041-210X.14396

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology, 5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Chakraborty, S., Dann, C., Mandal, A., Dann, B., Paul, M., & Hafeez-Baig, A. (2021). Effects of rubric quality on marker variation in higher education. *Studies in Educational Evaluation, 70,* 100997. https://doi.org/10.1016/j.stueduc.2021.100997

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. International *Journal of Educational Technology in Higher Education, 20*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J. L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative Analysis. *Journal of Medical Internet Research, 26*(1), e53164. https://doi.org/10.2196/53164

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2022). ChatGPT goes to law school. *Journal of Legal Education, 71*(3), 387–400. https://doi.org/10.2139/ssrn.4335905

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J. N., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S. C., Unger, M., Veldhuizen, G. P., Wagner, S. J., & Kather, J. N. (2023). The future landscape of large language models in medicine. *Communications Medicine, 3*(1), 1–8. https://doi.org/10.1038/s43856-023-00370-1

Coen, A., & Cuddapah, J. L. (2026). Artificial intelligence in action: How preservice teachers embark on the fast-paced journey of mindful AI use. *American Journal of STEM Education, 20*, 111–132. https://doi.org/10.32674/keqjjt69

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science, 47*(3), e13256. https://doi.org/10.1111/cogs.13256

Cooper, N., Clark, A. T., Lecomte, N., Qiao, H., & Ellison, A. M. (2024). Harnessing large language models for coding, teaching and inclusion to empower research in ecology and evolution. *Methods in Ecology and Evolution, 15*(10), 1757–1763. https://doi.org/10.1111/2041-210X.14325

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, 61*(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148

Deep, P. D., Edgington, W. D., Ghosh, N., & Rahaman, M. S. (2025). Evaluating the Effectiveness and Ethical Implications of AI Detection Tools in Higher Education. *Information, 16*(10), 905. https://doi.org/10.3390/info16100905

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805). *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Dilling, M., & Owen, L. (2024). Designing politics and IR assessments in the era of AI: an empirical investigation into ChatGPT's output across Bloom's revised taxonomy. *Journal of Political Science Education, 21*(2), 290-309. https://doi.org/10.1080/15512169.2024.2408767

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology, 13*, 100060. https://doi.org/10.1016/j.jrt.2023.100060

El Fathi, T., Saad, A., Larhzil, H., Lamri, D., & Al Ibrahmi, E. M. (2025). Integrating generative AI into STEM education: Enhancing conceptual understanding, addressing misconceptions, and assessing student acceptance. *Disciplinary and Interdisciplinary Science Education Research, 7*(1), 6. https://doi.org/10.1186/s43031-025-00125-z

Erez, Y., & Hazzan, O. (2025, May 22). Students in advanced computational fields are accelerated early adapters of generative AI technology – *Communications of the ACM*. https://cacm.acm.org/blogcacm/students-in-advanced-computational-fields-are-accelerated-early-adapters-of-generative-ai-technology/

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

Farrelly, T., & Baker, N. (2023). Generative Artificial Intelligence: Implications and Considerations for Higher Education Practice. *Education Sciences, 13*(11), Article 1109. https://doi.org/10.3390/educsci13111109

Feng, H., Ronzano, F., LaFleur, J., Garber, M., de Oliveira, R., Rough, K., Roth, K., Nanavati, J., Abidine, K. Z. E., & Mack, C. (2024). Evaluation of large language model performance on the biomedical language understanding and reasoning benchmark: comparative study. *medRxiv*. https://doi.org/10.1101/2024.05.17.24307411

Giannakopoulos, K., Kavadella, A., Aaqel Salim, A., Stamatopoulos, V., & Kaklamanos, E. G. (2023). Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing chat in supporting evidence-based dentistry: comparative mixed methods study. *Journal of Medical Internet Research, 25*, e51580. https://doi.org/10.2196/51580

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education, 9*(1), e45312. https://doi.org/10.2196/45312

Giuffrè, M., You, K., & Shung, D. L. (2024). Evaluating ChatGPT in Medical Contexts: The Imperative to Guard Against Hallucinations and Partial Accuracies. *Clinical Gastroenterology and Hepatology, 22*(5), 1145–1146. https://doi.org/10.1016/j.cgh.2023.09.035

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience, 25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Google AI. (2024). *Gemini. Language Model.* https://gemini.google.com/app

Google LLC. (2024a). Generative AI. Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/adjust-parameter-values

Google LLC. (2024b). Gemini Apps' release updates and improvements. Gemini. https://gemini.google.com/updates

Graham, A. I., Harner, C., & Marsham, S. (2022). Can assessment-specific marking criteria and electronic comment libraries increase student engagement with assessment and feedback? *Assessment & Evaluation in Higher Education, 47*(7), 1071–1086. https://doi.org/10.1080/02602938.2021.1986468

Gravel, J., D'Amours-Gravel, M., & Osmanlliu, E. (2023). Learning to Fake It: Limited Responses and Fabricated References Provided by ChatGPT for Medical Questions. *Mayo Clinic Proceedings: Digital Health, 1*(3), 226–234. https://doi.org/10.1016/j.mcpdig.2023.05.004

Hadra, M., Cambridge, K., & Mesbah, M. (2026). Evaluating the accuracy and reliability of AI content detectors in academic contexts. *International Journal for Educational Integrity, 22*, 4. https://doi.org/10.1007/s40979-026-00213-1

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024).

Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine, 30*(9), 2613–2622. https://doi.org/10.1038/s41591-024-03097-1

Haider, J., Söderström, K. R., Ekström, B., & Rödl, M. (2024). GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation. *Harvard Kennedy School Misinformation Review, 5*(5), 1-16. https://doi.org/10.37016/mr-2020-156

Hellas, A., Leinonen, J., Sarsa, S., Koutcheme, C., Kujanpää, L., & Sorva, J. (2023). Exploring the responses of large language models to beginner programmers' help requests. *ICER '23 Proceedings of the 2023 ACM Conference on International Computing Education Research, 1,* 93–105. https://doi.org/10.1145/3568813.3600139

Huang, X., & Qiao, C. (2024). Enhancing computational thinking Skills through artificial intelligence education at a STEAM high school. *Science & Education, 33*(2), 383–403. https://doi.org/10.1007/s11191-022-00392-6

Huang, Y., Gomaa, A., Semrau, S., Haderlein, M., Lettmaier, S., Weissmann, T., Grigo, J., Tkhayat, H. B., Frey, B., Gaipl, U., Distel, L., Maier, A., Fietkau, R., Bert, C., & Putz, F. (2023). Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: Potentials and challenges for AI-assisted medical education and decision making in radiation oncology. *Frontiers in Oncology, 13*, 1265024. https://doi.org/10.3389/fonc.2023.1265024

Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports, 14*(1), 3440. https://doi.org/10.1038/s41598-024-53303-w

Jiang, J., Wang, F., Shen, J., Kim, S., & Kim, S. (2024). A survey on large language models for code generation (arXiv:2406.00515). *arXiv*. https://doi.org/10.48550/arXiv.2406.00515

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Khalil, M., & Er, E. (2023). Will ChatGPT get you caught? Rethinking of plagiarism detection. In P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies* (pp. 475–487). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-34411-4_32

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2024). Delving into ChatGPT usage in academic writing through excess vocabulary (arXiv:2406.07016). *arXiv*. https://doi.org/10.48550/arXiv.2406.07016

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Lubiana, T., Lopes, R., Medeiros, P., Silva, J. C., Goncalves, A. N. A., Maracaja-Coutinho, V., & Nakaya, H. I. (2023). Ten quick tips for harnessing the power of ChatGPT in computational biology. *PLoS Computational Biology, 19*(8), e1011319. https://doi.org/10.1371/journal.pcbi.1011319

Lund, B., Mannuru, N. R., Teel, Z. A., Lee, T. H., Ortega, N. J., Simmons, S., & Ward, E. (2026). Student perceptions of AI-assisted writing and academic integrity: ethical concerns, academic misconduct, and use of generative AI in higher education. *AI in Education, 1*(1), 2. https://doi.org/10.3390/aieduc1010002

Lyness, S. A., Peterson, K., & Yates, K. (2021). Low inter-rater reliability of a high stakes performance assessment of teacher candidates. *Education Sciences, 11*(10), 648. https://doi.org/10.3390/educsci11100648

Mammides, C., & Papadopoulos, H. (2024). The role of large language models in interdisciplinary research: Opportunities, challenges and ways forward. *Methods in Ecology and Evolution, 15*(10), 1774–1776. https://doi.org/10.1111/2041-210X.14398

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Mao, J., Chen, B., & Liu, J. C. (2024). Generative Artificial Intelligence in Education and Its Implications for Assessment. *TechTrends, 68*(1), 58–66. https://doi.org/10.1007/s11528-023-00911-4

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences of the United States of America*, 121(41), e2322420121. https://doi.org/10.1073/pnas.2322420121

Meissner, R., Pögelt, A., Ihsberner, K., Grüttmüller, M., Tornack, S., Thor, A., Pengel, N., Wollersheim, H. W., & Hardt, W. (2024). LLM-generated competence-based e-assessment items for higher education mathematics: Methodology and evaluation. *Frontiers in Education, 9,* 1427502. https://doi.org/10.3389/feduc.2024.1427502

Microsoft Corporation. (2024a). *Microsoft CoPilot.* https://copilot.microsoft.com

Microsoft Corporation. (2024b). *Model selection and temperature settings.* https://learn.microsoft.com/en-us/ai-builder/prompt-modelsettings

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(5880), 1191–1195. https://doi.org/10.1126/science.1152876

Mugaanyi, J., Cai, L., Cheng, S., Lu, C., & Huang, J. (2024). Evaluation of large language model performance and reliability for citations and references in

scholarly writing: cross-disciplinary study. *Journal of Medical Internet Research, 26*, e52935. https://doi.org/10.2196/52935
Nam, D., Macvean, A., Hellendoorn, V., Vasilescu, B., & Myers, B. (2024). Using an LLM to help with code understanding (arXiv:2307.08177). *arXiv*. https://doi.org/10.48550/arXiv.2307.08177
Nejjar, M., Zacharias, L., Stiehle, F., & Weber, I. (2024). LLMs for science: Usage for code generation and data analysis. *Journal of Software: Evolution and Process*, *35*(1) e2723. https://doi.org/10.1002/smr.2723
OpenAI. (2022, November 30). *Introducing ChatGPT.* https://openai.com/index/chatgpt/
OpenAI. (2024). *API reference*. https://platform.openai.com
OpenAI. (2024, May 13). *Introducing GPT-4o and more tools to ChatGPT free users.* https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/
Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: evidence from GitHub Copilot (arXiv:2302.06590). *arXiv*. https://doi.org/10.48550/arXiv.2302.06590
Petingola, M., Zhang, Y., Yan, Y., & Lin, W. (2025). Integrating ethical AI tools into educational practices for enhancing academic integrity. *CUI '25 Proceedings of the 7th ACM Conference on Conversational User Interfaces*, *6*, 1–6. https://doi.org/10.1145/3719160.3737626
Polyportis, A. (2023). A longitudinal study on artificial intelligence adoption: Understanding the drivers of ChatGPT usage behavior change in higher education. *Frontiers in Artificial Intelligence, 6*, 1324398. https://doi.org/10.3389/frai.2023.1324398
Poudel, P., Chouchen, M., Subedi, S., Gaulee, U., Bista, K., Bista, K., & Bhattarai, U. (2026). AI adoption and student productivity in higher education: A cross-institutional study. *American Journal of STEM Education, 20*, 117–134. https://doi.org/10.32674/2wda3n83
R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/
Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training. https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035
Redmond-Sanogo, A., Burton, M., Ivy, J., & Maiorca, C. (2026). Generative AI in mathematics, science, and STEM education: research, applications, and emerging themes. *School Science and Mathematics, 126*(1), 3–8. https://doi.org/10.1111/ssm.70002
Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of

language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118
Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591–611. https://doi.org/10.2307/2333709
Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models are double-edged swords. *Radiology, 307*(2), e230163. https://doi.org/10.1148/radiol.230163
Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62*(318), 626–633. https://doi.org/10.2307/2283989
Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2
Spataro, J. (2024, September 16). *Microsoft 365 Copilot Wave 2: Pages, Python in Excel, and agents. Microsoft 365 Blog*. https://www.microsoft.com/en-us/microsoft-365/blog/2024/09/16/microsoft-365-copilot-wave-2-pages-python-in-excel-and-agents/
Stone, A. (2022). Student perceptions of academic integrity: a qualitative study of understanding, consequences, and impact. *Journal of Academic Ethics, 21*, 357-375. https://doi.org/10.1007/s10805-022-09461-5
Strzelecki, A. (2024). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments, 32*(9), 5142–5155. https://doi.org/10.1080/10494820.2023.2209881
Susnjak, T., & McIntosh, T. R. (2024). ChatGPT: the end of online exam integrity? *Education Sciences, 14*(6), 656. https://doi.org/10.3390/educsci14060656
Tao, Z., Lin, T.-E., Chen, X., Li, H., Wu, Y., Li, Y., Jin, Z., Huang, F., Tao, D., & Zhou, J. (2024). A survey on self-evolution of large language models (arXiv:2404.14387). *arXiv*. https://doi.org/10.48550/arXiv.2404.14387
Tarabanis, C., Zahid, S., Mamalis, M., Zhang, K., Kalampokis, E., & Jankelson, L. (2024). Performance of publicly available large language models on internal medicine board-style questions. *PLOS Digital Health, 3*(9), e0000604. https://doi.org/10.1371/journal.pdig.0000604
Taylor, B., Kisby, F., & Reedy, A. (2024). Rubrics in higher education: An exploration of undergraduate students' understanding and perspectives.

*Assessment & Evaluation in Higher Education, 49*(6), 799–809. https://doi.org/10.1080/02602938.2023.2299330

Tian, S. (2024). The integration of computational thinking and artificial intelligence serves to enhance the cognitive processes and skill acquisition of students. *ISAIE '24 Proceedings of the 2024 International Symposium on Artificial Intelligence for Education*, 564–567. https://doi.org/10.1145/3700297.3700394

Trenker, J., Menon, S. S., & Blumtritt, C. (2023, July). *A steam-engine moment to the computer age, generative AI is boosting efficiency and creativity to unprecedented heights.* Statista. https://www.statista.com/site/insights-compass-ai-generative-ai

Truhn, D., Weber, C. D., Braun, B. J., Bressem, K., Kather, J. N., Kuhl, C., & Nebelung, S. (2023). A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. *Scientific Reports, 13*(1), 20159. https://doi.org/10.1038/s41598-023-47500-2

Tsao, J. (2025). Trajectories of AI policy in higher education: Interpretations, discourses, and enactments of students and teachers. *Computers and Education: Artificial Intelligence, 9*, 100496. https://doi.org/10.1016/j.caeai.2025.100496

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*(2), 99–114. https://doi.org/10.2307/3001913

Ülkü, A. (2023). Artificial intelligence-based large language models and integrity of exams and assignments in higher education: The case of tourism courses. *Tourism & Management Studies, 19*(4), 21–34. https://doi.org/10.18089/tms.2023.190402

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *NIPS'17 Proceeding of the 31st International Conference on Neural Information Processing Systems*, 6000-6010. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*(1), 14045. https://doi.org/10.1038/s41598-023-41032-5

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38*(3/4), 330–336. https://doi.org/10.2307/2332579

Wolfe, C. R. (2024, August 22). *LLMs evolve quickly. Their underlying architecture, not so much*. https://stackoverflow.blog/2024/08/22/llms-evolve-quickly-their-underlying-architecture-not-so-much/

Xing, W., Zhu, T., Wang, J., & Liu, B. (2024). A survey on MLLMs in education: Application and future directions. *Future Internet, 16*(12), 467. https://doi.org/10.3390/fi16120467

Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Zhuang, C., Slack, D., Lyu, Q., Hendryx, S., Kaplan, R., Lunati, M., & Yue, S. (2024). A careful examination of large language model performance on grade school arithmetic (arXiv:2405.00332). *arXiv*. https://doi.org/10.48550/arXiv.2405.00332

---

**Bios**

**DAVID HUNT**, M.A., is a Senior Lecturer in Secondary Education (Computer Science) at the University of Worcester. Email: d.hunt@worc.ac.uk

**MATHIEU DI MICELI**, Ph.D., is a Lecturer in Human Anatomy and Physiology at the University of Leeds. Email: m.c.dimiceli@leeds.ac.uk