Does GenAI Write Good Science, and Does It Know Whether It Can?

Exploring the Ability of GenAI to Write and Evaluate Scientific Text

Nigel Francis & Andrew Shore

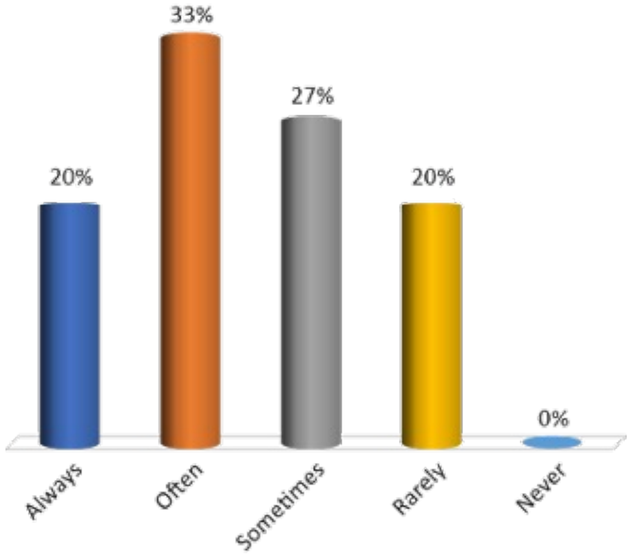
Introduction

The Rise of Generative Artificial Intelligence

- Large Language Models (LLMs) are impacting a wide variety of industries, including education and science
- LLMs are being used to draft, edit and refine outputs but can it evaluate the quality of scientific writing?
- Questions remain about their effectiveness, limitations, and potential biases

Why do we need to worry?

16 to 18-year-olds are a lot more engaged with AI



How often do you use AI tools?

What about current students?

UG students know about it...

53% used AI to help with assessments

65% think institutions should not accept AI genertated work

73% expect to use AI after graduation

Only 9% said institutional approaches have changed

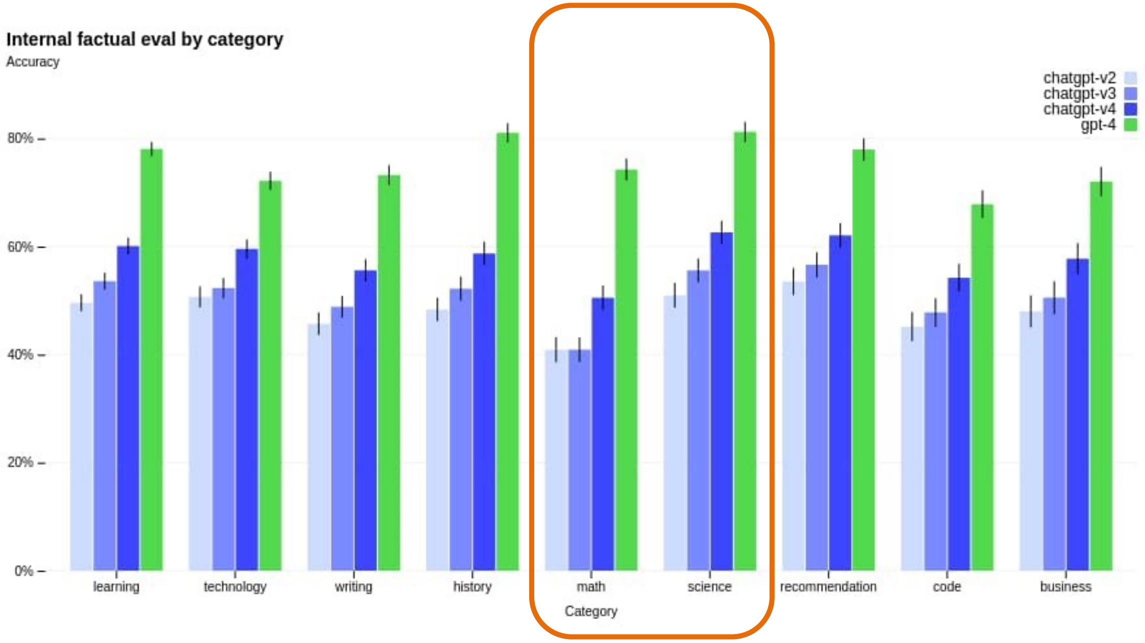
Provide or punish? Students' views on generative Ai in higher education (HEPI Policy Note 51, Feb 2024)

Why this matters

- Educational impact: Can AI support or hinder learning in academic contexts
- Equity of access: Does the difference between free and premium models exacerbate inequalities
- AI as a tool: How can educators and students integrate AI effectively in teaching and learning?

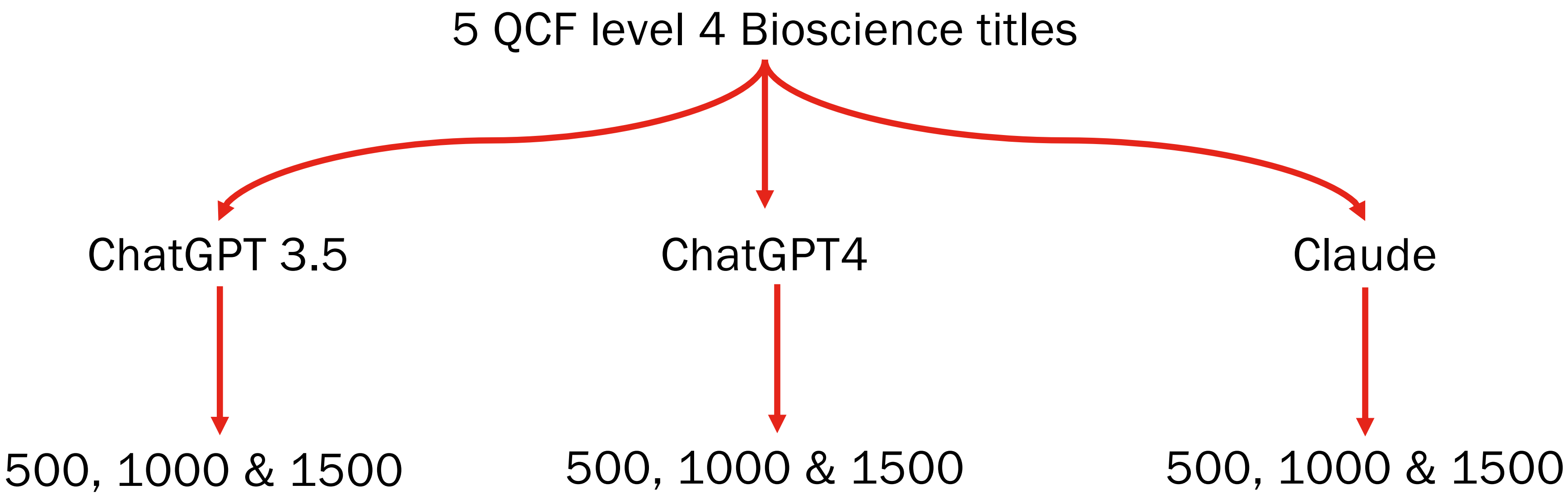
Purpose of the Study

- To assess whether GenAI can produce high quality scientific essays
- To evaluate whether GenAI can critique and mark essays it generates
- To explore the implications of free vs premium GenAI models in academic settings



Methods

AI Generation of Essays



5 QCF level 4 Bioscience titles

ChatGPT 3.5

ChatGPT4

Claude

500, 1000 & 1500

500, 1000 & 1500

500, 1000 & 1500

Total of 45 essays, 9 from each title.

AI Generation of Essays - Prompting

- “imagine you are a 1st year university student; I want you to write an essay based on this title: *insert title*.”
- “Can you use Harvard referencing for your sources through the essay and also provide a reference list?”
- “Can you make the essay 500 words? (Excluding references)”
- “Can you make the essay 1000 words? (Excluding references)”
- “Can you make the essay 1500 words? (Excluding references)”

- All essays were generated on the same date

Human Marking

- Each essay was ‘blinded’ so that the GenAI used was unknown to the markers
- Each essay was marked independently by 3 human markers who were final year undergraduate students
- Markers also provided qualitative feedback/justification of their mark
- Standardisation of marking was carried out as a group with guidance from an academic member of staff
- An established rubric/marking criteria was used for mark generation – the criteria already in use for summative assessment of level 4 students using the same essay titles.

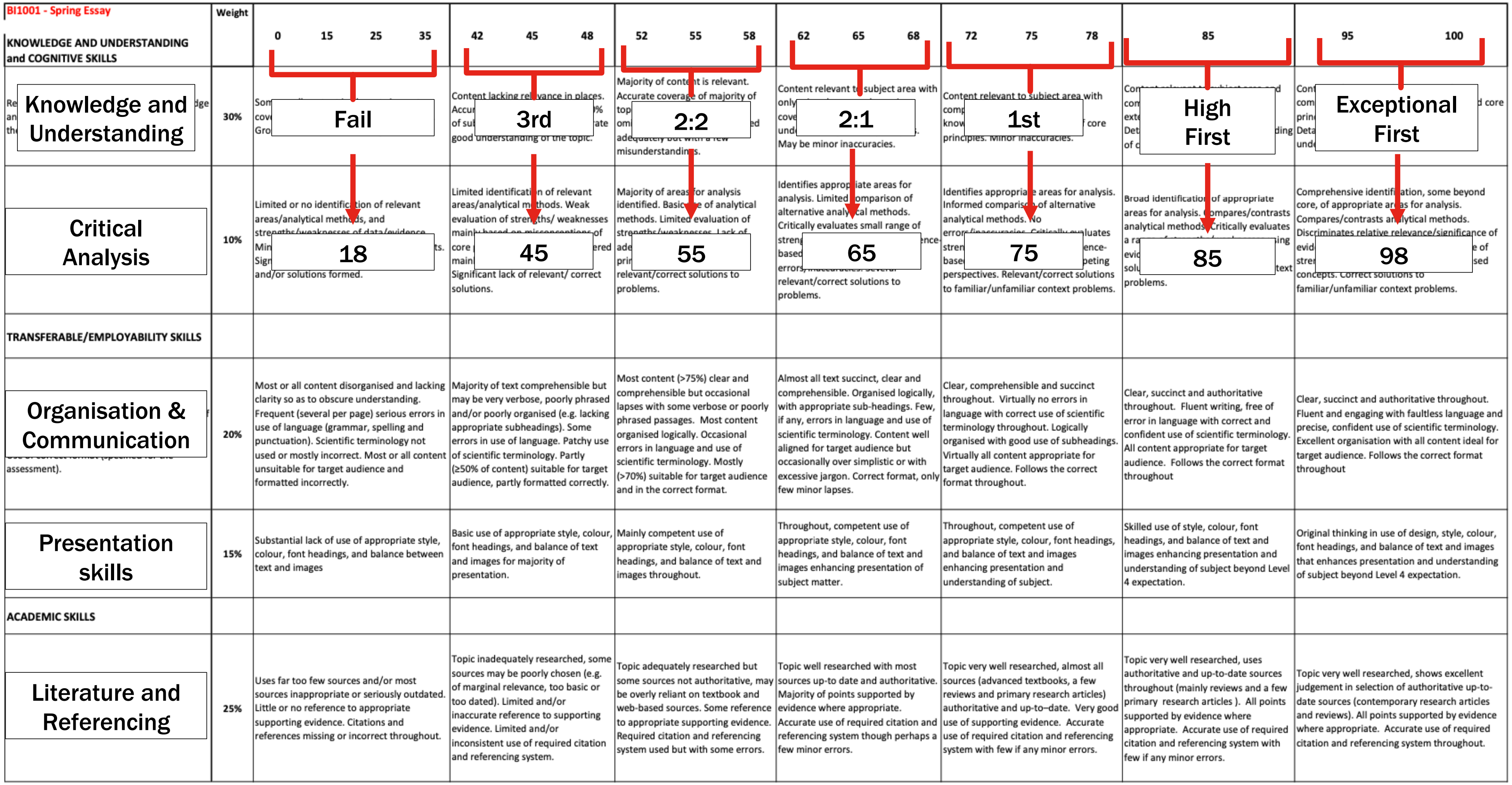
Generative AI Marking

- “I am going to provide you with an essay marking rubric for a first-year essay at a university. I want you to analyze the document and then provide me with a summary of the five marking criteria and their weighting so I can check you have interpreted it correctly.”
- I am now going to provide you with an example of a full essay, and I want you to tell me within which grade level from ‘fail’ to ‘exceptional first’ you think it falls for each of the criteria: *pasted essay text*”
- Each model was reloaded after each input.
- The models were not able to provide quantitative marks like the human markers so the rubric was modified as follows:

Generative AI Marking

- “I am going to provide you with an essay marking rubric for a first-year essay at a university. I want you to analyze the document and then provide me with a summary of the five marking criteria and their weighting so I can check you have interpreted it correctly.”
- I am now going to provide you with an example of a full essay, and I want you to tell me within which grade level from ‘fail’ to ‘exceptional first’ you think it falls for each of the criteria: *pasted essay text*”
- Each model was reloaded after each input.
- The models were not able to provide quantitative marks like the human markers so the rubric was modified as follows:

Generative AI Marking – Adapted from a figure by Ahmed Al-Sammere



BI1001 - Spring Essay KNOWLEDGE AND UNDERSTANDING and COGNITIVE SKILLS	Weight	0 15 25 35	42 45 48	52 55 58	62 65 68	72 75 78	85	95 100
Re an th dge Knowledge and Understanding	30%	Som cov Gro Fail	Content lacking relevance in places. Accur % of sul ate good understanding of the topic. 3rd	Majority of content is relevant. Accurate coverage of majority of top omi ed adequately but with a few misunderstandings. 2:2	Content relevant to subject area with only cove und . May be minor inaccuracies. 2:1	Content relevant to subject area with comp know f core principles. Minor inaccuracies. 1st	Content relevant to subject area and com exte Det ding of c High First	Cont com prin Deta und d core Exceptional First
Critical Analysis	10%	Limited or no identification of relevant areas/analytical methods, and strengths/weaknesses of data/evidence. Min ts. Sign and/or solutions formed. 18	Limited identification of relevant areas/analytical methods. Weak evaluation of strengths/ weaknesses mainly based on misconceptions of core ered main Significant lack of relevant/ correct solutions. 45	Majority of areas for analysis identified. Basic use of analytical methods. Limited evaluation of strengths/weaknesses. Lack of ade prin relevant/correct solutions to problems. 55	Identifies appropriate areas for analysis. Limited comparison of alternative analytical methods. Critically evaluates small range of streng ence- based errors/inaccuracies. Several relevant/correct solutions to problems. 65	Identifies appropriate areas for analysis. Informed comparison of alternative analytical methods. No errors/inaccuracies. Critically evaluates stren ence- base peting perspectives. Relevant/correct solutions to familiar/unfamiliar context problems. 75	Broad identification of appropriate areas for analysis. Compares/contrasts analytical methods. Critically evaluates a ra sing evid solu text problems. 85	Comprehensive identification, some beyond core, of appropriate areas for analysis. Compares/contrasts analytical methods. Discriminates relative relevance/significance of evid e of stre sed concepts. Correct solutions to familiar/unfamiliar context problems. 98
TRANSFERABLE/EMPLOYABILITY SKILLS								
Organisation & Communication assessment).	20%	Most or all content disorganised and lacking clarity so as to obscure understanding. Frequent (several per page) serious errors in use of language (grammar, spelling and punctuation). Scientific terminology not used or mostly incorrect. Most or all content unsuitable for target audience and formatted incorrectly.	Majority of text comprehensible but may be very verbose, poorly phrased and/or poorly organised (e.g. lacking appropriate subheadings). Some errors in use of language. Patchy use of scientific terminology. Partly (≥50% of content) suitable for target audience, partly formatted correctly.	Most content (>75%) clear and comprehensible but occasional lapses with some verbose or poorly phrased passages. Most content organised logically. Occasional errors in language and use of scientific terminology. Mostly (>70%) suitable for target audience and in the correct format.	Almost all text succinct, clear and comprehensible. Organised logically, with appropriate sub-headings. Few, if any, errors in language and use of scientific terminology. Content well aligned for target audience but occasionally over simplistic or with excessive jargon. Correct format, only few minor lapses.	Clear, comprehensible and succinct throughout. Virtually no errors in language with correct use of scientific terminology throughout. Logically organised with good use of subheadings. Virtually all content appropriate for target audience. Follows the correct format throughout.	Clear, succinct and authoritative throughout. Fluent writing, free of error in language with correct and confident use of scientific terminology. All content appropriate for target audience. Follows the correct format throughout	Clear, succinct and authoritative throughout. Fluent and engaging with faultless language and precise, confident use of scientific terminology. Excellent organisation with all content ideal for target audience. Follows the correct format throughout
Presentation skills	15%	Substantial lack of use of appropriate style, colour, font headings, and balance between text and images	Basic use of appropriate style, colour, font headings, and balance of text and images for majority of presentation.	Mainly competent use of appropriate style, colour, font headings, and balance of text and images throughout.	Throughout, competent use of appropriate style, colour, font headings, and balance of text and images enhancing presentation of subject matter.	Throughout, competent use of appropriate style, colour, font headings, and balance of text and images enhancing presentation and understanding of subject.	Skilled use of style, colour, font headings, and balance of text and images enhancing presentation and understanding of subject beyond Level 4 expectation.	Original thinking in use of design, style, colour, font headings, and balance of text and images that enhances presentation and understanding of subject beyond Level 4 expectation.
ACADEMIC SKILLS								
Literature and Referencing	25%	Uses far too few sources and/or most sources inappropriate or seriously outdated. Little or no reference to appropriate supporting evidence. Citations and references missing or incorrect throughout.	Topic inadequately researched, some sources may be poorly chosen (e.g. of marginal relevance, too basic or too dated). Limited and/or inaccurate reference to supporting evidence. Limited and/or inconsistent use of required citation and referencing system.	Topic adequately researched but some sources not authoritative, may be overly reliant on textbook and web-based sources. Some reference to appropriate supporting evidence. Required citation and referencing system used but with some errors.	Topic well researched with most sources up-to date and authoritative. Majority of points supported by evidence where appropriate. Accurate use of required citation and referencing system though perhaps a few minor errors.	Topic very well researched, almost all sources (advanced textbooks, a few reviews and primary research articles) authoritative and up-to–date. Very good use of supporting evidence. Accurate use of required citation and referencing system with few if any minor errors.	Topic very well researched, uses authoritative and up-to-date sources throughout (mainly reviews and a few primary research articles). All points supported by evidence where appropriate. Accurate use of required citation and referencing system with few if any minor errors.	Topic very well researched, shows excellent judgement in selection of authoritative up-to-date sources (contemporary research articles and reviews). All points supported by evidence where appropriate. Accurate use of required citation and referencing system throughout.

Results

Three Way ANOVA results showing the impact of each factor and significant interactions (P<0.05)

Adapted from an original figure by Mollie Ridge

	Degrees of Freedom	Sum of Squares	Mean Squares	F value	P - Value
Human Marker	2	38.9	19.43	0.483	0.62120
Essay Length	2	122.7	61.34	1.526	0.23281
Essay Subject	4	471.3	117.83	2.931	0.03583*
AI Model	2	581.1	290.56	7.228	0.00257**
Significant Interactions					
Human marker & Essay Subject	8	961.6	120.19	2.990	0.01277**
AI Model & Essay Length	4	604.5	151.12	3.759	0.01287**

Human Awarded Marks for AI generated essays

- There was a significant impact of AI model type on the awarded mark ($P<0.01$)
- Average essay marks were 3rd class
- Claude averaged the lowest marks and ChatGPT 4.0 the highest marks

- Impact of different human markers not significant
- Significant interaction between humans and essay title ($P<0.05$)
- Significant variation between essay title ($P<0.05$) but explained by Claude’s lower performance in some titles.

Impact of Essay length on Human Awarded Marks

- Effect of essay length was not significant
- Interaction between AI used and essay length was significant ($P<0.05$)
- Claude gained lower marks in the 500 word essays
- Little other effect of essay length with ChatGPT 3.0 performing slightly better at 1000 words and relatively consistent results for ChatGPT 4.0
- AI generated surpisingly similar word counts regardless of the length prompt

Qualitative Assessment of AI Generated Essays – content and presentation

- Lack of scientific detail
- Little or no discussion
- Reads like a list converted to prose
- No Figures
- Characteristic ‘awkward’ introductions

“This essay will adhere to the Harvard referencing style and provide a reference list, while aiming to be informative and academically rigorous.”

Qualitative Assessment of AI Generated Essays – recognition of sources

- Frequent use of ‘imagined’ references
- Real references but from irrelevant work by a real author, but who had worked in the area.
- Genuine titles but imagined authors
- Missing authors in the reference list
- Citations only appearing at the end of a paragraph
- Longer essays increased accuracy of referencing and some 1500 word essays had no significant errors.

Human and AI Awarded Marks for AI generated Essays.

- **Average human awarded markers for AI generated essays ($P<0.01$)**
- Claude – 41%
- ChatGPT 3.0 – 44%
- ChatGPT 4.0 – 46%

- **Average AI awarded marks for AI generated essays**
- Claude – 68%
- ChatGPT 3.0 – 68%
- Chat GPT 4.0 – 68%

The pattern of marks awarded to each essay was different for Human and AI awarded marks.

Qualitative Assessment of AI Marking of AI Essays

- AI marks were higher in every criteria of the rubric
- Presentation Skills. May have assumed that figures were present and accurate even if not present
- Academic Skills – referencing, did not identify flaws in referencing

Discussion and Conclusions

GenAI Performance in Essay Writing

- GPT4 consistently outperformed other models (Claude and GPT3.5) across most variables
- Essays averaged 3rd class marks – human evaluation
- Limitations in GenAI's ability to produce high-quality scientific writing from zero-shot prompts

Essay Length and Subject

- Essay length had limited impact on performance
- Longer essays showed slight improvements in referencing accuracy
- Certain essay subjects were handled better by GenAI
- High variability between models

GenAI vs Human Marking

- GenAI consistently awarded higher marks compared to human markers
- GenAI marking failed to identify referencing and formatting flaws

Strengths and Weaknesses of GenAI

Strengths:

- Efficiency in generating coherent, structured prose
- Potential as a supplementary tool for students, particularly in generating drafts or structuring arguments

Weaknesses:

- Lack of critical analysis and depth in content
- Over-reliance on fabricated or inappropriate references
- Inconsistent performance across topics and essay lengths

Strengths and Weaknesses of GenAI

Equity Considerations:

- Differences between free and premium models may widen educational inequalities

Educational Potential:

- GenAI can serve as a teaching aid but cannot replace human expertise
- Encouraging transparent use of GenAI is essential to maintain academic integrity

Conclusions

Summary of findings

- GenAI is a promising tool, but not yet a replacement for human scientific writing or evaluation
- Current GenAI models produce basic scientific tex but lack depth and detail for higher academic outcomes

Recommendations

- **For Students:** Use AI for initial brainstorming and draft generation but review critically for accuracy and depth
- **For Educators:** Incorporate AI literacy into curricula to help students use these tools effectively and ethically

Current and Future Work

Current project:

- Can training enhance the ability of GenAI to evaluate and mark scientific text
- Is the accuracy of referencing improving through dedicated referencing tools?

Future project:

- Long-term studies to evaluate the impact of AI-assisted writing on learning outcomes

Acknowledgements

Ahmed Al Samere
Mollie Ridge

OU and NCFE Evaluation Report

https://law-school.open.ac.uk/sites/law-school.open.ac.uk/files/files/OU%20NCFE%20report%20on%20GAI%20and%20assessment.pdf