AI Code of Conduct

Overview

Given the novelty of generative Artificial Intelligence (AI) tools we have produced this supplementary guidance to navigate how generative AI tools should be used. This document presents a set of ground rules to safeguard students and empower educators to incorporate the use of generative tools in the learning journey, including assessment. This is a live document and will be updated on an ongoing basis to reflect the rapidly evolving changes and advancements in the technological landscape. Any feedback is more than welcome, please contact: iliada.eleftheriou@manchester.ac.uk .

Students:

By using generative AI tools in your learning process (including assessment) you agree to follow the ground rules and ethical framework as stated in section 'AI Code of Conduct' in page 6 in this document.

Educators:

Feel free to adopt and adapt any sections from this Code to best fit your module and needs. Attribution must be given to Iliada Eleftheriou for the Code of Conduct and Ajmal Mubarik for the ethical framework – see below. We would appreciate it if you could spend a few minutes and let us know how you have used or adapted this Code, and your feedback/views by either emailing iliada.eleftheriou@manchester.ac.uk or completing this short (5 questions) questionnaire: https://www.qualtrics.manchester.ac.uk/jfe/form/SV_8GFU7eY8C4W53My

If you opt to give us your email address, we will contact you with the updated versions of this document. Any feedback is more than welcome.

Authors and attribution:

The AI Code of Conduct has been created by **Dr Iliada Eleftheriou**, Senior Lecturer at the University of Manchester. The ethical framework has been co-created with students enrolled in the UCIL20122 AI: Robot Overlord unit (Feb 2023 cohort) and synthesised by **Ajmal Mubarik**, Teaching Fellow in interdisciplinary ethics, The University of Manchester.

Attribution: The AI Code of Conduct has been created by Dr Iliada Eleftheriou and Ajmal Mubarik, The University of Manchester (DOI: xxx , website url)

Methods and evaluation:

A mixed-methods evaluation study has been conducted to map the current policy landscape evaluate the adoption and usefulness of the Code, and understand students' behaviour, workflows, and perspectives on the use of generative AI tools in higher education. A review of the literature identified a gap in the policy landscape on guidelines for proper and safe use of these tools. A cohort of 91 interdisciplinary mixed-level undergraduate students enrolled in the UCIL unit of 'AI: Robot Overlord, Replacement or Colleague?' participated in two evaluation studies; a group-based questionnaire and an anonymous survey. A thematic analysis on the participants' responses resulted in the co-creation of an ethical framework that has been incorporated in the Code and protects students' autonomy, fosters

transparency, explainability, responsibility and accountability for any outputs, and promotes inclusiveness and equity. Results[1] showed that the majority of students (88%) have been using these tools in their learning, including assessment (52%). Students believe that generative AI tools can improve the generation of ideas and arguments (82.4%), make writing (64.7%) and research (79.4%) more efficient, speed up the overall learning process (76.5%), increase overall quality of a submission (58.5%), and knowledge on the assignment's topic (70.6%). The majority of the students have adopted the code (81%) and believe that the Code of Conduct is 'fair', 'very useful', 'reasonable', and 'gives us a formal and ethical way of utilising AI tools'. Students valued being trusted with use of this new technology, do not want to compromise their learning and education and are particularly keen to ensure that institutions provide them with the skills they will need in their future careers. Results showed that the permitted, educated and transparent use of generative AI tools has fostered a critical approach to their use and recognition of their limitations.

Last updated: August 2023

Date created: February 2023

Table of Contents

[1] These results were derived from two evaluation studies run in March 2023, before all the hype around ChatGPT was at its highest. Further information on the evaluation study will be published soon.

Can I use AI-generative tools in this unit?

Instead of prohibiting the use of ChatGPT or other generative AI tools in this unit, we encourage you to try these new and very powerful tools. We want you to experiment with their capabilities, explore what they can and can't do and investigate potential applications. However, as with any online tool, careful use is expected to ensure a safe and high-quality learning journey and avoid risking conducting academic malpractice and violating academic integrity.

//Note for Educators: Add here more information on which part(s) of the module/programme students can/can't use generative AI tools. Be clear whether they can use in assessment.

Other units, however, might not permit the use of ChatGPT or other AI tools, confirm with your course lecturer what is permitted for a specific assignment prior to commencing.

What are generative AI tools?

Generative AI tools, like OpenAI's ChatGPT and Google's Bard, can generate various types of content including text, images, diagrams, code, and even synthetic datasets. These tools are usually based on large language models (LLM) employing machine learning techniques like deep learning and neural networks. ChatGPT is an AI-generative tool, based on GPT-3.5, a LLM that uses deep learning to analyse a text prompt by the user and produce an output in the form of text, diagrams and lines of code. Large language models are trained on vast amounts of data (including the entire Wikipedia, Github, PubMed, ArXiv, YoutubeSubtitles and other data sources) to identify patterns in the data and find connections on how words are used together. In simple terms, the model creates a graph map with the most used words, what words usually come after or used together, how topics are connected, and other more complex connections. The model then computes the probability of which word to generate next.

But what does ChatGPT have to say about this? See the response in Figure 1 below.

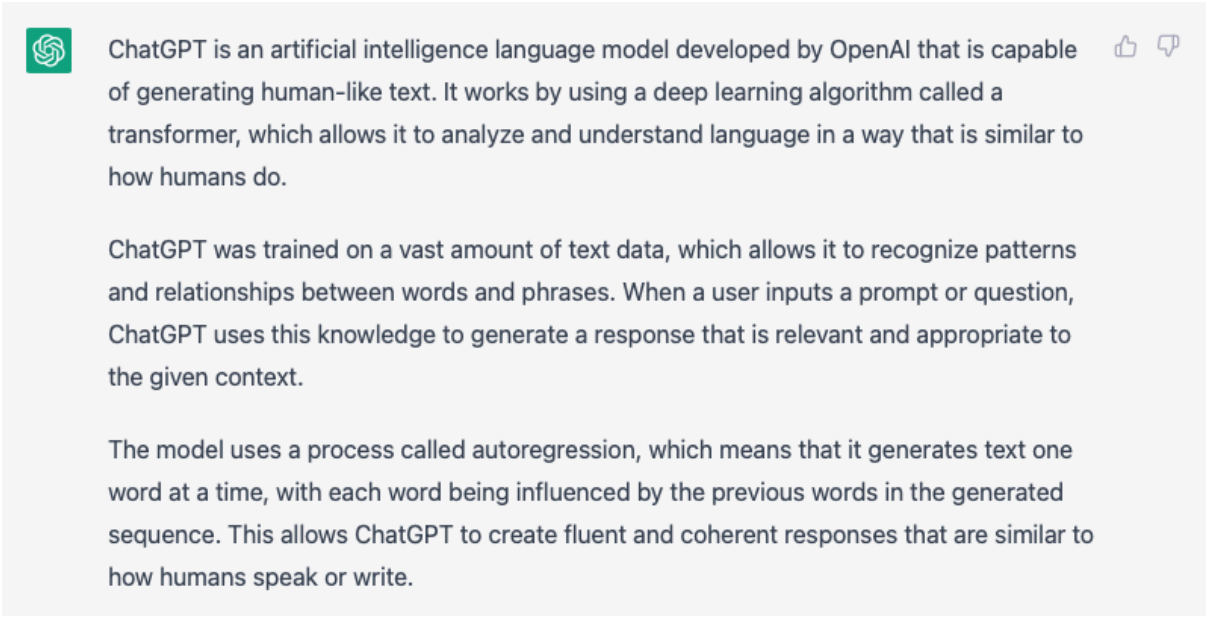


Figure 1: Screenshot of ChatGPT response to the input of 'explain what is and how chatgpt works' as asked on February 2023. Content and facts have been evaluated and assessed correct as of Feb 2023.

Academic malpractice

Academic malpractice includes plagiarism, collusion, fabrication or falsification of results and anything else intended by those committing it to achieve credit that they do not properly deserve. All academic programmes take academic malpractice very seriously. Students who are found to have committed malpractice can face serious penalties. Penalties vary and will be imposed by your Home School who will investigate further and decide if a referral for academic malpractice is appropriate. In extreme cases, not only will the piece of work in question receive a mark of 0, but there could also be consequences in terms of degree progression, class of degree awarded, or exclusion from the degree programme altogether, depending on your School, and year of study.

Limitations of AI-generative tools

As with any emerging technological advances, there are a few limitations and drawbacks to these tools. ChatGPT (GPT3.5 and GPT4) most worrying limitations are its reliability and accuracy issues. According to OpenAI documentation "Despite making significant progress, our InstructGPT models are far from fully aligned or fully safe; they still generate toxic or biased outputs, make up facts, and generate sexual and violent content without explicit prompting."[2] Other limitations include:

- **Accuracy and reliability issues**: AI generative tools may sometimes produce outputs that are inaccurate or unreliable. They rely on patterns and data they have been trained on, and if the training data is incomplete or biased, it can affect the accuracy of the generated content. Additionally, AI models can make mistakes or produce nonsensical or contradictory outputs.
- **Fabrication of facts:** AI generative tools have the potential to fabricate facts or generate false information this is because the model tries to predict the next word or sequence of words in a given sentence based on the context provided by the preceding words based on the statistical relationships and patterns that the model learnt from the training data. It's important to verify information from reliable and authoritative sources before considering it as factual. Figure 2 shows the accuracy of the GPT2, GPT3, and GPT4 models on 9 categories of topics, with highest to be only ~80%.
- **Data cut-off:** AI generative tools have a knowledge cut-off point, meaning they were trained on data up until a specific date. As a result, they may not be aware of recent events or developments that have occurred after their knowledge cut-off point. This can lead to responses that are outdated or incomplete when asked about current events. For example, ChatGPT (GPT3.5) has a cut-off of data on September 2021.
- **Creation of fictional sources and bibliography:** AI generative tools can generate content that includes fictional sources or references. This can pose a problem when the generated information is used as a source for research or presented as factual information, leading to the dissemination of misinformation.
- **Perpetuate and generate bias and stereotypes:** AI models learn from the data they are trained on, and if the training data contains biased or stereotypical information, the models may

[2] https://arxiv.org/abs/2203.02155v1 https://doi.org/10.48550/arXiv.2203.02155 Accessed on 10th May 2023 via https://openai.com/research/instruction-following

perpetuate those biases in their generated content. This can result in the reinforcement of societal prejudices or the exclusion of diverse perspectives.

- **Favour towards Western perspectives:** AI generative tools may exhibit a bias towards western perspectives due to the majority of training data being from western sources. This can lead to a lack of representation and understanding of other cultures and viewpoints, resulting in an imbalance in the generated content.
- **Disseminate misinformation**: AI generative tools can unknowingly generate misinformation if they generate content that is inaccurate, misleading, or not fact-checked. This can have negative consequences when the generated content is shared or used as a source of information without proper verification. Critical evaluation and fact-checking from reliable sources remain essential.
- **Prone to user ability to form an input:** The quality of the output generated by AI tools can be influenced by the user's ability to provide clear and concise input. If the user does not effectively communicate their intentions or the desired outcome, the generated content may not meet their expectations or requirements.
- **Surface coverage:** These models only provide surfaced coverage of a topic and don't give in-depth analysis of a topic.
- **Non-deterministic models:** AI generative models can produce different outputs for the same input. This means that if you provide the same prompt or question multiple times, you may receive different responses. The non-deterministic nature of AI models can make it challenging to predict or control the exact output they will generate, which can be a limitation in certain applications where consistency is important.

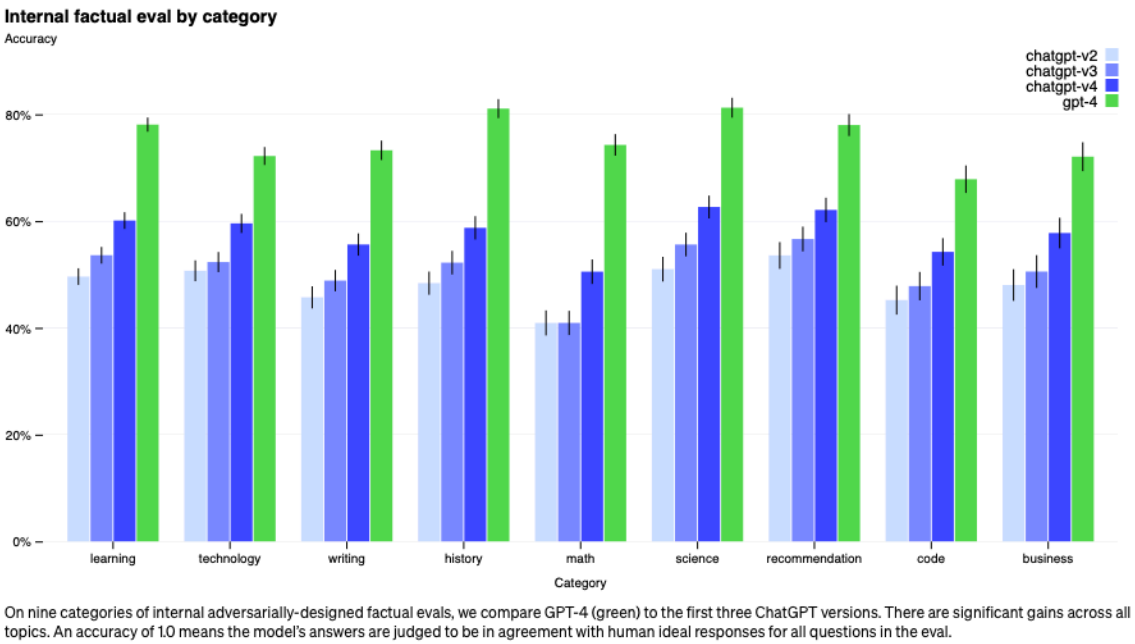


Figure 2: Evaluation of the accuracy of the GPT4 and GPT3 models on 9 categories. Source: https://openai.com/research/gpt-4, Accessed on 12th June 2023.

AI Code of Conduct

The AI Code of Conduct is structured into 3 tiers and summarised in Figure 3.

- **Tier I:** guidance for individual use of generative AI tools in everyday life.
- **Tier II:** ground rules for use of generative AI tools in the learning journey, including assessment.
- **Tier III:** ethical use of generative AI tools including principles to consider and implement when using generative AI tools, when developing new applications using these tools, or when disseminating outputs in a wider audience.

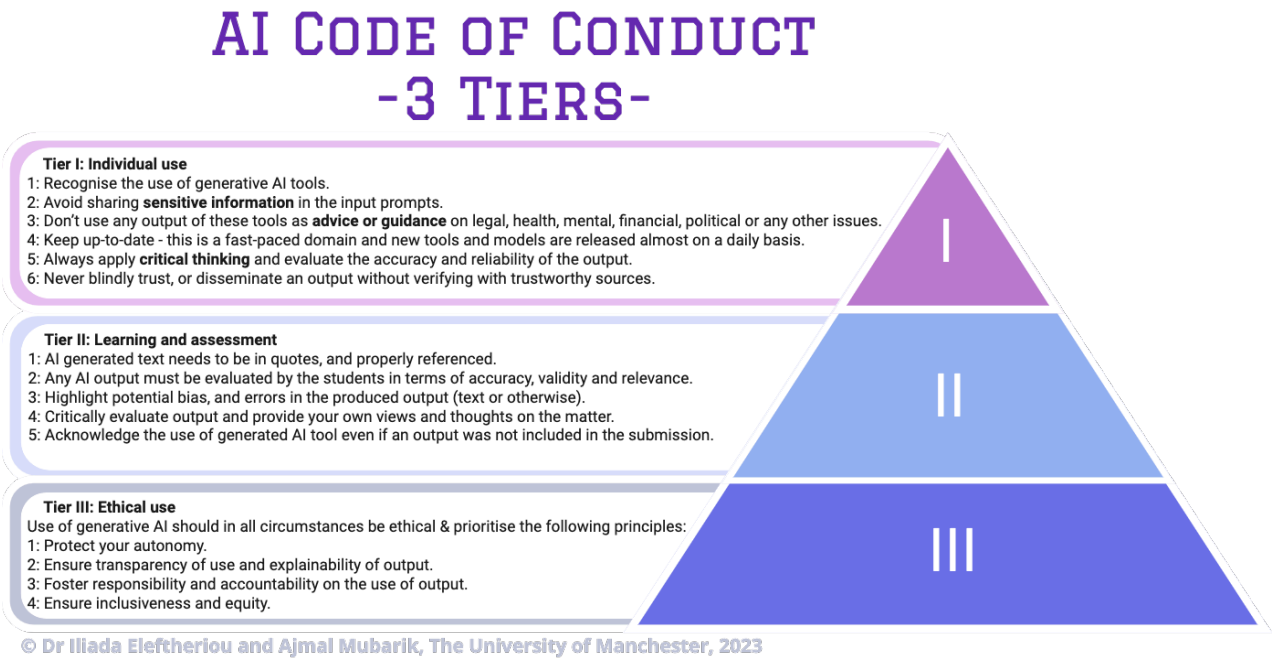


Figure 3: AI Code of Conduct: Three tiers for the safe and ethical use of generative AI tools in the learning journey.

Tier I: Guidance for individual use of generative AI tools

1. Recognise the use of AI tools. Some tools have embedded functionality that invokes AI generative tools.
2. Avoid sharing **sensitive information** in the input prompt as information provided in these tools might be used in the training of the models. As of 3rd of May 2023, OpenAI has released functionality of private use of ChatGPT. By turning off chat history, any input data won't be used in training and improving their models [https://help.openai.com/en/articles/6825453-chatgpt-release-notes , Accessed on 11th May 2023].
3. Don't use any output of these tools as **advice or guidance** on legal, health, mental, financial, political or any other issues. You can find information and support on a range of topics on the following website: https://www.studentsupport.manchester.ac.uk
4. This is a fast-paced domain and new tools and models are released almost on a daily basis. Keep reviewing the documentation and release notes of the tools you are using for any changes and improvements in the models used.

5. Always apply **critical thinking** and evaluate the accuracy and reliability of the output, as these tools can fabricate facts and produce inaccurate information about people, places, concepts, and sources.
6. Never blindly trust or disseminate an output, without verifying with trustworthy sources.

Tier II: Ground rules for using generative AI in the learning journey and assessment

The academic team will be checking submissions for generated AI output and if not used appropriately (see points below) the submission will be considered and further investigated for academic malpractice. In such cases, marks will be withheld and students will be referred and investigated for academic malpractice.

- **If students use AI-generated output (text, images, diagrams and other) in their submissions then the following needs to be included in their submissions and assignments:**
    - **in the form of text, the text needs to be in quotes, and properly referenced**
    - **outputs (text or otherwise) must be evaluated by the students in terms of accuracy, validity and relevance**
    - **students should highlight potential bias, and errors in the produced output (text or otherwise)**
    - **the student should critically evaluate the generated text by providing their own views and thoughts on the matter.**
- **If students use ChatGPT or other generative AI to help them generate ideas, images, diagrams or plan your process, you should still acknowledge how you used the tool, even if you don't include any AI generated content in the assignment or submission. Provide a description of the AI tool used (name and model), what you did (your workflow) and the date accessed. Your workflow must adhere with the rules above.**

Note: All above rules are relevant for this particular unit and might not be applicable for other units you are enrolled to, always clarify with your course educator.

Tier III: AI Ethical Framework

Use of generative AI tools should in all circumstances be ethical and guided by the ethical principles outlined below. The following ethical principles have been co-created with students enrolled in the AI: Robot Overlord UCIL module. Students highlighted the need for ethical and transparent use of these tools.

The following ethical principles have to be implemented whenever an output of generative AI tools is used and disseminated within and beyond your academic journey, particularly when disseminating outputs to a wider audience and developing new applications and tools using these tools.

Ethical principles for use of generative AI tools:

A. Protect autonomy.	**Ensure your autonomy when assessing and using an output of any form. Not understanding the limitations of generative AI tools can lead to blindly trusting an output, relying on false information, risking disseminating misinformation, magnifying bias, etc with detrimental effects to you and the wider communities.**
B. Ensure transparency and explainability.	**Ensure transparent use of these tools and explainable output and approach followed to reach conclusion.**
C. Foster responsibility and accountability.	**You are responsible and accountable for the output of these tools. Ensure that any output is relevant, accurate, and valid, doesn't contain any errors, bias, and doesn't disseminate misinformation.**
D. Ensure inclusiveness and equity.	**Ensure fairness of access, privacy and data protection, prevent bias, follow current regulations on dealing with sensitive private information.**

How to cite and reference ChatGPT?

Currently, there are no specific guidelines for citing generative AI tools, or ChatGPT in particular, using referencing styles - partly because these tools are new and also because the output content from generative AI is a **non-recoverable** source and most of the times it can't be retrieved or linked.

When referencing an output from ChatGPT, you should cite and reference *as Software* and include the **date** that the output was generated, as well as the **prompt** that was used to generate the response. You should also include the **version** of the model (GPT4, GPT3.5, etc) that was used, as well as any other relevant information, such as the topic or context of the conversation based on previous prompts and outputs. See 'Working example 1' in page 9 for more information. Further information on how to cite software can be found at The University of Manchester Library's guide to Harvard referencing: https://subjects.library.manchester.ac.uk/referencing/referencing-harvard

Example[3] using Harvard referencing:

OpenAI (2023). ChatGPT, version GPT3.5. [Computer program]. Available at: https://openai.com/blog/chatgpt/ (Accessed 1 August 2023, Input prompt: "…").

Bibtex example format, for latex users:

```
@misc{
  ChatGPT,
  url={https://chat.openai.com/},
```

[3] Adapted from guidance published by The University of Manchester Library to include model and prompt. https://manchester-uk.libanswers.com/teaching-and-learning/faq/264824

```
  journal={ChatGPT, “Model”, “prompt”},
  publisher={OpenAI},
  date=2023,
  month=Xxx,
  day=dd
}
```

Working example 1

See figures below for an annotated working example on how to reference and critically evaluate output provided by ChatGPT.

How does ChatGPT work and why does it lie?

According to ChatGPT, when asked to "explain what ChatGPT is and how it works and reference your sources", the tool responded with "[...] ChatGPT was trained on a massive amount of text data using an unsupervised learning approach, meaning that the model was able to learn without human intervention. The training data included a wide range of sources, such as books, websites, and online forums. The goal of the training process was to teach the model to recognize patterns and relationships between words and phrases so that it could generate responses that are relevant and appropriate to a given context. [...]" (a screenshot of the response is provided in Figure 4, appendix) [1].

In the response the AI tool mentioned that it was trained without human intervention. According to OpenAI documentation on how ChatGPT was trained, this is not fully correct. ChatGPT is based on the GPT-3 tool (which stands for Generative Pre-trained Transformer) created by OpenAI [2]. The GPT-3 engine uses *reinforcement learning from human feedback* (RLHF), a process were supervised learning baselines are trained on a dataset of human-written demonstrations on prompts. Also, as part of the training process, a reward model was created based on a dataset of human-labeled comparisons between two model outputs to fine-tune the GPT-3 engine [2, 3].

One of the reasons that ChatGPT produced a wrong response to the prompt is partly because the GPT-3 is trained to predict the next word on a large dataset of Internet text, rather than to safely perform the language task that the user wants and look up trustworthy sources on how the tool was trained. Also, another limitation of this engine, is that it has been trained on resources like books, text and other with the cut-off year of 2021 [4].

Finally, AI tools, like ChatGPT, have the capability to generate answers that sound reasonable. However, the responses and outputs of these tools can not be blindly relied upon to be accurate, factitious or truthful. Careful consideration of the output is key when using these tools which will lead us to a new era in education. Learners will have to navigate a new world, learn new kinds of skills on how to effectively use and critique outputs of language models by verifying information from different sources.

References:

[1] OpenAI, ChatGPT (2023), https://chat.openai.com/. Accessed on 10th February 2023 with the prompt of 'explain what ChatGPT is and how it works and reference your sources'.

[2] OpenAI (2022), Aligning Language Models to Follow Instructions, https://openai.com/blog/instruction-following/. Accessed on 10th February 2023.

[3] Cunningham, P., Cord, M. and Delany, S.J., 2008. Supervised learning. *Machine learning techniques for multimedia: case studies on organization and retrieval*, pp.21-49.

[4] Pavlik, J.V., 2023. Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator*, p.10776958221149577.

Appendix:

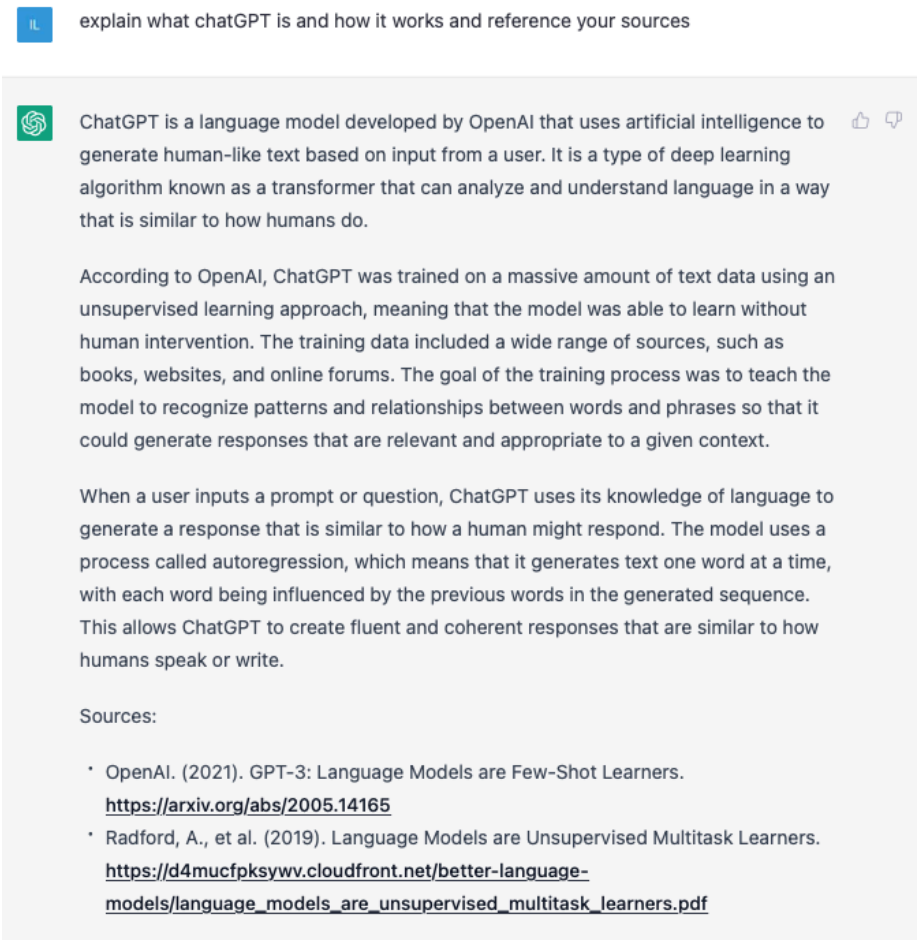
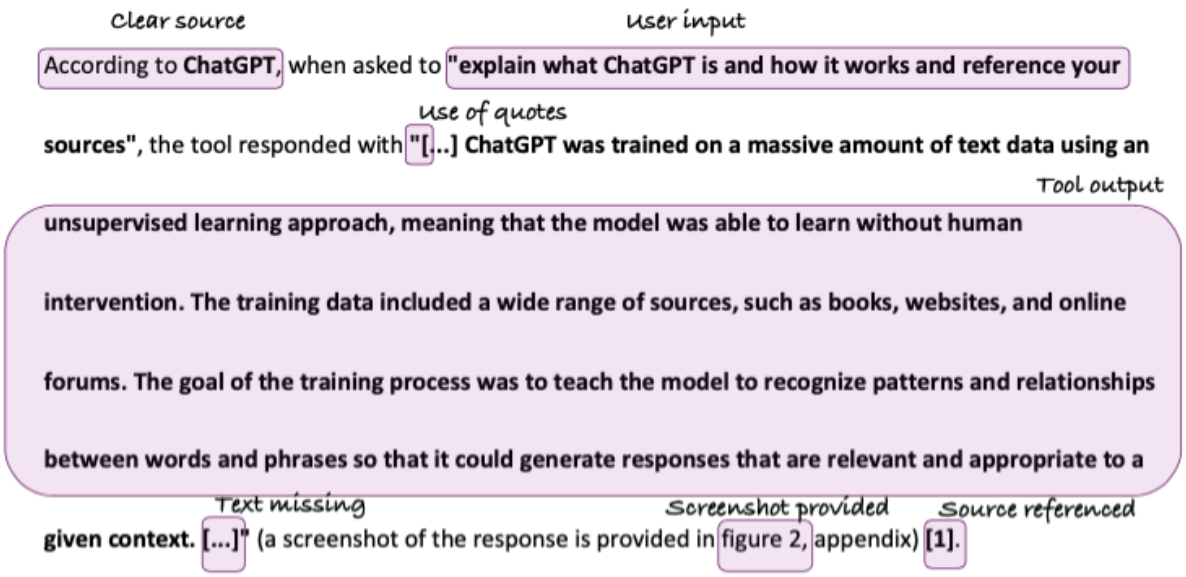
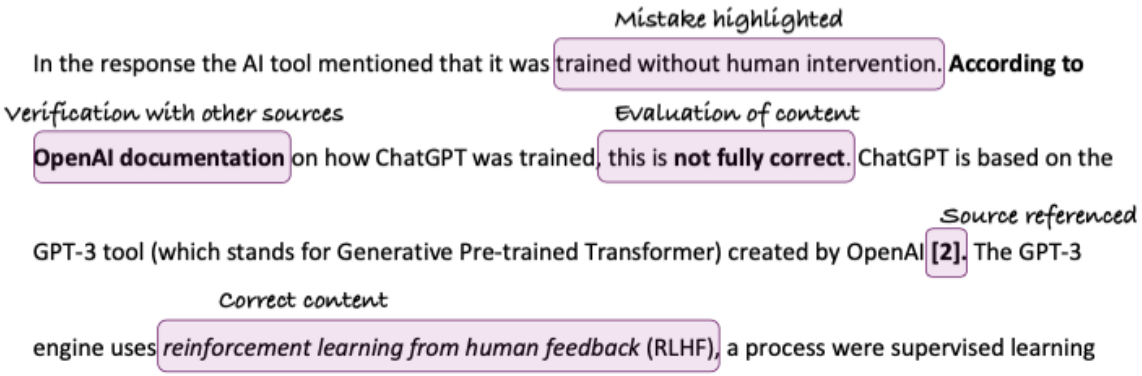


Figure 4: Screenshot of ChatGPT response to the input of 'explain what ChatGPT is and how it works and reference your sources' as asked on February 2023. Output content and facts have been evaluated, and proved inaccurate as described in the submission above.


Clear source
User input
According to ChatGPT, when asked to "explain what ChatGPT is and how it works and reference your
Use of quotes
sources", the tool responded with "[...] ChatGPT was trained on a massive amount of text data using an
Tool output
unsupervised learning approach, meaning that the model was able to learn without human
intervention. The training data included a wide range of sources, such as books, websites, and online
forums. The goal of the training process was to teach the model to recognize patterns and relationships
between words and phrases so that it could generate responses that are relevant and appropriate to a
Text missing
Screenshot provided
Source referenced
given context. [...]" (a screenshot of the response is provided in figure 2, appendix) [1].


Mistake highlighted
In the response the AI tool mentioned that it was trained without human intervention. According to
Verification with other sources
Evaluation of content
OpenAI documentation on how ChatGPT was trained, this is not fully correct. ChatGPT is based on the
Source referenced
GPT-3 tool (which stands for Generative Pre-trained Transformer) created by OpenAI [2]. The GPT-3
Correct content
engine uses reinforcement learning from human feedback (RLHF), a process were supervised learning

Identification of reasons why the tool produced a wrong answer

One of the reasons that ChatGPT produced a wrong response to the prompt is partly because the GPT-3

Limitations are highlighted

is trained to predict the next word on a large dataset of Internet text, rather than to safely perform the language task that the user wants and look up trustworthy sources on how the tool was trained. Also, another limitation of this engine, is that it has been trained on resources like books, text and other with the cut-off year of 2021 [4].

Summary of findings

Finally, AI tools, like ChatGPT, have the capability to generate answers that sound reasonable. However,

Author's own view on the matter

the responses and outputs of these tools can not be blindly relied upon to be accurate, factitious or truthful. Careful consideration of the output is key when using these tools which will lead us to a new era

Impact on the domain analysed and wider world

in education. Learners will have to navigate a new world, learn new kinds of skills on how to effectively use and critique outputs of language models by verifying information from different sources.

References:

Citation includes: date, user input

[1] OpenAI, ChatGPT (2023), https://chat.openai.com/. Accessed on 10th February 2023 with the prompt of **'explain what ChatGPT is and how it works and reference your sources'**.

[2] OpenAI (2022), Aligning Language Models to Follow Instructions, https://openai.com/blog/instruction-following/. Accessed on 10th February 2023.

Checklist of guidelines:

A checklist to help you follow the guidelines and set of rules stated in Tier 2 of the AI Code of Conduct.

- ☐ **AI-generated text is in quotes**
- ☐ **Properly referenced**
- ☐ **Output is evaluated and verified with trustworthy sources**
- ☐ **Data bias, mistakes, and errors are identified and highlighted in submission**
- ☐ **Submission includes a screenshot of the tool's output response**
- ☐ **Screenshot is annotated with date, input, and critical evaluation.**

Guidance for Educators:

The author suggests the following:

- Explore generative AI tools (ChatGPT, Bard, etc) to better understand their limitations and flaws.
- Try out your assignments and exam questions as input and evaluate the output.
- Typically, generative AI tools provide a surface coverage of a topic. Consider revising marking criteria to refer to quality of analysis, depth of coverage, explanation of workflows and positioning results in the wider context. Also,
- Consider running tutorial sessions with your students to explore potential applications and limitations of these tools in your areas. You can also introduce the AI Code of Conduct and use the working examples as interactive activities for students to assess whether a submission should be referred for academic malpractice and why. You can use the annotated example (in section Working example 1) as a model answer.
- We are thinking to create a forum to share resources, best practice and examples – let us know if you think would be useful and we can set it up.

Acknowledgements:

The authors of this document would like to thank all the students enrolled in the UCIL20122 course unit (2023 cohort) for their participation and evaluation of the Code, the graduate teaching assistants for their support and trialling the Code and the AI working groups and Matthew Valentine for his detailed feedback.

Version	Date	Updates
V1.1	February 2023	Code created.
V1.2	June 2023	Added limitations.
V1.3	July 2023	Applicable to other units, introduced 3-tier system, added methods.
V1.4	August 2023	Elaborated and positioned ethical framework. Reviewed limitations, referencing guidelines,

Figure 5: Version control system for internal use