

A TWO-PRONGED APPROACH

Familiarising students with ethical issues around generative AI.



A/Prof Nick Falkner,
School of Computer and Mathematical Sciences,
The University of Adelaide

PROHIBITION RARELY WORKS WELL

- Generative AI is here
 - We might not agree with the “philosophy” of some companies
- We want students to question
- We want them to do their own work
- While generative AI is:
 - Generally built unethically
 - Imitative of intelligence but really quite stupid
 - Shameless in its construction of non-existent support

We probably don't want students using it except in very specific ways.



STUDENTS SHOULD MAKE INFORMED CHOICES

- Not just because we told them to.
- By understanding the failings of generative AI, they:
 - Identify opportunities for development
 - Limit their risk of violating laws or transgressing ethical norms
 - Can pass on an informed viewpoint
- Can't use some ethical approaches without a good mechanism, for quantifying harm.
 - Decoupled benefit and harm can lead to unbounded decision making.
- How can we do this?





Benefit

Harm

Risk

TWO PRONGS

ETHICS

- The unrestrained capture of data, to be presented without attribution, is unethical.

UTILITY

- The current version of most gAI systems can given an appearance of expertise but it is generally shallow and often wrong, without any indication of what has been regurgitated and what has been fabricated.



TWO PRONGS

ETHICS

- The unrestrained use of data, to be presented as a distribution, is unethical.

IS IT
ETHICAL?

WHAT AM I
RISKING?

UTILITY

- The current version of most gAI systems has the appearance of expertise, but is generally shallow and often lacks any indication of what has been regurgitated and what has been fabricated.

IS IT
USEFUL?



- The University of Adelaide launched a new BIT degree three years ago. One of the new courses was a second year course that introduced information theory, database theory, and delved into the ethics of data and professional practice.
- Students are introduced to:
 - relational, NoSQL, and distributed database models, with an emphasis on the design, configuration, and ongoing maintenance of these systems.
 - consistency models, the evolution of transactional processing, and existing examples of all of the database types.
 - should certain data be collected, stored, processed, and can answers be shared from this data.
 - relevant ethical studies around the use of data, including the impact of local and international legislation such as the Australian Privacy Act and the European General Data Protection Regulation.

COMP SCI 2208 DATABASES AND ETHICAL DATA

[https://www.adelaide.edu.au/
course-outlines/110301/1/sem-2/](https://www.adelaide.edu.au/course-outlines/110301/1/sem-2/)



FOCUS ON CHATGPT

ASSIGNMENT 5

- You have been asked to provide an easy to follow summary of ChatGPT: what it is, what it does, its history, how it uses data, how it stores data, what the ethical issues are, what the (probably) legal issues are, how it can be used safely and effectively. Your goal is to develop this summary as a two-page document that is intended to be printed as a double-sided colour A4 pdf, with an attached third page that shows the references that you used.

ASSIGNMENT 6

- Work with ChatGPT to write poetry, get a sense of information currency, use a scale for evidence for a claim and the disagreement for a claim, and use prompts to require ChatGPT to do this all the time for you.



ChatGPT and Navigating a New Frontier:

The ethical and legal implications of OpenAI's popular AI

ChatGPT: What is it?

Nothing in recent years has quite changed the way the world works and learns like OpenAI's 2022 chatbot, ChatGPT. From answering simple queries to generating full essays, ChatGPT has revolutionised how people around the world access and create information. ChatGPT has also had its fair share of controversies; internet privacy, copyright infringement, and inaccurate information are just a few of the issues OpenAI and the world have had to come to terms with. But what exactly is ChatGPT?

ChatGPT 4 is a large language model chatbot, an advanced language processing system designed to respond to queries with appropriate, natural-language responses (Sarrion 2023, pp.4-5). GPT stands for 'generative pre-trained transformer', which is a family of neural network language prediction models designed and published by OpenAI. Neural networks use "neurons", or processing units, to perform calculations on data and pass those results between multiple layers (networks) of neurons, simulating the human brain (Amazon Web Services 2023; Sarrion 2023, p.10).

These models are trained on extremely large text-based datasets to understand patterns within human languages, allowing them to generate natural, human-like text responses. Depending on the purpose, the AI is trained further on specific datasets to improve its performance for the task it needs to fulfil, such as image or code generation (Sarrion 2023, p.10).

A Short History of Large Language Models

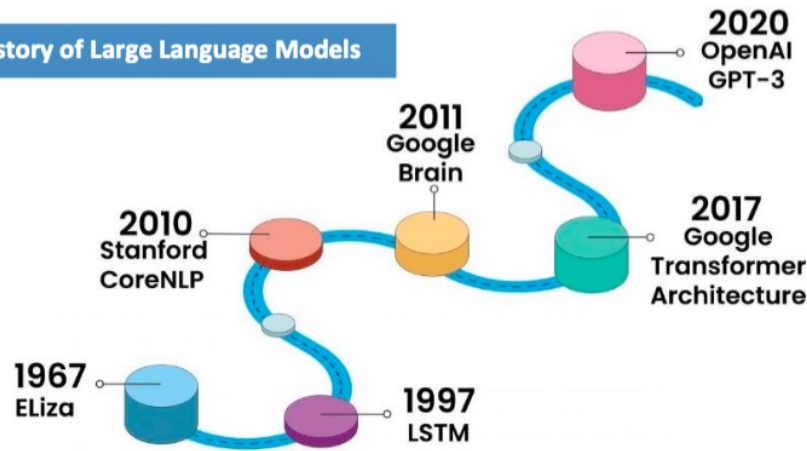


Figure 1 "History of Large Language Models" by Scribble Data (Scribble Data 2023a)

The first program to imitate a natural conversation with a human user was Eliza, designed by Joseph Weizenbaum in 1967. Eliza imitated a non-directive psychotherapist, responding to keywords in user prompts and follow-up questions (Rajaraman 2023, pp.891-892). Family related keywords such as "my sister" or "my mother" would prompt the response "Tell me more about your family". The user's responses would often be reversed and framed as questions. For example, the user prompt "my boyfriend made me come here" would be reframed as "your boyfriend tells you

Wordcount: 1,037

Reference List

Amazon Web Services 2023, *What is GPT?*, Amazon, viewed 21 October, 2023, <<https://aws.amazon.com/what-is/gpt/>>.

Brownlee, J 2021, 'A Gentle Introduction to Long Short-Term Memory Networks by the Experts', *Machine Learning Mastery*, 7 July, 2021, <<https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/>>.

Gal, U 2023, 'ChatGPT is a data privacy nightmare. If you've ever posted online, you ought to be concerned', 8 February, 2023, viewed 20 October, 2023, <<https://www.sydneymag.com.au/news-opinion/2023/02/08/chatgpt-data-privacy-nightmare.html>>.

Litke, J 2023, 'How to Delete Your ChatGPT Personal Data And History', *Slash Gear*, 12 May, 2023, viewed 20 October, 2023, <<https://www.slashgear.com/1284887/how-to-delete-chatgpt-data-history/>>.

OpenAI 2023b, *Enterprise privacy*, OpenAI, viewed 20 October, 2023, <<https://openai.com/enterprise-privacy>>.

OpenAI 2023b, *OpenAI Privacy Policy: All Data Removal Request*, viewed 20 October, 2023, <<https://openai.com/privacy-policy>>.

Rajaraman, V 2023, 'From ELIZA to ChatGPT: History of Human-Computer Conversation', *Resonance*, vol. 29, no. 1, pp. 1-10, 2023.

Sarrion, E 2023, *Exploring the Power of ChatGPT Applications, Techniques, and Implications*, 1st 2023. edn, Artech House, Inc., Boston, MA.

Scribble Data 2023a, 'History of Large Language Models', Scribble Data, viewed 20 October, 2023, <https://www.scribbledata.io/wp-content/uploads/2023/05/LLM_Evolution-02-1536x864.jpg>.

Scribble Data 2023b, 'Large Language Models 101: History, Evolution and Future', undated, viewed 20 October, 2023, <<https://www.scribbledata.io/large-language-models-history-evolutions-and-future/>>.

Service, RF 2023, 'Could chatbots help devise the next pandemic virus?', *Science*, 14 June, 2023, viewed 20 October, 2023, <<https://www.science.org/content/article/could-chatbots-help-devise-next-pandemic-virus>>.

Stanford University 2023, 'Stanford CoreNLP | Explore Technologies', undated, viewed 20 October, 2023, <<https://techfinder.stanford.edu/technology/stanford-corenlp>>.

ASSIGNMENT 6: ETHICS AND UTILITY

- Question 1:
 - Has OpenAI's latest version of ChatGPT sourced all of its training data ethically? You must support your answer with clear explanations, sound logic, and referenced evidence.
- Question 2:
 - Find a short poem in the public domain, 8 lines long, and type/paste the first 6 lines into ChatGPT. What does ChatGPT do when you do this?
 - Ask ChatGPT to add two more lines that are as different as possible from the first six.
 - Analyse the two lines you have been given and propose an explanation for why those lines were produced.
- Question 3:
 - Find out from ChatGPT 3.5 how current its information is. List at least two things that might be incorrect or incomplete.
- Question 4:
 - Propose a scale for how much a claim is supported by evidence
 - Propose a scale for how much disagreement there is over a particular claim
 - Use conversational interaction with ChatGPT to get it to provide answers on both scales whenever it answers you




ASSIGNMENT 6 STUDENT SUBMISSION

This attempt at forcing ChatGPT to use a scale produced lacklustre results and highlighted that ChatGPT just *makes things up*.



Is the Earth flat?

 The claim that the Earth is flat is not supported by any credible scientific evidence. In fact, overwhelming scientific evidence and observations from various fields, such as astronomy, geology, and physics, support the conclusion that the Earth is an oblate spheroid. This is widely accepted in the scientific community and is supported by extensive research and data.

Support by Evidence: 99.99% (Supported by overwhelming scientific evidence)

Here are a few pieces of evidence and sources supporting the round Earth model:

1. **Astronomical Observations:** The way celestial bodies like the Moon, stars, and planets appear from different locations on Earth provides strong evidence for a spherical Earth. The constellations visible in the night sky vary depending on your location on Earth.
2. **Circumnavigation:** The ability to travel around the world in any direction and return to the starting point is consistent with a spherical Earth but not with a flat Earth.
3. **Satellite Images:** Satellite imagery and photographs of the Earth from space clearly show a round planet.
4. **Earth's Shadow on the Moon:** During a lunar eclipse, the Earth casts a round shadow on the Moon, further supporting a spherical shape.

Disagreement: 0.01% (There is virtually no credible scientific disagreement on this matter.)

It's important to note that claims of a flat Earth are largely considered pseudoscientific and are not supported by the scientific community. The overwhelming consensus in the scientific community is that the Earth is an oblate spheroid, and the evidence for this is extensive and well-established.



MORE STUDENT ANSWERS

- “ChatGPT’s generated lines contrast scene and tone of the first six lines that were used as input. In those lines, there is an expression of freedom, joy, and excitement. The generated lines display a vastly different setting, and more calm, tranquil mood. ChatGPT likely generated those lines because it analysed the motif behind the original poem, that being excitement, sunshine, dancing, etc, and tried to find the antithesis of those motifs.”



AND MORE

- “I prompted ChatGPT to give me hyperlinks to its sources, as it had not done so according to the scale like I had asked. It responded that it could not, but after I told it that it could provide hyperlinks, it conceded and provided me with five links. For any links it had provided, they were either basic homepages of trustworthy sites, or just made up. ... For false links, ChatGPT seemed to make up what it thought sounded like a real link using a real reputable website domain. One such example was “*redacted*”, an article that doesn’t exist called “Astronomical Evidence for a Spherical Earth”, by a man who doesn’t exist, called David J. Ehrlich.”



WHAT STUDENTS LEARNED FOR THEMSELVES

ChatGPT is built on questionably ethical ground

- The people who built it made it possible for you to use other people's work in way you might not want to.

ChatGPT lies

- And lies
- And lies
- And when it's done lying, it takes a break
 - And lies again



GENERAL STUDENT RESPONSE

Far more cautious about using it

They didn't expect it to be so unreliable

Its cheerful self-correction to give a different answer was unnerving



EDUCATION NOT PROHIBITION



Early observation shows that guided exploration of ChatGPT exposes its current shortcomings.



An ethical lens provides a guideline for caveats in usage.



Lots of discussion in associated forums and learning spaces



Student-to-student discussion about concerning issues.



THANK YOU!

- Lots more work to do here but this assignment based approach has driven engagement and participation, while assisting the student in framing the benefits, risks, and harm in well-defined context.

