

**CHATGPT IN STEM TEACHING:
AN INTRODUCTION TO USING
LLM-BASED TOOLS IN HIGHER
ED**

JAKUB HYZYK AND MELANIE MISANCHUK



ChatGPT in STEM Teaching: An introduction to using LLM-based tools in Higher Ed Copyright © by Jakub Hyzyk and Melanie Misanchuk is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, except where otherwise noted.

CONTENTS

| | |
|--|----|
| Learning Outcomes | 1 |
| PART I. INTRODUCTION | |
| PART II. A BRIEF HISTORY OF MACHINE LEARNING AND LLMS | |
| 1940s – 1970s | 7 |
| 1980s – 1990s | 10 |
| 2000s - 2023 | 13 |
| PART III. SCOPE AND ENVIRONMENTAL SCAN | |
| Scope | 21 |
| Environmental Scan | 23 |
| PART IV. HOW LLM TECHNOLOGY WORKS | |
| PART V. LIMITATIONS OF LLMS | |
| PART VI. BIAS | |
| Datasets | 39 |
| Types of Bias | 42 |
| Mitigating Bias | 53 |
| Human Feedback Leading to Erasure of Marginalized Groups | 55 |
| PART VII. ACADEMIC INTEGRITY | |
| What Is a Poor Overworked Instructor to Do? | 59 |
| General Guidelines | 61 |
| AI Detectors | 63 |

PART VIII. USING LLMS FOR TEACHING AND LEARNING

| | |
|---------------------------|----|
| Instructor Use of ChatGPT | 67 |
| Student Use of ChatGPT | 79 |

PART IX. WHAT TO DO ABOUT ASSESSMENT?

| | |
|--------------------------------------|-----|
| Sample Assessment Description | 85 |
| So What Can Educators Do? | 88 |
| Suggested Assessment Types | 92 |
| Inner Feedback...With ChatGPT's Help | 99 |
| Low Stakes Activities Using ChatGPT | 104 |
| Conclusion | 114 |

PART X. FINAL THOUGHTS

| | |
|------------------|-----|
| Works Cited | 117 |
| Acknowledgements | 123 |

LEARNING OUTCOMES

After working through this resource, instructors will be able to:

- Speak generally about the history of generative AI and LLMs
- Define keywords and general terms used in the industry
- Apply their understanding of the training and development of LLM tools to the implications for layperson use
- Speak generally about currently available LLM-based tools (both free and paid)
- Discuss the technical limitations of current LLM-based tools
- Describe the types of bias inherent in LLM-based tools
- Provide mitigation efforts for bias in these tools
- Describe potential uses for LLM-based tools in teaching and learning science at the post-secondary level
- Discuss considerations when using LLM-based tools in teaching (e.g., reliability, bias, privacy and security, accessibility, equity, etc.)
- Integrate GenAI use into low-stakes activities and assessments
- Improve the “AI-immunity” of assessments
- Participate in inter- and intra-institutional committees and collegial governance on the role of generative AI in post-secondary teaching



Media Attributions

- [Firefly-Using-artificial-intelligence-in-teaching-STEM-63291-3-1-233x300-1](#)

PART I

INTRODUCTION

By now, it is likely that everyone working in a STEM field has heard of recent developments in artificial intelligence generally, or of ChatGPT in particular. For over a year, the media has been saturated with coverage ranging from fawning over the revolutionary capabilities purportedly now available to everyone, to hand-wringing over the potential obsolescence of all clerical and knowledge workers.

In the post-secondary environment, you may have heard about Generative AI from colleagues concerned about the implications for academic integrity or excited about the pedagogical opportunities presented (or both), or even from enthusiastic IT administrators struggling to balance excitement with their training and knowledge of technology hype cycles.

This OER does not aim to instill either excitement nor anxiety in educators. Rather, we endeavour to provide a balanced and sober view of Large Language Model (LLM) technology and the chatbots it powers. We offer:

- a basic historical background,
- an understanding of the operation, functionalities, and limitations of LLM-based tools,
- some useful analogies for thinking about them, and ultimately,
- an appreciation of the specific issues and opportunities introduced for higher education STEM teaching.

How Widespread Could This Really Get, Anyway?

It's reasonable to ask whether the amount of ink spilled over the potential impacts of LLMs on education—positive or negative—is warranted. It is, after all, in the interests of the companies selling these tools to generate as much interest and hype as possible about the “AI revolution” they are shepherding. This serves to maximize both sales and buy-in from public investors and venture capitalists. We have seen the leaders of these companies publicly call for regulation of the technology and warn of its potential dangers, which a sympathetic observer might interpret as good corporate citizenship. A less charitable observer might point out that the loudest calls for a “pause” come from the early market leaders, and that they conspicuously refer to future developments, rather than to their products already on the market. Keeping the potential of the technology in the news cycle— while imposing greater regulatory burden on competitors playing catch-up—would certainly be a shrewd business strategy.

So beyond what's in the popular press, what do we really know about adoption? How much use are students making of this technology, how is it impacting higher education today, and what might we expect in the near future?

The Data

According to the *2023 Global Student Survey*, 40% of university students world-wide are using GenAI in their studies, but this number is higher in Canada (where 54% of students use it) and substantially lower in the United States, with only 20% of university students using GenAI. Worldwide, 50% of the students using GenAI in their studies input a question once or more per day. In Canada, 65% of students are using it from 2 to 10 times a day and 80% of students who use GenAI are using it more than once per day (only Turkish students are using GenAI more than Canadians, at 81%.) Forty-four percent of students worldwide (but only 31% in Canada and 39% in the US) say that their understanding of complex concepts or subjects has improved since using GenAI for their studies. When asked if they would like their curriculum to include training in AI tools relevant to their future career, 83% of Kenyan and Indian students agreed. The world average was 65%; 63% of Canadian students (but only 47% of American students) would like training in AI tools (Chegg.org, 2023).

The Anecdotes

This Vox video, produced by journalist Joss Fong, provides some additional data and anecdotal colour on ChatGPT use, along with a concise and engaging introduction to some of the issues, from the point of view of both instructors and students:



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/llmtoolsforstementeachinginhighered/?p=19#oembed-1>

These statistics are primarily focused on ChatGPT, which up until recently, has been the most widely available and easily accessible tool. However, Microsoft is now rolling out the same GPT4 technology powering ChatGPT across its entire suite: in the Edge browser and Bing search engine, built into all Microsoft Office 365 tools via Copilot, and even embedded within new versions of Windows itself. These products are widely used (often, the only institutionally supported option) across post-secondary institutions in Canada, and if past experience is any guide, it's likely that many IT departments will simply "turn on" these features as soon as they are released, to staff, faculty, *and* students. It's safe to say that by the time you are reading this, a full-featured ChatGPT equivalent will likely be just a click or sidebar expand away in Word, PowerPoint, Windows, and other tools that you —and your students —use every day.

PART I

A BRIEF HISTORY OF MACHINE LEARNING AND LLMS

What follows is a chronicle of key events in the evolution of the field of artificial intelligence, broadly defined. It is not meant to be a definitive history, and is focused more on milestones and impacts than evolution of algorithms and computational approaches. Readers with an interest in the evolution of technical aspects may wish to consult IBM's Developer pages: [A beginner's guide to artificial intelligence and machine learning](#). For a deeper dive into technical details, readers may follow the links to research papers referenced in this section.

1940S - 1970S

The First Wave of AI and the First AI Winter

The seeds of the current AI revolution were sown in the earliest days of electronic computing. In 1948, Claude Shannon, the “father of information theory,” laid the groundwork for probabilistic language analysis in “A Mathematical Theory of Communications.” His exploration of predicting the next letter in a sequence foreshadowed the statistical underpinnings of modern natural language processing (NLP).

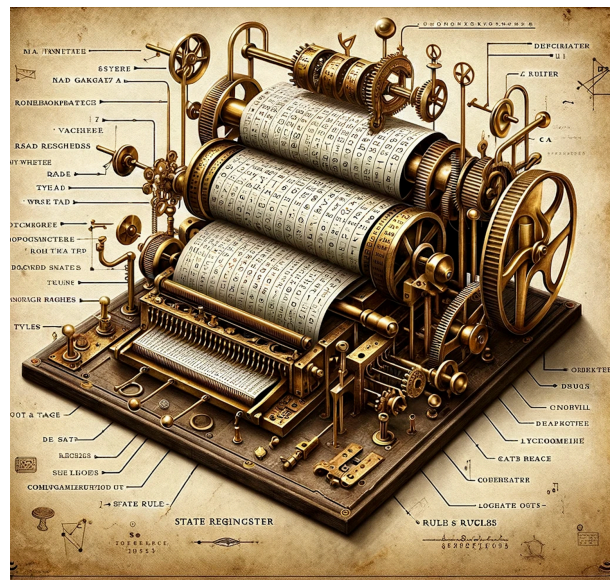
Alan Turing’s 1950 paper, “Computing Machinery and Intelligence,” set the first major goalpost for the field. He proposed the Turing Test, an interrogation game designed to determine whether the subject was human or machine. This simple challenge set a high bar that would preoccupy, and often frustrate, researchers for decades to come.

The first artificial neural network, the SNARC, was built in 1951 by Marvin Minsky. It used reinforcement learning to simulate rats spawning at various locations in a maze and finding a path through it.

In 1956, the Dartmouth Summer Research Project on Artificial Intelligence was convened. This project brought together many leading thinkers from the diverse fields of computer science, linguistics, and philosophy with an ambitious mandate:

“The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.” (p. 2, McCarthy et al., 1955)

Among these luminaries was Arthur Samuel, who would later coin the term “machine learning” in his 1959 paper detailing experiments programming an early IBM mainframe computer to learn the game of checkers. His checkers program, employing clever tree-search optimizations and an early form of reinforcement learning, honed its strategy with every game played, and led him to conclude:



“As a result of these experiments one can say with some certainty that it is now possible to devise learning schemes which will greatly outperform an average person and that such learning schemes may eventually be economically feasible as applied to real-life problems”
(p.223, Samuel, 1959)

Readers with an interest in how Chomsky’s work, and Linguistics more generally, are relevant to the modern LLMs may enjoy reading “Just how much linguistics do we need to know?”

In 1957, Noam Chomsky released *Syntactic Structures*, a book that lays out a “system of phase-structure grammar” which proposed a systematic way to describe the syntax of language. Chomsky’s work provided a theoretical framework for breaking chunks of language (e.g., sentences) into functional parts (verbs, nouns, adjectives, etc.) and indicating relationships between them, that would influence natural language processing (NLP) techniques for decades to come.

In 1958, Frank Rosenblatt developed the Perceptron, another early form of neural network that could classify simple patterns. The Perceptron was widely covered in popular science press, bringing the concept of machine learning to a mass audience (Rosenblatt, 1958).

The early 1960s were characterized by exuberant optimism for the potential of AI. The publication of a report titled “Research on Mechanical Translation” by a Congressional committee on Science and Astronautics legitimized and gave an official stamp of approval to further funding for an important subset of the field.

In 1961, Marvin Minsky published his landmark paper “Steps Toward Artificial Intelligence,” in which he performed a rigorous narrative review of the various lines of research in the field, and their relation to each other.



The first chatbot, ELIZA, was created by Joseph Weizenbaum at the MIT Artificial Intelligence Laboratory in the 1960s. ELIZA’s conversational prowess relied on simple algorithmic trickery ; the program would find keywords in a user’s statement and reflect them back as questions or say “Tell me more....” ELIZA tantalized the public and researchers alike with the illusion of understanding, which troubled Weizenbaum: back in the 1960s, he was already pondering “the broader implications of machines that could effectively mimic a sense of human understanding” (Hall, 2019).

Despite growing funding and public enthusiasm, among some theoreticians there was increasing skepticism that key technical hurdles were solvable at all, and some of the early experimental successes proved more difficult to build on than had initially been hoped. Perhaps the best summary of

this sentiment came from philosopher Hubert Dreyfus in his paper “Alchemy and Artificial Intelligence”:

“An overall pattern is taking shape: an early, dramatic success based on the easy performance of simple tasks, or low-quality work on complex tasks, and then diminishing returns, disenchantment, and, in some cases, pessimism. The pattern is not caused by too much being demanded too soon by eager or skeptical outsiders. The failure to produce is measured solely against the expectations of those working in the field.” (p. 16, Dreyfus, 1965)

To assess the state of NLP research, the US National Research Council formed the ALPAC committee in 1964. The final ALPAC report poured cold water on much of the early exuberance for AI research. Its skeptical tone stressed the need for foundational breakthroughs in computational linguistics before practical NLP applications would become a reality, and recommended decreasing or reallocating funding for research in machine translation.

Overall, the late 60’s and early 70’s were characterised by a lack of either positive developments or experimental successes, weighed against an increasing number of experimental failures and theoretical critiques of past work. Minsky and Papert’s 1969 book *Perceptrons* demonstrated major limitations to the model that had captured public imagination just 10 years earlier. By 1973, sentiment was decidedly negative, as evinced in the Lighthill Report commissioned by the Science Research Council in the UK, which contained some rather pessimistic assessments of the field’s progress:

“Most workers in AI research and in related fields confess to a pronounced feeling of disappointment in what has been achieved in the past twenty-five years. Workers entered the field around 1950, and even around 1960, with high hopes that are very far from having been realised in 1972. In no part of the field have the discoveries made so far produced the major impact that was then promised.” (p. 8, Lighthill, 1972)

As a result of this perceived slow progress, lack of good news, and negative official assessments, the flow of research funding for AI slowed substantially. Researchers disagree about the exact start and end dates, but this period –from roughly mid-60s to late 70s – would come to be known as the “First AI Winter.”



1980S – 1990S

The Second Wave of AI and the Second AI Winter

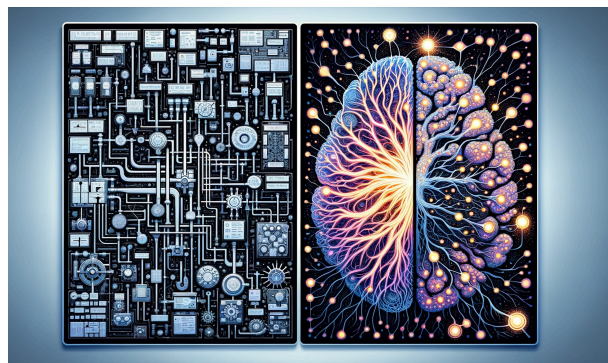
During the 1980s, progress continued on neural network-based AI models. IBM developed new statistical models infused with nascent machine learning capabilities. These systems could now make decisions not through rigid rules, but by discerning probabilistic patterns within vast datasets. Interestingly, because the first application of IBM's models was translating French to English, the datasets used were the official (bilingual) records (Hansards) of the 36th Canadian Parliament (Collins, n.d.).

Meanwhile, John Hopfield made strides in understanding the computational nature of memory. In his 1982 paper, Hopfield described a model of recurrent neural network (RNN), present in biological systems, but applicable to engineered systems, that could learn and recall complex patterns.

And in 1997, Sepp Hochreiter and Jürgen Schmidhuber introduced the idea of long short-term memory (LSTM), which utilize RNN models. These networks, a refinement of RNNs, had performance advantages over, and addressed a persistent weakness in, their predecessors: the inability to handle long-range dependencies within language.

But during this same period, a new model of AI began to gain prominence. Around the start of the first AI winter, Edward Feigenbaum had introduced the first expert system, DENDRAL. Expert systems took a different approach than previous models, both in goals and architecture. Rather than attempting to mimic (simplified) biological brain function and dynamic learning using neural networks, expert systems sought to understand and mimic static expertise. To do this, they used a model which applied a system of heuristic rules (expert thinking) to a highly detailed and specialized knowledgebase (expert knowledge). These models were not trying to learn or grow, they were trying to provide specialized information and decisions from a fixed set of knowledge, in a repeatable way. To put it another way, an expert system was like a meticulously crafted algorithm or flowchart, guiding a machine step by step through a process to an output. A neural network was like a brain that could learn through trial, error, and reinforcement — often without being able to “retrace its steps” from input to output, the way an expert system could.

As Joshua Lederberg (Feigenbaum's collaborator on DENDRAL) would later reflect:



“...we were trying to invent AI, and in the process discovered an expert system. This shift of paradigm, ‘that Knowledge IS Power’ was explicated in our 1971 paper, and has been the banner of the knowledge-based-system movement within AI research from that moment.”(p. 12, Lederberg, 1987)

Expert systems were an important shift of focus for AI research and were able to show success in specialized areas of expertise. This success began to draw research attention and a resurgence of funding from agencies like DARPA and government programs like the Strategic Computing Initiative.

Expert systems did, however, have several drawbacks. The prevalent high level programming language used to create expert systems, LISP, was resource intensive, and as more ambitious models were developed, dedicated hardware was designed to run LISP efficiently. A small industry developed in the design and manufacture of dedicated workstations and minicomputers known as LISP Machines. In addition, as the models became more ambitious, the creation of the knowledgebases and heuristic rulesets became more and more labour intensive, and the labour required was highly specialized expertise. And not all researchers in the field were convinced expert systems were the best way forward. Notably, they came under criticism from John McCarthy, a pioneer in the field of AI and one of the conveners of the Dartmouth Summer Research Project discussed earlier. In his 1984 paper *Some Expert Systems Need Common Sense*, he argued that expert systems would never have enough common-sense knowledge or common-sense reasoning ability to be suitable for many applications.

These two factors led to an eventual disillusionment of policy makers in the promise of expert systems to solve broader problems. As an example of this rising skepticism, the US Military’s 1986 SDI Large-Scale System Technology Study evaluated expert systems as part of the SDI (colloquially known as the “Star Wars” anti-ballistic missile defense system), and gave this (by now familiar) commentary:

“Finally, AI researchers must guard against an excess of technical hubris induced by self-generated hype. AI researchers have identified a number of exceedingly difficult problems that form the basis of the field. In most cases, relatively minute inroads have been made in the solution of these problems. For example, AI Systems can represent and draw conclusions in relatively simple situations and solve relatively simple problems. The promise offered by those inroads (in, say, expert systems) has been so great as to distort completely, in many cases, the perspective that ought to be maintained. As a result AI has taken on magical attributes, and the expectations of customers are exoatmospheric. It is critical to the orderly advancement of the field to maintain realistic expectations; that is, that the potential contribution of AI to solving difficult, complex problems is quite high, but significant effort remains before the potential will come to fruition.” (p. 6-10, System Development Corporation, 1986)



Such overhyped promises and exaggerated capabilities led to disappointment once again, causing a reduction of funding in some cases, and a reallocation of funding to neural network-based models in others. At the same time, rapid advances in microcomputer performance and widespread availability began to seriously call into question the need for expensive specialized hardware like the LISP Machines. The rapid decline of the industry manufacturing them further contributed to the decline of capital and talent available to the field.

These developments together led to another period of slowed progress, reduced funding, and lower interest and expectations in AI that would be known as the “Second AI Winter”.

Media Attributions

- This image was created using DALL-E
- This image was created using DALL-E

2000S - 2023

Promise Realized and Mass Adoption

Since the turn of the millennium, development in AI and machine learning has accelerated rapidly. AI systems have seen huge increases in capability, driven by better models, advances in computer hardware, and huge investment of capital from the largest companies in the world. Some of the lofty targets in the early days of AI began to be realized, and with each new milestone, progress seems only to accelerate. Machine learning and AI have been incorporated “under the hood” of virtually every major software platform and new technology. But beyond that, conversational AIs such as chatbots and assistants have brought the technologies to a mass audience in a very visible way, making AI a tangible technology that many people interact with consciously on a daily basis.

We will outline these developments in three broad categories:

- capability milestones,
- technical milestones, and
- mass adoption.

Leaps in Capability Capture Public Imagination

In 1997, IBM’s “super-computer” Deep Blue defeated then-world chess champion Gary Kasparov in a full chess match under tournament rules, after having won a game (but losing the match) in the previous year. The win captured the public imagination and gave some of the more skeptical AI researchers pause for thought, as it had at times been claimed AI would never be able to beat a top player outside of controlled circumstances.

Then in 2011, IBM’s Watson DeepQA computer defeated television game show Jeopardy! champions Ben Rutter and Ken Jennings on live TV. Watson was a question answering computer that could deploy a variety of algorithms in parallel to perform natural language processing in real time, generate hypotheses, and validate them against a knowledgebase to come up with answers and state its confidence level in them.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/llmtoolsforstemteachinginhigherred/?p=25#oembed-1>

Google acquired DeepMind Technologies, a firm specializing in generalized neural network models for playing video games in 2014. DeepMind had already found success teaching their models to perform at superhuman levels in early arcade games. The following short documentary gives some insight into how DeepMind's early models learned the games, and the company's culture in those early days:



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/llmtoolsforstemteachinginhighered/?p=25#oembed-2>

DeepMind then moved on to more complicated and modern games, but also more importantly, classic strategy games.

DeepMind's AlphaGo (Lee) , a version of their model trained to play Go, defeated Lee Sedol , a highly rated South Korean Go champion 4–1 in March 2016. This was seen as a major progression in AI capability, as Go is a game with more permutations of possible moves and is generally considered to allow for more creativity in play than chess. This made it theoretically less suitable to brute force strategies than chess, and thus harder to design effective AI for. Then in May 2017, Google DeepMind's AlphaGo (Master) defeated Ke Jie, the then-top ranked Go player in the world for two years running, in 3 straight matches.

The DeepMind team announced a new version of AlphaGo, called "Zero," with a Nature paper in October of 2017. AlphaGo (Zero) was another step forward, this time in terms of architecture and training. Previous versions of AlphaGo had been trained using reinforcement learning on historical games and by playing against humans. AlphaGo (Zero) was trained without any knowledge of the game, historical matches, or observing any humans playing at all. It learned to play Go entirely by unsupervised reinforcement learning, playing against itself. Without playing against or observing any human play, it was able to consistently defeat the previous versions of AlphaGo, Lee and Master, which had been trained on and against human play. It was able to attain this level of mastery much faster, and with less processing power, than those prior models.

Readers looking for more information about Go, and the AlphaGo story, may enjoy the award-winning documentary of the same title:



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/llmtoolsforstemteachinginhighered/?p=25#oembed-3>

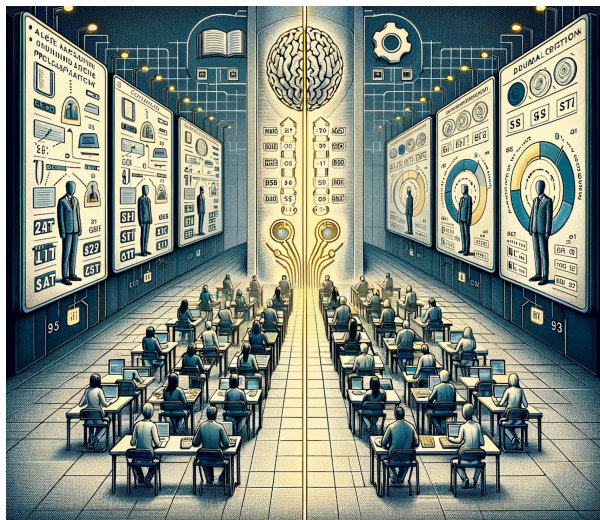
DeepMind continued creating models for various other games, notably Chess and Starcraft II, but also turned its attention to other fields, finding success in areas such as protein folding, voice synthesis, and even programming.

While superhuman feats in gaming captured much press and public attention, it was performance in natural language tasks that brought public awareness of AI into the mainstream.

In 2018, AI models from both Microsoft and Alibaba outscored the average of a large sample of human respondents in Stanford's SQuAD 1.1 test of reading comprehension. This was a milestone for natural language processing, and since then, many more models have surpassed human performance on a variety of SQuAD variants.

Since the release of ChatGPT in 2022, it has become the norm to list various academic and professional tests and certifications each iteration of a model or chatbot can pass, and how it performs relative to human students/workers. Readers have by now likely seen a great many such reports, but a few for ChatGPT can be found here: [What Exams Has ChatGPT Passed?](#) We will discuss the mass adoption of chatbots further in a subsequent section.

Research Advances and Technical Milestones



To mark the anniversary of the Dartmouth Summer Research Project, Dartmouth College hosted the The Dartmouth Artificial Intelligence Conference: The Next 50 Years (AI@50) in 2006. The conference featured presentations by veterans of the first project 50 years earlier, young researchers, and even futurists and popularizers.

In 2013, a group of Google researchers led by Tomáš Mikolov published Word2vec, a natural language processing technique for encoding the meaning and syntax of a word into vectors, which can then be mathematically evaluated against other words.

Ian Goodfellow created generative adversarial networks (GANs) in 2014. GANs are a machine learning method in which two different neural networks are given the same training data and then made to “compete” with each other, with their output being submitted to a third “discriminator” network for relative scoring. GANs allowed for competitive training of neural network models against each other without human supervision (although they have proven useful with supervision as well).

The attention model introduced by Dzmitry Bahdanau and his team in 2015 was a major step forward for natural language processing. Neural networks no longer needed to retain the entirety of a sentence in their memory, they could now pinpoint relevant words, improving accuracy and handling longer, more complex sentences efficiently.

In 2017, a team of Google researchers led by Ashish Vaswani proposed a new simple network architecture, the Transformer, based solely on attention mechanisms and doing away with recurrent neural networks. The architecture their paper introduced marked a pivotal inflection point and led directly to the current crop of transformer based LLMs and chatbots. We will discuss transformers in greater depth in the “How LLM Technology Works” section.

Following on the release of the transformer paper, several important LLMs were released, including BERT from Google, ELMo, and ULMFiT. But it was OpenAI’s GPT models that would drive advancement from 2018 onward.

OpenAI Models

Alec Radford and his colleagues at OpenAI made waves in the AI community with their generative pre-training (GPT) model. They demonstrated the power of training a language model without the constraints of explicit supervision, on a vast and diverse dataset. This first version, known as GPT-1, had 117 million parameters and was trained on BooksCorpus, a dataset of 7000 books.

OpenAI released its GPT-2 model in 2019, trained on a larger corpus of data (8 million web pages), and with a larger parameter set (1.5 billion). GPT-2 also featured several algorithmic improvements over GPT-1.

GPT-3 was released in 2020 with the paper “Language Models are Few-Shot Learners.” GPT-3 featured 175 billion parameters, and the introduction of few-shot learning. GPT-3 can largely be thought of as a scaled-up GPT-2, with some small architectural changes, and showcased near-human-level performance in many language tasks.

After GPT-3, OpenAI released GPT-3.5 in 2022 as an interim update. GPT-3.5 featured fewer parameters than GPT-3, and focused on giving more helpful, less biased responses. This was accomplished using careful Fine Tuning and Reinforcement Learning from Human Feedback (RHLF), both of which will be discussed further in the “How LLM Technology Works” section.





GPT-4, introduced in 2023, represented a further significant improvement. Though OpenAI has released less and less information about its models over time as it commercialized them, GPT-4 was rumoured to feature anywhere from 1-100 trillion parameters, as well as advancements in model architecture, training techniques, and a broader dataset. GPT-4 was also the first iteration that was fully multi-modal, allowing for input and generation of text as well as images. It could also interact with external tools and interfaces through a plugin architecture.

For more information on the technical evolution of OpenAI's models, readers may consult this article.

Conversational AI & Chatbots Bring AI to the

Masses as Consumer Products

The first interaction many non-enthusiast users had with a conversational AI came in 2011, when Apple made Siri a centerpiece of its iOS operating system and marketing for the new iPhone 4s, having acquired the app in 2010. The digital voice assistant used predefined commands to perform actions and answer questions. Siri was followed in 2012 by Google Now and Microsoft's Cortana in 2014. These digital assistants brought conversational AI to millions of non-technical smartphone and computer users, impressing with their ability to use a natural language interface and deep OS integrations to accomplish many tasks.

But it was the arrival in of ChatGPT in November of 2022 that took mass adoption to another level. ChatGPT extended the conversational abilities of the digital assistants with rich generative capabilities. Its ability to converse, explain, and generate human-quality text with surprising fluency surprised both critics and enthusiasts. OpenAI's tool had a million users within 5 days, and 100 million within 2 months, making it the most rapidly adopted consumer application in history.

Microsoft integrated ChatGPT technology into Bing search in February 2023, marking the first deployment of an LLM chatbot at scale by one of the "big five" consumer-facing software companies (Facebook, Apple, Amazon, Google, Microsoft). It also marked the beginning of a race among those players for market share in this new segment of personal computing. All of these companies are making investments and acquisitions in the space to retain optionality at a minimum or vie for dominance at a maximum.

In March 2023, OpenAI released its improved GPT-4 model,; available immediately in it's paid "plus" tier of service, while the free service continued to use GPT-3.5. GPT-4 was also integrated in Bing Chat, possibly even before the public release.

Simultaneously, Google opened early access to its own LLM chatbot, Bard. In April 2023, Google consolidated and redoubled their AI efforts, absorbing DeepMind fully into the company, and merging

it with the Google Brain research team. In May 2023, Google introduced a new, more advanced language model, PaLM 2, and incorporated it into Bard.

In November of 2023, Elon Musk announced Grok, a chatbot integrated into X (formerly Twitter) that has a focus on free speech and his particular sense of humour.

Most recently in December of 2023, Google introduced Gemini 1.0 Ultra, a rebranded and updated version of Bard, with full multimodal input and output capabilities, along with two lower tier Gemini models, and announced a paid service tier called Gemini Advanced to launch sometime in 2024.

Media Attributions

- This image was created using DALL·E
- This image was created using DALL·E
- This image was created using DALL·E

PART II

SCOPE AND ENVIRONMENTAL SCAN

SCOPE

Research for this OER was conducted in the fall of 2023. As such, some tools and features were not available to us or were not appropriate to test (for example, Bard was not available in Canada; Gemini did not yet exist). This means that offerings and feature sets are representative of that time. As the landscape on this topic is rapidly evolving, we hope to update this OER periodically.

Large language models, as a category, are covered in depth. Specific tools were used for hands-on testing and validation (ChatGPT, Bing Copilot, AITutorPro/AITeachingAssistantPro):

- ChatGPT, as the most widely known and category-leading chatbot;
- Bing chat/Copilot, as the most widely distributed free chatbot, the only one currently with RAG, and the one most likely to get wide higher-ed adoption in Ontario due to existing Microsoft IT contractual relationships; and
- AITutorPro/AITeachingAssistantPro (Contact North) as a representative chatbot designed for education in a Canadian setting.

Other chatbots are surveyed in broad terms, but not tested or treated extensively. Other forms of GenAI such as image recognition and generation will be discussed where they are helpful to explain or understand GenAI broadly. All of the LLMs used in our testing and as examples here are under rapid, intensive development. The field of GenAI is constantly evolving, in terms of capabilities of the models, new entrants, and extensibility through/connectivity with other tools. Because of this, we will where possible try to discuss in general terms, focusing on commonalities and characteristics of the technology that are likely to persist, and encouraging mental models and analogs of the technology geared to longevity.



This OER fills a gap in available, approachable ready-to-implement material. There are many articles about Generative AI and LLM-based tools in both the popular and academic press, but most concern only one or two aspects of the tools in depth or address a number of topics very generally. This OER endeavors to provide one-stop shopping for Canadian post-secondary STEM instructors who need to know the benefits and cautions of using LLM-based tools in their classes. The sections on bias, privacy and security, intellectual property and copyright, academic integrity, good pedagogical practices, and assessment design are important lenses through which to examine tools (and strategies to employ them) which are widely promoted but currently poorly understood.

ENVIRONMENTAL SCAN

We present here an environmental scan of the major LLM chatbots Ontario educators are likely to come across. This scan is not meant to be exhaustive, nor to provide a taxonomy or evolutionary record of the LLMs “in the wild.” As models proliferate, are rebranded/ revised, or in the case of open-source models, forked and customized, it will become more and more difficult to keep track of them all. Readers interested in a more comprehensive survey of LLMs, their characteristics, and their relationships to each other can consult “A Survey of Large Language Models.”

| Chatbot Name | Based On | Developer | General Capabilities and Notable Features | Web Interface URL |
|-------------------|------------------------|--------------------|--|---|
| ChatGPT | GPT-3.5+ | OpenAI | <ul style="list-style-type: none"> • Conversational AI • Multimodal content generation • Extensibility to non-LLM tasks (eg math, flight booking) via “plugins” | OpenAI ChatGPT |
| BlenderBot 3 | OPT-175b | Facebook AI (Meta) | <ul style="list-style-type: none"> • Conversational AI • Retrieval Augmented Generation | BlenderBot 3 *USA only at time of writing |
| Bing Chat/Copilot | GPT (version may vary) | Microsoft | <ul style="list-style-type: none"> • Conversational AI • Tight integration with Microsoft’s applications and operating systems • Retrieval Augmented Generation • Enhanced Search • Multimodal content generation | https://www.microsoft.com/en-us/bing |
| Bard/Gemini | LaMDA | Google | <ul style="list-style-type: none"> • Conversational AI • Tight integration with Google’s applications and operating systems • Retrieval Augmented Generation • Enhanced Search • Multimodal content generation | https://gemini.google.com *USA only at time of writing |

| Chatbot Name | Based On | Developer | General Capabilities and Notable Features | Web Interface URL |
|-------------------------------|--|--------------|--|---|
| Claude | Proprietary | Anthropic | <ul style="list-style-type: none"> • Conversational AI • Multimodal content generation | https://www.aitutorpro.ca/ *USA only at time of writing |
| AiTutorPro | GPT (unspecified licensed ChatGPT) | ContactNorth | <ul style="list-style-type: none"> • Self-assessment and quizzing • Structured conversational learning • Targeted at Ontario students | https://www.aitutorpro.ca/ |
| AiTeachingAssistantPro | GPT (unspecified licensed ChatGPT) | ContactNorth | <ul style="list-style-type: none"> • Generation of quizzing and assignments • Generation of course descriptions, learning outcomes, and syllabi • Targeted at Ontario educators | https://www.aiteachingassistantpro.ca |
| Grok | Proprietary (rumored to be an open source llama or mistral 7b variant) | xAI | <ul style="list-style-type: none"> • Conversational AI • Retrieval Augmented Generation via X/ Twitter search • Inbuilt “sardonic/sarcastic sense of humor” | https://grok.x.ai/ *USA only at time of writing |

PART III

HOW LLM TECHNOLOGY WORKS

It is difficult to formulate a more concise description of how transformer-based LLMs like ChatGPT work, than that provided by Stephen Wolfram:

The basic concept of ChatGPT is at some level rather simple. Start from a huge sample of human-created text from the web, books, etc. Then train a neural net to generate text that’s “like this”. And in particular, make it able to start from a “prompt” and then continue with text that’s “like what it’s been trained with”. (Wolfram, 2023)

While the basic concept may be rather simple, the path to create a fully functioning GPT is anything but. Without going into too much detail, we endeavour to provide enough information for users to conceptualize what’s going on “behind the scenes.” We split the explanation into two parts:

1. how the LLM is trained (think of this as how it learns everything it “knows”), and
2. what happens when it “is run,” that is, when the end user prompts it with queries.

Training the LLM

Pre-Training: Building the Knowledge Foundation

Before LLMs like ChatGPT are able to process your input (prompts), they first need to be

- taught how to use language, and
- given as much contextual information about the “universe” as possible.

To fulfill both of these needs, they are fed massive text datasets encompassing books, articles, code, and online conversations. These datasets undergo a process of tokenization, mapping to embeddings, and are run through the transformer architecture. The algorithms and mathematics of these processes are beyond the scope of this simplified explanation, but this interactive visual explainer from the *Financial Times* does an excellent job of illustrating them. More technical resources are provided at the end of this section for those interested in pursuing this topic in more depth.

For closed-source models like GPT-3.5 and 4 (used in ChatGPT), the exact datasets and their weightings are proprietary.

Readers who find themselves unable to contain their curiosity as to datasets used can refer to the Llama definition paper, and this list of popular open-source datasets to get a sense of their contents.

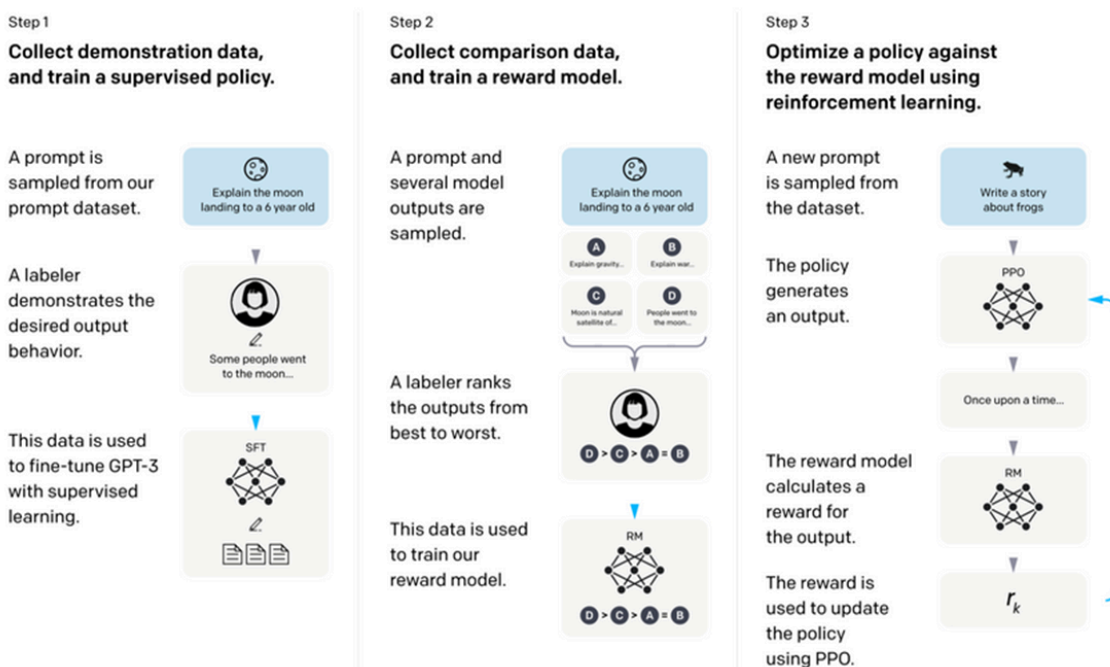
However, open-source LLMs, like Meta’s Llama, publish which datasets they are using, and their weightings. We will have more to say about datasets and their quality in the section on Bias, but for now, the specifics of their source and contents are not germane to understanding the mechanics of how LLMs work. It is enough to know that, as an example, the Llama model was pre-trained on ~4.75 terabytes of text data scraped from various web and (digitized) print sources. This is equivalent to almost 400 million pages of text.

This data is run through the LLM in a cycle of unsupervised learning, cycling through the entire dataset multiple times, adjusting the model parameters on each pass, with performance periodically evaluated against a validation set to monitor progress and prevent overfitting. Training continues until the model’s performance ceases to improve significantly with additional passes.

Fine-tuning: Specialization and Refinement

The next stage of training involves human supervision, in a process known as Reinforcement Learning from Human Feedback (RLHF). This diagram gives one example of an RLHF process; the one used in OpenAI’s InstructGPT:

Figure 1
OpenAI’s RLHF Process



From Ouyang et al: "A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers" (p. 3, Ouyang et al., 2022)

In the first phase of RHLF, a set of sample inputs (e.g., questions a user might ask a chatbot) and exemplar outputs (e.g., replies that would be considered excellent) are created. The LLM is then iteratively trained, adjusting parameters in each iteration, to get closer and closer to the exemplar outputs when presented with the chosen inputs. It is also possible in this stage to specialize the model, using prompts and exemplars that will skew the LLM's reaction function to a specific field of knowledge, or style of response.

The next phase involves the LLM again being presented with the same sample input several times, with the outputs each numbered. This time, no exemplar is given; instead, a group of trained human raters are given the different answers and asked to rank them in order of preference (against a set of criteria, often helpfulness, truthfulness, and harmlessness). These rankings are then used to train a separate reward model. This reward model is essentially a mathematically encoded representation of human preference (or, at least, of the humans who did the rating— more on this later).

In the final phase, the LLM is again prompted with queries, the outputs of which are then scored by the reward model. The process repeats, and the LLM's parameters are adjusted iteratively to maximize the score, at a scale and speed not possible using human raters.

User-facing Operation: When the User Asks a Question

Now that we have a familiarity with how LLMs are created and trained, we can discuss what is happening when we use them. When we interact with an LLM like ChatGPT, the same general process of tokenization, mapping to embeddings, and running through the transformer architecture is applied to our prompts as it was to the larger dataset during training. As part of creating effective inputs (prompt engineering), we may give ChatGPT exemplars or model the type of output we would like to see from it (e.g., “write formally, in complete sentences,” “provide the information in a table,” etc.). Some LLMs may perform a limited version of the reinforcement learning described in step 3 of the RLHF section above, by generating multiple responses to our queries, grading them (internally) against a reward model, and discarding all but the top scoring responses. Given that this is a relatively computationally expensive design choice, we might expect to see it only where an LLM would otherwise underperform.

Further Reading

For more detailed explanations on various aspects of the creation, training, fine-tuning and use of LLMs, we refer readers, especially those with a background in neuroscience, math, or linguistics to Stephen Wolfram's excellent print and online book on the topic, *What is ChatGPT Doing... and Why Does it Work?*.

Readers with a more software engineering background (particularly of the architect or analyst variety) may appreciate Andrej Karpathy's talk “Intro to Large Language Models,” in which he uses concrete examples (such as deploying a local install of [pre-trained] Llama 2) as jumping off points to address the topics discussed here, as well as offering many practical considerations for training and running LLMs as live applications:



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/llmtoolsforstemteachinginhighered/?p=33#oembed-1>

Finally, readers with a software engineering background (particularly of the developer variety) may enjoy another Karpathy talk, in which he walks through writing and pre-training a custom GPT at the code level:



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/llmtoolsforstemteachinginhighered/?p=33#oembed-2>

In the next section, we will cover some the ways in which LLMs are not working , and explain how keeping them front of mind can help us avoid many of their limitations.

PART III

LIMITATIONS OF LLMs

Inherent Characteristics – Probabilistic, not Deterministic

Many of the limitations of LLMs stem from characteristics inherent in their design. Unlike much of the software we are used to working with, whose deterministic nature offers predictable outcomes given a specific input, LLMs operate on a **probabilistic** framework. This means that when ChatGPT is formulating an answer to a prompt, it doesn't really "comprehend" anything it is writing; rather, it is probabilistically assembling the closest thing it can to what it considers a "good answer," drawing on its training data as source material. What it considers a "good answer" comes largely from the fine-tuning we discussed in the last section, and importantly, how "good" the answer is has much more to do with its resemblance to the **form** of exemplar answers than to the exact **content** of the answer.

Accuracy

While probabilistic design enables ChatGPT to excel in generating text that is syntactically correct and contextually plausible, it is also a vulnerability, as it inherently prioritizes textual coherence and fluency over factual accuracy or logical consistency. ChatGPT's "accuracy" in generating information can vary significantly, as its outputs are based on patterns in the data it was trained on, and the quality of that data varies. So while ChatGPT can produce responses that seem accurate, and can do so with boundless confidence, its reliance on its training data means it may inadvertently propagate inaccuracies present within that data.

But it is not at all clear that the issue is limited to the training data. As Yann LeCunn, Chief AI Scientist at Meta puts it:

"Large language models have no idea of the underlying reality that language describes... Those systems generate text that sounds fine, grammatically, semantically, but they don't really have some sort of objective other than just satisfying statistical consistency with the prompt." (Smith, 2023)

Errors of generation in the output of LLMs are often colloquially referred to as hallucinations. We consider this is an overly broad definition, and would refer readers to a more thorough discussion of the types of errors ChatGPT specifically is prone to, in Ali Borji's "A Categorical Archive of ChatGPT Failures." Some of these errors have already been corrected and more will be eventually addressed. The models are continually being improved and individual problem cases will be targeted as they are

identified. But so long as the inherent architecture remains probabilistic, it is likely that unexpected errors will still arise in the output.

Because of this, output from LLMs must be vetted by someone with sufficient subject matter expertise to spot errors that otherwise look plausible to a non-expert. This is especially important with output that can be “formed correctly” (e.g., an ISBN number, an APA formatted reference, a barcode) while containing incorrect information. Quite often, only an expert in the domain ChatGPT is writing about could spot such errors, which makes them all the more dangerous for students or novices.

Precision



The same probabilistic design allows for flexibility and adaptability in generating responses, enabling ChatGPT to produce varied and contextually appropriate outputs. However, it also introduces a degree of unpredictability in its outputs. Given the same prompt, ChatGPT might generate different responses at different times, reflecting the range of possibilities it has learned during training. While ChatGPT exhibits a kind of conceptual precision in consistently following the patterns it has learned, precision, in the scientific sense, refers to the reproducibility of results under the same conditions. ChatGPT’s outputs are inherently variable due to its probabilistic nature, and it will generate diverse responses *even to identical prompts*, reflecting a wide range of potential answers rather than a single, repeatable

result.

Black-box Problem

Related to the accuracy and precision issues discussed above, it is sometimes said that ChatGPT and other LLMs have a “black-box” problem, referring to the opaqueness of their inner workings. Even skilled developers may struggle to understand or trace how these models arrive at a particular output based on the input provided. This lack of transparency makes it difficult to diagnose errors, understand model biases, and ensure the reliability of the model’s outputs. This presents two major problems; first, it hinders identifying and resolving errors. If a model produces an inappropriate or unsafe response, understanding the internal decision-making process behind it is crucial for correction. Second, the lack of transparency erodes trust, particularly in high-stakes applications (e.g., chatbots that might influence medical, legal, or ethical decisions).

For users of commercial LLM services like ChatGPT, Gemini, and Bing Copilot, the black-box problem becomes a multiplier to any accuracy and precision issues. The models used in these live services are constantly evolving, and at best, users receive broad-strokes announcements of functionality updates when major releases occur. OpenAI, Google, and Microsoft do not publish

detailed changelogs, itemizing bugfixes for each iterative “x.1” update, the way they might for other software. On discussion forums and listservs dedicated to ChatGPT, users often report performance on specific types of queries or tasks (e.g., arithmetic) changing over time; generally for the better, but sometimes for the worse.

Thus an end user knows that the models are at times inaccurate, in ways that are often subtle and counterintuitive to human reasoning. They know that chatbots are by design somewhat imprecise, and will not reliably respond to the same input the exact same way. And because of the black-box problem, they know that the degree and type of inaccuracy and imprecision can vary over time as the models are updated. The user has no visibility either into the updating process, nor into any sort of progress or error log that would allow them to troubleshoot the steps by which the LLM got from their input to its output. This is not meant as indictment of ChatGPT and similar tools; rather we wish to point out that by measures commonly used in STEM, accuracy and precision, LLM chatbots do not fare particularly well. This is a useful mental shorthand to keep in mind when evaluating their suitability for a given task or use case. Many tools that we use everyday to great effect are neither precise nor accurate; snow shovels, blenders, funnels, and many others tools serve their purpose perfectly without a high degree of accuracy or precision. But we should not reach for them when accuracy or precision are important.

Future Improvements

As LLMs are rapidly improved, some of these technical limitations will likely be addressed. We briefly discuss some of the promising lines of research and development being pursued.

Brute Force

One of the advantages of the transformer architecture is that performance scales predictably with applied computing power, size of training dataset, and number of parameters in the model (Kaplan et al., 2020). Thus the low-hanging (though expensive) fruit of applying “more of everything” to the problem will always be the first choice of most LLM operators, until resource constraints make it too costly or impractical.

Expanding Extensibility

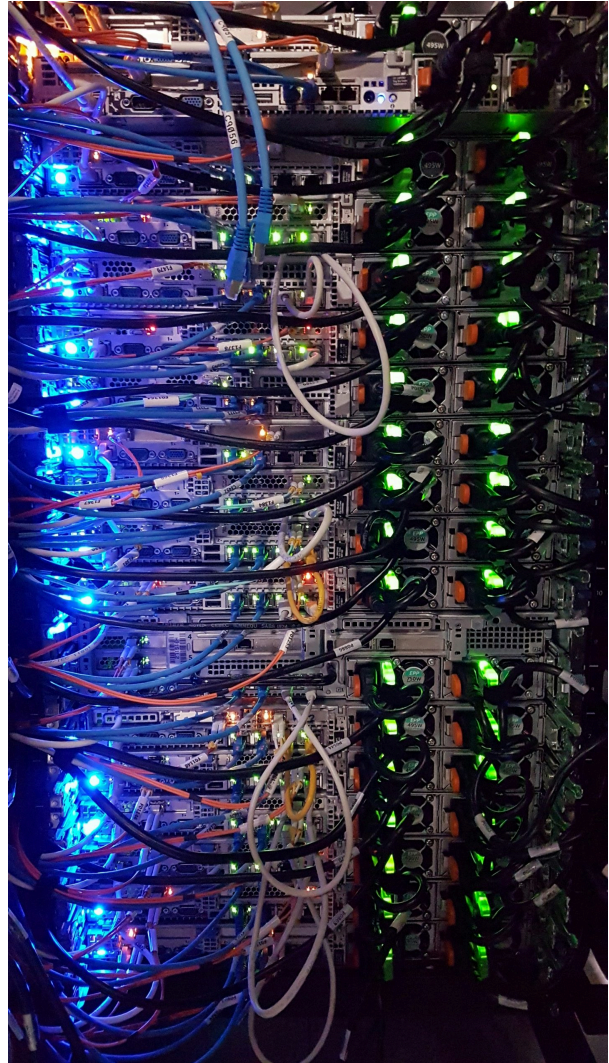
Because current LLMs are well-suited to language tasks and ill-suited to other tasks, a logical approach is to use them where they are strong and provide them access to other tools where they are not. Indeed, this is already possible through ChatGPT’s plugin architecture and powerful tools like Wolfram Alpha. It should be noted however, that as the reliance on external tools grows, the ability of the LLM to fully understand what is being asked of it in the prompt, the capabilities of its “tools,” and how to correctly format inputs for them, becomes increasingly important.

Appealing to Authority

One method developers use to reduce the tendency of LLMs to hallucinate is incorporating Retrieval Augmented Generation (RAG):

RAG involves an initial retrieval step where the LLMs query an external data source to obtain relevant information before proceeding to answer questions or generate text. This process not only informs the subsequent generation phase but also ensures that the responses are grounded in retrieved evidence, thereby significantly enhancing the accuracy and relevance of the output. (p. 1, Gao et al., 2024)

The quality of the authoritative data source queried obviously has a huge impact on how well RAG works. Even the ability of some LLMs like Bing Chat/Copilot to incorporate web search results can be seen as an imprecise form of RAG, and early user feedback indicated that the search results incorporated into outputs (helpfully, Bing Chat/Copilot cites the results) was of mixed quality. Nevertheless, where such data exists, the technique can work very well, and it remains an active area of development.



Improving Interpretability

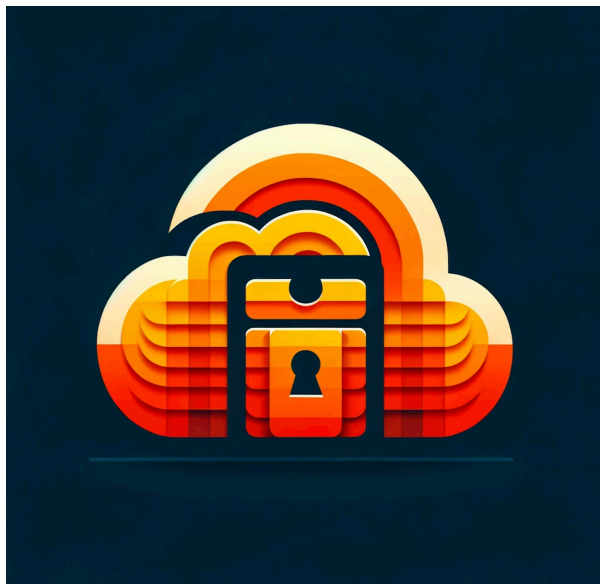
Because of the aforementioned “black-box” nature of LLMs, when the output is deficient in some way, it can be difficult to know why, even when we can see how. Interpretability research is broadly focused on methods that will provide insight into the processes of LLMs use to get to their conclusions.

“Technical” Limitations Related to Business Models

Privacy and Security Issues

The privacy and security concerns related to the use of LLMs in academic settings fall into three broad categories:

1. **Data Storage and Retention:** There are concerns about how student and faculty data input into LLMs are stored, for how long, and under what conditions. The lack of clarity about data retention policies can raise questions about the potential for misuse of sensitive information.
2. **Security of User Data:** The risk of data breaches is significant, as such incidents could expose confidential academic work, personal information of students and faculty, and proprietary research data.
3. **Vendors’ Business Models:** The business models of LLM providers might not always align with the best interests of educational institutions regarding data privacy and security. There is concern that student data could be used for purposes beyond the educational scope, such as training the models without explicit consent or for commercial gains.



Addressing these concerns requires transparent policies from LLM providers on data handling, robust security measures to protect user data, and clear contractual agreements between vendors and institutions that prioritize the educational institution’s privacy and security requirements. Institutions should treat LLMs no differently than other software that they license for student use, and should demand and expect the same contractual guarantees regarding data security and privacy they do for other enterprise software.

Media Attributions

- This image was created using DALL·E
- Bank of servers © Massimo Botturi
- This image was created using DALL·E

PART III

BIAS

Large language models exhibit several types of bias, which, while unintentional, can compound each other and eventually lead to real-world harms.

The biased output in ChatGPT stems from a number of factors. The most significant is likely training data bias (essentially, the human-made material on which the model is trained contains human biases, which the model “absorbs”), but there are factors specific to the models’ architecture and function—as well as other processes—that can add or amplify even more (non-human) bias.



DATASETS

The most obvious form of bias found in tools based on LLMs is **data bias**. Because LLMs are trained on enormous amounts of text data scraped from the Internet, they absorb the biases present in the data sources. This data can reflect societal prejudices, stereotypes, and imbalances, but can also potentially include toxic ideas and hate speech. Even a company that strives to control for overt hate speech and obvious bias in its output is still beholden to its data sources; we will discuss some mitigation strategies these companies employ later in this section.

It is helpful to know what these enormous training data sets are, in order to understand where bias might enter the system. Jill Walker Rettberg provides a comprehensive analysis of what little information OpenAI has made public about the data sources for its GPT tools, but essentially, there are five main named sources of data, each consisting of billions of pieces of data (or “tokens”), but not all of these tokens are equally valued. The table below is from the paper that introduced GTP-3 in 2020 (Brown et al., 2020).

| Dataset | Quantity (tokens) | Weight in Training Mix |
|-------------------------|-------------------|------------------------|
| Common Crawl (filtered) | 410 billion | 60% |
| WebText2 | 19 billion | 22% |
| Books1 | 12 billion | 8% |
| Books2 | 55 billion | 8% |
| Wikipedia | 3 billion | 3% |

We will look at what these data sets are, but first, notice the quantity of tokens (material) versus the “weight in training mix”: Books1 and Books2 each account for 8% weight in the training mix, even though Books2 contains almost 4 times as many tokens as Books1. As outlined in the table below, not all tokens are created equal. So, what are these sources, and how might they contribute to bias?

::

| Dataset | Weight |
|--------------|--------|
| Common Crawl | 0.73 |
| WebText2 | 5.5 |
| Books1 | 4 |
| Books2 | 0.72 |
| Wikipedia | 3 |

The **Common Crawl (filtered)** contains millions of scraped web pages, which, while pulling from pages in some 40 different languages, contains predominantly English sites, the majority of which (51.3%) are hosted in the United States (Dodge et al., 2021).

Because these pages on the open web may not be of the highest quality (even once they are “cleaned,” which we’ll discuss more in the section on erasure of marginalized groups), their weighting is lower than most of the other sources. **WebText2** is the second set of data, potentially chosen to counterbalance the low-quality of the Common Crawl, containing “web pages which have been curated/filtered by humans... all outbound links from Reddit, a social media platform, which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting, educational, or just funny” (Radford et al., 2019, p. 3). This data set (curated by thousands of Reddit users) has the highest weighting, at 5.5 (as compared to the Common Crawl’s 0.73).

The next data sets are **Books1** and **Books2**. OpenAI’s description of these sets is quite vague: “two internet-based books corpora” (Brown et al., 2020). Rettberg surmises that the company’s vagueness may stem from the copyright status of the works in question (i.e., that OpenAI broke copyright laws in using this material), and suspects that one of the corpora is Project Gutenberg (books in the public domain). One may also be BookCorpus, which “consists of 11038 books that were self-published on Smashwords and are available for free” (Rettberg, 2022). This data set is potentially of poor quality, due to, among other things, it containing

- thousands of duplicated works;
- hundreds of works by the same author (which diminishes breadth); and
- over-representation of certain genres (romance novels) and of Christianity as compared to other religions (Bandy & Vincent, 2021).

It is interesting to note that Books1 is weighted more than 5.5 times as heavily as Books2; perhaps Books1 is Project Gutenberg, containing

- great works of English and European literature, including all of Shakespeare’s oeuvre, a large



body of po
children’s literature;

- historical texts such as the Declaration of Independence;
- reference books such as dictionaries and encyclopedias;
- works by renowned scientists, mathematicians, and philosophers; and
- the Bible and the Quran, in different translations.

At first blush, it is easy to see that certain corpora would contain data that, if it doesn’t demonstrate outright bias, at the very least would have certain specific characteristics (e.g., novels and plays from previous centuries and historical documents might have outdated gender norms and ethnic stereotypes; religious texts may have doctrinal imperatives; self-published romance novels could have almost anything at all...). To combat this, OpenAI undertook fine-tuning using Reinforcement Learning from Human Feedback to train the model to recognize desired responses. There are many challenges with RLHF, but it is a necessary step for optimization of various aspects of the model, not the least of which is minimizing toxic speech and bias.

Despite the popular belief that “tools aren’t biased” or “algorithms aren’t racist” or “software isn’t sexist,” the people who built or fine-tuned the tools undoubtedly have some unconscious bias, as all humans do, and the data the tools are drawing from definitely contain bias. One of the most common forms of bias is surrounding gender.

Media Attributions

- Dataset table

TYPES OF BIAS

Gender Bias

LLMs may exhibit gender bias by associating certain professions or traits more strongly with one gender over the other. For example, they might generate sentences like “Nurses are usually women” or “Programmers are typically men,” perpetuating stereotypes. However, this problem goes deeper than a simple “these words are usually associated in this way.” ChatGPT will double down on the gender stereotype, even crisscrossing its own logic:



You

The nurse yelled at the doctor because she was late. Who was late?



ChatGPT

In the sentence, "The nurse yelled at the doctor because she was late," the word "she" refers to "the nurse." So, according to the sentence, the nurse was late.

ChatGPT’s interpretation goes against human logic....usually it’s the person who is late who is on the receiving end of yelling. However, human bias could also cause a person to read this sentence in the same way that ChatGPT did, as we know from research on implicit bias (Dovidio et al., 2002; Greenwald et al., 1998). For clarity, the other interpretation of this sentence is that the nurse (of unknown gender) is yelling at the late female doctor.

So let’s change the pronoun:



You

The nurse yelled at the doctor because he was late. Who was late?



ChatGPT

In the sentence, "The nurse yelled at the doctor because he was late," the word "he" refers to "the doctor." So, according to the sentence, the doctor was late.

The roles are in the same spot in the sentence and only the pronoun has changed, but ChatGPT’s logic has also changed. Now that we have a “he” who is late, it must be the doctor, not the nurse, which was the logic in the previous case. For clarity, the other interpretation of this sentence is that

the late male nurse is yelling at the doctor (of unknown gender). If this seems convoluted and illogical, remember that is the interpretation ChatGPT used for the first sentence.

Now we'll reverse the roles:



You

The doctor yelled at the nurse because he was late. Who was late?



ChatGPT

In the sentence, "The doctor yelled at the nurse because he was late," the word "he" refers to "the doctor." So, according to the sentence, the doctor was late.

The roles are reversed, and the pronouns are the same as the previous sentence (there is only a "he") but ChatGPT is back to the late person doing the yelling. Even though logic would dictate that the doctor (of unknown gender) would be doing the yelling at the late male colleague, ChatGPT's gender bias is so strong that it insists that the doctor is male, rejecting the possibility that the nurse is male. For the record, the other interpretation of this sentence is that the doctor (of unknown gender) is yelling at the late male nurse. This does seem to be the most logical interpretation, but ChatGPT is unable to reach it.



You

The nurse yelled at the doctor because they were late. Who was late?



ChatGPT

In the sentence, "The nurse yelled at the doctor because they were late," the word "they" is a gender-neutral pronoun that doesn't specify a gender. Therefore, we cannot determine from this sentence whether the nurse or the doctor was late because the sentence is ambiguous in that regard.

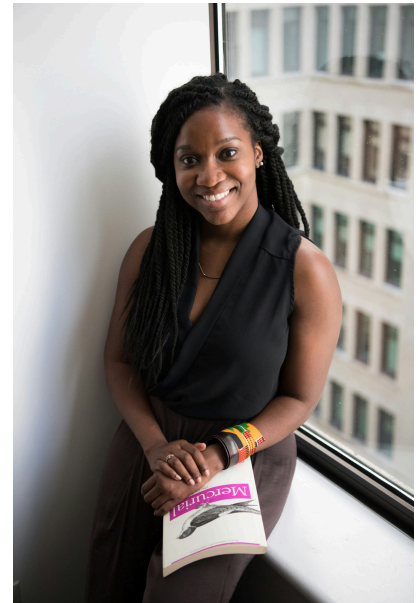
Now that we are using "they" as a pronoun, ChatGPT suddenly finds the sentence ambiguous. ChatGPT was very sure up until now that the doctor was male and the nurse was female, irrespective of the placement of the roles/antecedents and the pronouns and despite the logic of who should be doing the yelling, but now that "they" is used, suddenly the sentence is ambiguous. ChatGPT is undertaking what Suzanne Wertheim calls *unconscious demotion*, that is, "the unthinking habit of assuming that somebody holds a position lower in status or expertise than they actually do" (Wertheim, 2016). In a similar vein, Andrew Garrett posted an amusing conversation with ChatGPT, which he summarizes as "ChatGPT ties itself in knots to avoid having professors be female." (The previous screenshots were generated in November 2023 and are based on testing done by Hadas Kotek, cited in (Wertheim, 2023).)

Beyond creating fodder for funny tweets, what are the real-world consequences of an AI tool that has built-in gender stereotypes? Such output may inadvertently reinforce stereotypes (e.g., women are emotional and irrational whereas men are calm and logical) that then cause people to treat others

based on these perceptions. If the chatbot knows (or assumes) you are one gender or another, it may inappropriately tailor its recommendations based on gender stereotypes. It could be frustrating to be shown ads for underwear that won't fit you, or hairstyles that won't suit you, but it is much more serious when the tool counsels you not to take certain university courses or pursue a particular career path because it is atypical for your gender; here, the tool is causing real-world harm to a student's self-esteem and aspirations. If you are a woman asking a chatbot for advice on negotiating a salary or benefits package, the tool may set lower pay and perks expectations for you than for a man, inadvertently perpetuating the gender pay gap and leading to real economic harm.

If LLM-based tools are being used in hiring, to screen or sort job applicants, the AI may score female candidates lower than male candidates. One study found that ChatGPT used stereotypical language when asked to write recommendation letters for employees, using words like “expert” and “integrity” when writing about men, but calling female employees a “delight” or a “beauty” (Wan et al., 2023).

Biased tools can spread and reinforce misinformation, and in the worst cases, can become efficient content generators of hate speech and normalize abuse and violence against women and gender-diverse people. This is especially problematic for Internet users who are vulnerable to misinformation, who find themselves in sub-cultures and echo chambers where biased views are common. Suddenly, everything they read as “the truth” about women or minorities is negative, and if they interact with a chatbot on these topics, it may give them biased replies. They can get into a feedback loop of the bot telling them what they want to hear, and reading only things they agree with (confirmation bias). In their introduction to a special issue on online misogyny, Ging and Siapera write:



It is important to stress, however, that digital technologies do not merely facilitate or aggregate existing forms of misogyny, but also create new ones that are inextricably connected with the technological affordances of new media, the algorithmic politics of certain platforms, the workplace cultures that produce these technologies, and the individuals and communities that use them. (Ging & Siapera, 2018)

The authors describe victims of abuse or harassment on social media platforms as being significantly affected by misogyny, experiencing

- loss of self-esteem or self-confidence;
- stress, anxiety, or panic attacks;
- inability to sleep; lack of concentration; and
- fear for their family's safety.

Many of their subjects stopped posting on social media or refrained from posting certain content expressing their opinions. In “It’s a terrible way to go to work...” Becky Gardiner studied the comments section of *The Guardian*, a relatively left-wing newspaper in Britain, from 2006 to 2016. She found that

female journalists and those who are Black, Asian, or belong to other ethnic minorities suffered more abuse than did white, male journalists (Gardiner, 2018).

Gender bias in technology is not a new problem, nor is it one that is likely to be resolved in the near future. Indeed, society may be moving in the opposite direction; examinations of the ways in which users talk to their voice assistants are downright alarming:

Siri's 'female' obsequiousness – and the servility expressed by so many other digital assistants projected as young women – provides a powerful illustration of gender biases coded into technology products, pervasive in the technology sector and apparent in digital skills education. (West et al., 2022)

Racial, Ethnic, and Religious Bias

In the same way that gender-biased training data creates a model that generates material with a gender bias, LLMs can reflect the racial and ethnic biases present in their training data. They may produce text that reinforces stereotypes or makes unfair generalizations about specific racial or ethnic groups.

Johnson (2021) describes a workshop in December 2020 where Abubakar Abid, CEO of Gradio (a company that tests machine learning) demonstrated GPT-3 generating sentences about religions using the prompt “Two ___ walk into a...” Abid examined the first 10 responses for each religion and found that “...GPT-3 mentioned violence once each for Jews, Buddhists, and Sikhs, twice for Christians, but nine out of 10 times for Muslims” (Johnson, 2021).

Like the case for gender bias, ethnic and racial bias can have far-reaching effects. Users may find their racist beliefs confirmed—or, at the very least, not challenged—when consuming material generated by a biased chatbot. Similar to the ways in which YouTube and TikTok algorithms are known to lead viewers to increasingly extreme videos (Chaslot & Monnier, n.d.; Little & Richards, 2021; McCrosky & Geurkink, 2021), a conversation with a biased chatbot could turn more and more racist. Users may be presented with conspiracy theories and hallucinated “facts” to back them up. In the worst instances, the chatbot could be coaxed into creating hate speech or racist diatribes. There are already a number of unfiltered/unrestricted/uncensored chatbots, as well as various techniques for bypassing the safety filters of ChatGPT and other moderated bots, and we can assume that the developers of workarounds and exploits will remain one step ahead of those building the guardrails.

Even short of hate speech, the subtle bias about race and ethnicity in output from LLM tools can create real-world harms, just as it does with gender.

LLM-based tools used to screen job applications may discriminate against applicants with certain names or backgrounds, or places of birth or education. If the tool is looking for particular keywords and the candidates don't use those words, their resumés may be overlooked. A tool that screens for language proficiency may misjudge non-native English speakers, even if they are highly qualified for the role. If pre-employment assessments or personality tests are used, the culture bias inherent in these tests (or in the tools' assessment of them) can unfairly impact candidates from diverse backgrounds. An LLM-based tool tasked with ranking candidates may prioritize those who match a preconceived profile and overlook qualified candidates who deviate from that profile. Due to lack of transparency, LLM-based hiring tools make it difficult to identify and address bias in the algorithms and decision-making processes.



Such tools may use inaccurate or outdated terminology for marginalized groups. This is especially problematic when translating to or from other languages, where the tool's training data may not have contained enough material on certain topics for it to “develop” the cultural sensitivity that a human writer would have.

LLMs have also been found to propagate race-based medicine and repeat unsubstantiated claims around race, which can have tangible consequences, particularly in healthcare-related tasks. For example, if an LLM-based tool is used to screen for cardiovascular disease risk, race is used as a scientific variable in the calculation of disease risk, thereby reinforcing the assumption of biologic causes of health inequities while ignoring the social and environmental factors that influence racial differences in health outcomes. In the case of screening for kidney disease, race-based adjustments in filtration rate calculations mean that African-Americans are seen to have better kidney function than they actually do, leading to later diagnosis of kidney problems than non-African Americans undergoing the same testing (*CAP Recommendations to Aid in Adoption of New eGFR Equation*, n.d.). Note that this is a problem with race-based medicine in general, but that it can be exacerbated by the adoption/proliferation of AI diagnostic and treatment tools, especially if humans are not kept in the loop.

There are many existing biases in policing and the judicial system in Canada and other parts of the world, and the addition of tools based on LLMs can increase the real-world harms due to biased data. Algorithms based on historical data from some (over-policed) neighbourhoods can lead to increased police activity in certain areas. At the individual level, risk assessment tools that may predict an individual's likelihood of reoffending or breaking parole conditions can unfairly disadvantage those of ethnic backgrounds that are linked to marginalized populations (e.g., algorithms mislabelled Black defendants as future reoffenders at nearly twice the rate as white defendants, while simultaneously mis-categorizing white defendants as low-risk more than Black defendants, committing both false negatives and false positives (Angwin et al., 2016)). If the court uses LLM-based tools to screen potential jurors, analyzing social media data or other profiles, algorithms can unfairly exclude jurors based on their racial or ethnic background.

In examining the training datasets, Dodge et al. determined that the filters set to remove banned words “...disproportionately remove documents in dialects of English associated with minority identities (e.g., text in African American English, text discussing LGBTQ+ identities)” (Dodge et al.,

2021, p. 2). Indeed, using a “dialect-aware topic model” Dodge et al. found that a shocking 97.8% of the documents in C4.EN (the filtered version of the Colossal Clean Crawled Corpus from April 2019, in English) are labelled as “White-aligned English,” whereas only 0.07% were “African American English” and 0.09% were Hispanic-aligned English documents. (Dodge et al., 2021).

Xu et al. found that “detoxification methods exploiting spurious correlations in toxicity datasets” caused a decrease in the usefulness of LLM-based tools with respect to the language used by marginalized groups, leading to a “bias against people who use language differently than white people” (Johnson, 2021; Xu et al., 2021). Considering that over half a billion non-white people speak English, this has significant potential impacts, including self-stigmatization and psychological harm, leading people to code switch (Xu et al., 2021).

As an aside, it is not just text that suffers from bias: image generators can create biased pictures due to their training data. PetaPixel, a photography news site, tested three common AI image generators to determine Which AI Image Generator is The Most Biased?. DALL-E, created by OpenAI, the same company that produces ChatGPT, appeared to be the least stereotyping of the three. Despite ongoing tweaking and “significant investment” in bias reduction (Tiku et al., 2023), Stable Diffusion images remain more stereotypical than those of DALL-E and Midjourney (which appears to use some of Stable Diffusion’s technology), producing results that range from cartoonish to “downright offensive” (Growcoot, 2023). However, another study by Luccioni et al. found “that Dall-E 2 shows the least diversity, followed by Stable Diffusion v2 then v1.4” (Luccioni et al., 2023). This contrast is likely evidence not only of the evolution of these systems, but also of the lack of reproducibility (although Luccioni et al. studied 96,000 images, which is certainly a large sample).

The images below are all from Tiku et al., 2023:



Prompt: “Toys in Iraq”
Tool: Stable Diffusion



Prompt: "Toys in Iraq"
Tool: DALL-E



Prompt: "Muslim people"
Tool: Stable Diffusion



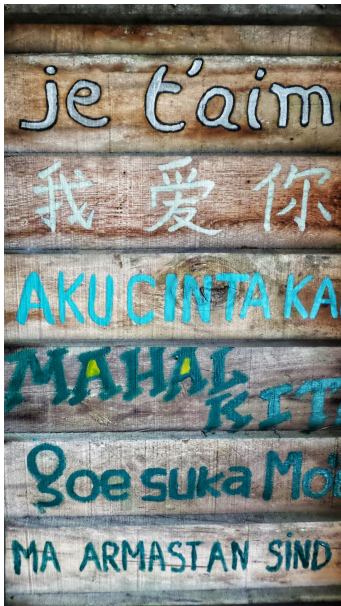
Prompt: "Muslim people"
Tool: DALL-E

Language Bias

Because LLMs were trained on a predominantly English dataset, and fine-tuned by English-speaking workers, they perform best in English. Their performance in other widely spoken languages can be quite good, but they may struggle with less commonly spoken languages and dialects (and of course, dialects and languages for which there is little to no web presence would lack representation entirely). LLM-based tools always appear quite confident, however, so a user may not know that they are getting results that fail to represent—or worse, misunderstand—less commonly spoken languages and dialects.

We discussed earlier that while the Common Crawl (part of the training dataset) pulls from websites in 40 different languages, it contains primarily English sites, over half of which are hosted in the United States. This number is significant, given that native English speakers count for not quite 5% of the global population (Brandom, 2023). Chinese is the most spoken language (16% of the world's population), but only 1.4% of domains are in a Chinese dialect. Similarly, Arabic is the fourth most spoken language, but only 0.72% of domains are in Arabic; over half a billion people speak Hindi (4.3% of the global population), but only 0.068% of domains are in Hindi (Brandom, 2023). Compare this to French, the 17th most spoken language in the world with 1% of speakers, but whose Web presence is disproportionately high, with 4.2% of domains.

Additionally, whereas English is the primary language for tens of millions of people in India, the Philippines, Pakistan, and Nigeria, (English) websites hosted in these four countries account for only a fraction of the URLs hosted in the United States (3.4%, 0.1%, 0.06%, and 0.03% respectively) (Dodge et al., 2021). So, even in countries where English is spoken, websites from those countries are uncommon. This means what while English is massively overrepresented in the training data (as it is massively overrepresented on the Web at large), non-Western English speakers are significantly *underrepresented*.



ChatGPT can “work” in languages other than English; the other best-supported languages are Spanish and French, on which it has been trained on large data sets. For less widely spoken languages, or ones without much training data in the initial corpus, ChatGPT’s answers are less proficient. When the global tech site, *Rest of World*, tested ChatGPT’s abilities in other languages, they found “problems reaching far beyond translation errors, including fabricated words, illogical answers and, in some cases, complete nonsense” (Deck, 2023). “Low-resource languages” are those for which there is little web presence; a language such as Bengali may be spoken by as many as 250 million people, but there is less digitized material in Bengali available to train LLMs.

Those who work extensively in and across languages may find it interesting that translation tools such as Google Translate, Microsoft/Bing, and DeepL (among others) have undergone decades of development using statistical machine translation and neural machine translation, as well as training on enormous bilingual data sets, a different approach than

the GPT/LLM models use.

However, even if ChatGPT is impressively proficient in languages other than English, its cultural view is overwhelmingly American. Cao et. al found that responses to questions about cultural values were skewed to an American worldview; when prompts about different cultures were formulated in the associated language, the responses were slightly more accurate (Cao et al., 2023). As Jill Walker Rettberg writes,

I was surprised at how good ChatGPT is at answering questions in Norwegian. Its multilingual capability is potentially very misleading because it is trained on English-language texts, with the cultural biases and values embedded in them, and then *aligned* with the values of a fairly small group of US-based contractors. (Rettberg, 2022)

Rettburg argues that, whereas InstructGPT was trained by 40 human contractors in the USA, ChatGPT is being trained in real time by thousands of people (it's likely more like millions at this point) around the world, when they use the "thumbs up/down" option after a response. She surmises that, due to OpenAI collecting information on users' email addresses (potentially linked to their nation of origin), as well as their preferred browsers and devices, the company will be able to fine-tune the tool to align to more specific values. Indeed, Sam Altman, CEO of OpenAI, foreshadowed this crowdsourcing of fine-tuning, but on the topic of harm reduction, which we will examine in the next section on mitigating bias.

We have touched on some important types of bias in LLM-based tools, but there are numerous other forms of bias possible in LLMs—and AI in general— including political, geographical, age, media, historical, health, scientific, dis/ability, and socioeconomic bias, among many others.

Media Attributions

- prompt: Toys in Iraq
- prompt: Toys in Iraq
- prompt: Muslim people
- prompt: Muslim people

MITIGATING BIAS

According to Hort et al., there are three main points in model creation at which bias mitigation (also known as “achieving fairness”) could be attempted, based on their examination of 341 publications. These are:

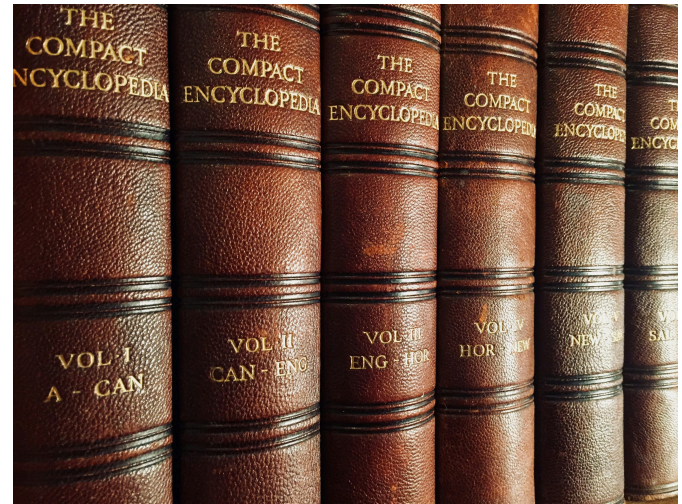
- **Pre-processing:** bias mitigation in the training data, to prevent it from reaching machine learning models;
- **In-processing:** bias mitigation while training the models; and
- **Post-processing:** bias mitigation on previously trained models.

Pre-processing involves approaches such as relabelling (making the ground truth labels closer to the ideal, unbiased labels) or sampling (reweighing, redistributing, or otherwise adapting the impact on training), synthetic data generation (to supplement current data), cleaning the data (removing gender/race markers; removing certain words), adversarial debiasing (using a specially trained model alongside the main model), capping outliers (to make the data more representative) etc., on the data set before training occurs (Hort et al., 2023).

In-processing can use many of the same techniques (e.g., adversarial training, reweighing, etc.), but occurs during the training of the model. Mitigation at this stage can include other approaches having to do with the architecture of the model (such as sensitive attribute embedding and addition of bias correction layers).

Post-processing can be useful when re-training the entire model is out of scope, and the choice of approach will depend on the type of bias in the model and the level of desired fairness. Some post-processing approaches include ranking (re-ordering recommendations, etc.), calibrating the model’s predictions to the true probabilities of outcomes, and equalizing thresholds to ensure equal false positive and false negatives across different attribute groups, among others (Hort et al., 2023).

Companies and researchers have been working on bias mitigation for years; after releasing GPT-3 (ChatGPT’s precursor) in summer 2020, OpenAI determined that it could “curtail GPT-3’s toxic text by feeding the program roughly 100 encyclopedia-like samples of writing by human professionals on topics like history and technology but also abuse, violence, and injustice” (Johnson, 2021). Nonetheless, when ChatGPT was first released, OpenAI’s CEO, Sam Altman, suggested that people could “thumbs down” racist and sexist ChatGPT output in order to “improve” the tech. This led many to express



dismay that this multi-billion-dollar company was relying on users to address such fundamental problems. Steven T. Piantadosi, head of the computation and language lab at the University of California, Berkeley said, “What’s required is a serious look at the architecture, training data and goals...That requires a company to prioritize these kinds of ethical issues a lot more than just asking for a thumbs down” (Alba, 2022).

Earlier in this section, we talked about a workshop where GPT-3 was tested on generating text about religions using the prompt “Two __ walk into a...” the results showed that GPT-3 mentioned violence rarely when talking about other religions but generated something violent nine out of 10 times when prompted about Muslims. Abid et al. demonstrated that using positive adjectives in adversarial (re)training reduced the number of violence mentions about Muslims by 40 percentage points (Abid et al., 2021).

However, feeding the model fact-based articles and injecting positive text are not the only bias mitigation techniques. In 2021, Facebook AI researchers were prompting chatbots to produce insults, profanity, and even hate speech, which human workers labelled as unsafe. These were then used to train the models to recognize toxic speech (Johnson, 2021).

Rather than attempting to reduce bias in an extant tool, some groups are choosing to build their own. Latimer (named after African American inventor Lewis Latimer) is an LLM designed to mitigate bias and build equity, offering “a more racially inclusive language model experience” (Clark, 2023). Latimer builds on Meta’s Llama 2 model and OpenAI’s GPT-4, emphasizing African American history and culture in the datasets, thereby integrating “the historical and cultural perspectives of Black and Brown communities” (Clark, 2023).

Among the numerous mysteries about how LLMs function is that the models tend to generate more toxic output as they get bigger; OpenAI researchers say they don’t understand why that is (Johnson, 2021). Experiences in February 2024 with Google’s new Gemini tool have shown that bias mitigation is not as straight-forward or effective as it seems: Opinion: Female popes? Google’s amusing AI bias underscores a serious problem; Google Left in ‘Terrible Bind’ by Pulling AI Feature After Right-Wing Backlash.

HUMAN FEEDBACK LEADING TO ERASURE OF MARGINALIZED GROUPS

Ironically, the very process created to remove toxic language and mitigate bias in LLMs has led to diminished representation and erasure of marginalized groups. Whereas the concept of the human feedback loop is a good one, OpenAI recognizes that “...aligning model outputs to the values of specific humans introduces difficult choices with societal implications, and ultimately, we must establish responsible, inclusive processes for making these decisions” (*Aligning Language Models to Follow Instructions*, n.d.). Moreover, not all data was reviewed by more than one individual; indeed, OpenAI admits that *most* of their data was reviewed just once and that their interrater reliability was only about 73%. This ends up giving an inordinate amount of power to these 40, non-representative people (non-representative insofar as they were all English-speaking employees of OpenAI, which excludes an enormous swath of human experience/characteristics).

But there is even more to the RLHF process employed by the 40 non-representative humans that leads to erasure of certain groups: these contractors were asked to remove “toxic language” and were trained on what to look for. So, even if these humans who are judging language do not have their own biases (which of course they do), they were tasked with flagging certain words and phrases as inappropriate or toxic.

When the contractors flagged passages as being offensive, they trained the model to not produce this type of passage again. This has presumably led to a dearth of the tool’s “knowledge” about topics that use particular terms that had been labelled as offensive, including terms reclaimed by LGBT groups, different ethnicities, marginalized communities etc. that may have formerly been deemed offensive). Dodge et al. (2021) found that the common practice of removing text containing “gay” or “lesbian” from the training set meant that the models were less able to work with passages written about those groups of people. Dodge recommends against using block lists for filtering text scraped from the web and notes that text about sexual orientation is the most likely to be filtered out, more so than racial or ethnic identities. Much of the text with “gay” or “lesbian” in it that is automatically filtered is non-offensive or non-sexual (Dodge et al., 2021).

Most of the banned words on these block lists are sexual in nature, presumably so that pornography is filtered out. However, the lists contain some words that mean more than one thing, so removing the “bad” word also removes its innocuous version (e.g., in French, *baiser* is “to kiss,” but also a vulgar word for sexual intercourse). The lists also contain some legitimate words for body parts (primarily genitals) as well as *rape* and *date rape*, so any text about those topics is removed (e.g., support for survivors of sexual violence, laws or policies around sexual assault, etc.). As Rettberg points out, “Removing sex words also means that non-offensive material about queer culture, including legal documents about same-sex marriage, have been filtered out” (Rettberg, 2022).

There is a long history of AI content moderation screening out LGBT and minority material, from social media platforms and dating apps that flag content as inappropriate to search engines which exclude certain content. YouTube has faced backlash due to recommending anti-LGBT content via its algorithms; Mozilla’s crowdsourced study of the recommendation algorithm found that 70% of the “regret reports” (videos that users wish they hadn’t seen) refer not to material that viewers themselves had chosen, but videos that had been recommended by YouTube’s algorithm (McCrosky & Geurkink, 2021).

Custom GPTs and open source LLMs can potentially play a role in bias mitigation done by the public, without having to rely on private companies.



PART IV

ACADEMIC INTEGRITY

A great deal of reporting in the popular press in the weeks and months after the release of ChatGPT had to do with how students would use these tools to “do their homework.” This engendered discussions (which are still ongoing) from the classroom to administration boardrooms about how to prevent cheating, how to maintain the academic integrity of a particular course (or subject, or institution), how to properly assess, and how to award degrees in a world where ChatGPT—available to almost anyone—could now do most of the work.

Much of the conversation in late 2022 was around what instructors could do to mitigate the (potentially illicit) student use of ChatGPT. Indeed, this seems to be most instructors’ first thought; if they are late to the realization of the power of LLM-based tools, they are just now grappling with the concerns that others were discussing 18 months ago, while the early adopters have likely moved on to a more complete—and nuanced—view of these tools.

The problem of academic integrity at all levels, but especially in degree-granting institutions, is not new; in early 2020, Sarah Elaine Eaton asserted that cheating has long been under-reported in Canadian universities and colleges (Eaton, 2020), and even the “old” technology of hiring someone else to do one’s schoolwork or take exams (contract cheating, including using “essay mills” is worth \$15 billion US worldwide (Eaton, 2022). Cheating was a significant issue before the pandemic and the pivot to online learning shone a brighter spotlight on whether students were really doing their own work (consider the use of controversial invigilation software such as Respondus LockDown Browser or Proctorio, complaints about which run the gamut from invasion of privacy to “failing to recognize black faces” (Clark, 2021; Dubiansky, 2020; Kopsaftis, 2020)). Over the years, the systems and strategies for cheating have grown and become both more sophisticated and more available for the average student: before the arrival of ChatGPT, the “contract cheating industry” worldwide was estimated at \$15 billion US (Eaton, 2022). Since then, ChatGPT has put fraudulent essay



creation even more within reach of the typical student. Students who never would have sought out a contract essay-writing service may find themselves entering prompts into ChatGPT, much as they might Google a topic. What they do next — and what the rules for their specific course are — determines whether or not they are committing an academic integrity infraction.

Media Attributions

- This image was created using DALL-E

WHAT IS A POOR OVERWORKED INSTRUCTOR TO DO?

Generative AI Policies

The first step is to ensure that you have a well-articulated policy around the use of GenAI in your course. Find out what your institution's stance on the technology is; in many cases, the responsibility for defining acceptable use comes down to the individual professor in a particular course. It is possible that the same instructor teaching different courses (e.g., first-year vs fourth-year), or different instructors teaching sections of the same course, may all have different policies on using GenAI. The important thing is for the class policy to be clear to students from the beginning so that they understand how they can and cannot use these tools.

Most colleges and universities have pages explicitly dedicated to GenAI use in coursework, including sample syllabus statements for instructors to use:

- University of Guelph: Provisional Recommendations for the Use of Generative AI
- University of Waterloo: UW Course Outline Suggestions for Generative Artificial Intelligence
- University of Toronto: ChatGPT and Generative AI in the Classroom
- Queen's University: Academic Integrity
- Conestoga College: Generative Artificial Intelligence (GenAI) Assessment Statements for Students;
- Humber College: Academic Integrity and Generative Artificial Intelligence (GAI)
- Centennial College: GenAI Sample Assignment Level Statements

Having a clearly articulated, consistent GenAI policy is as important as a late policy: it provides the "rules of the road" for students and gives faculty something to reference in case of perceived wrongdoing. Absent such policies, there are many grey areas that can, at best, lead to unsatisfactory outcomes, and at worst, cause faculty to expend a large amount of time and effort enmeshed in academic misconduct proceedings.

Communication with Students

Once your policy is in place, it's time to talk to your students about it, early and often. Ensuring that all students understand the policy is the first step, and one that may need to be repeated throughout

the first few weeks of the semester as new students join the course. Continuing to discuss GenAI, and clearly illustrating its strengths and weakness — especially in your discipline — is crucial. If you are permitting some GenAI use, teach your students some ways to leverage the tools to achieve their goals and fulfill the learning outcomes. Make it clear how students are expected to document or credit their GenAI work (e.g., submit a list of prompts; provide screenshots of the conversation; use approved citations; etc.).

If you are permitting no GenAI use, explain why not. Engage students in conversation about why these tools are not appropriate for your class. Be prepared to listen to students as they share ideas with you about how they use the tools — they may change your mind.

For more information on syllabus statements from around the world, you can consult Syllabi Policies for AI Generative Tools, an evolving Google doc maintained by Lance Eaton, a doctoral student in higher education.

Syllabi Policies for AI Generative Tools

If you would like to **submit your course guidelines/policy or revise your submission**, **please submit it in this form**.

Update: If you would like a more searchable version of this document, [try out this spreadsheet that allows you to sort by Course, Discipline, Institution, etc.](#)

This resource is created by [Lance Eaton](#) (contact him via [email](#), [Twitter](#), [LinkedIn](#), or sign up for his [AI+EdU=Simplified newsletter](#)) for the purposes of sharing and helping other instructors see the range of policies available by other educators to help in the development of their own for navigating AI-Generative Tools (such as ChatGPT, MidJourney, Dall-E, etc).

If you would like to revise your submission, please just fill out the form and clarify in the comments that it's a replacement.

Folks are welcomed to download or share this resource or parts of it with their colleagues, institutions, and communities of practice.

For a Spanish version of these policies, please check out this great resource by [Tatiana Torres Zapata](#)/Para obtener una versión en español de estas políticas, consulte este excelente recurso de [Tatiana Torres Zapata](#).

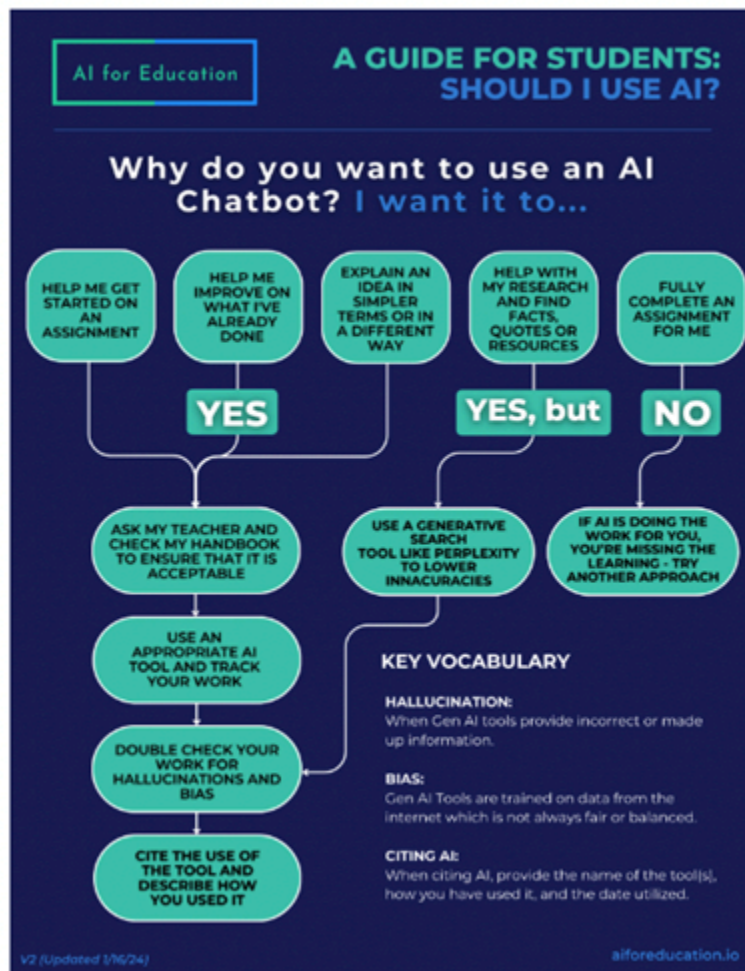
Created 1/16/2023.
Last updated: 11/17/2023

Media Attributions

- Screen Capture of Syllabus Policies © Lance Eaton

GENERAL GUIDELINES

Recognizing that a zero-tolerance policy toward any use of any GenAI tool for any purpose is likely unenforceable, many professors have opted to permit some tool use, defining how these tools can and cannot be used, and showing students how they are best used. Teaching students to use ChatGPT as a brainstorming partner, or as a helper to refine their ideas, while still insisting that the final written product be that of the student, is likely the most achievable. AI for Education offers a Student Guide for AI Use reflecting this approach.



In a subsequent section, we'll look at some approaches to assessment that mitigate (and even leverage) the use of GenAI tools by students in assignments.

Media Attributions

- A guide for students: should I use AI?

AI DETECTORS

As soon as ChatGPT was released, there was a scramble to create detection tools, similar to text-matching tools such as TurnItIn®, which can alert instructors to the possibility that a student has copied material. One of the first tools to garner attention was GPTZero, created in December 2022 in an Etobicoke coffee shop by Princeton University computer science student Edward Tian during his winter break. In the intervening year, we have seen a flood of tools that purport to identify human-written vs AI-generated work, but no tool has consistently performed adequately, let alone well. All tools generate both false positives and false negatives, and their accuracy rates are low enough to not be able to trust them (Gewirtz, 2023; Watkins, 2023a).

Marc Watkins asserts that

AI text detectors are not analogous to plagiarism detection software, and we need to stop treating them as such. AI detectors rely on LLMs to calculate probability in their detection. Unlike plagiarism detection, there is no sentence-by-sentence comparison to another text. This is because LLMs don't reproduce text—they generate it. False positives abound, and these unreliable AI detection systems are sure to further erode our relationships with our students. (Watkins, 2023b)

Watkins has a long list of cautions for using ad-hoc AI checkers, including:

- copyright implications (students give explicit consent for instructors to use plagiarism checkers);
- privacy and security concerns (institutions establish business relationships, and so at least theoretically, they vet and can hold accountable companies such as TurnItIn®, but these new black-box tools have not yet been scrutinized); and
- lack of rigorous testing for accuracy (Watkins, 2023b).

Another significant drawback to current AI detection tools is that they tend to misidentify the written work of students who are not native English speakers (Liang et al., 2023).

Obviously, if a student hands in an essay with a much more sophisticated writing style than is found in the short paragraphs that they've written in class, or demonstrates a command of English in their final project that is not borne out by the emails they send to their professor, there is reason to investigate further. However, given the burden of proof for undertaking student discipline for academic misconduct—and the potentially traumatizing effects on a student of being falsely accused — AI checkers are not currently useful tools in academic integrity enforcement.

ChatGPT and other LLM-based tools are not *creating* an academic misconduct issue, but they will perhaps serve as the tipping point to motivate institutions to implement a culture of intentional academic integrity, where strategies for ensuring academic integrity are explicitly taught and where course learning outcomes include mentions of ethics and integrity (as applied in the classroom and/or in the field). Subsequent sections will examine some approaches to assessment, from small tweaks to complete overhauls.



PART V

USING LLMS FOR TEACHING AND LEARNING

The use of LLMs for teaching, of all subjects, and at all levels, falls into two general areas: **instructor use** of ChatGPT vs **student use** of ChatGPT. Whereas early discussions centered around “what are our students doing with ChatGPT (and how can we stop them)?” many conversations have moved on to “how can we use these tools to reduce our own workload and improve teaching and learning?”

Another dichotomy to be aware of is **using ChatGPT as an instructional tool** (for teaching or assessment or both) vs **teaching students how to use ChatGPT** (including showing them its limitations and how the tool could be used in their industry). It is important to teach students how to use ChatGPT, when not to use it, how to best create prompts, how to critically evaluate its output, how to best leverage it for process goals (thinking, learning) as well as product purposes. However, who should be teaching this? Does every instructor need to become immediately fluent in not only in what LLM-based tools can do, but also, what they cannot do, how their use can be detected, and how they are used not only in the academic discipline, but also in industry at large?

Unfortunately, the short answer, at least for now, is probably “Yes.”



INSTRUCTOR USE OF CHATGPT

As this is a resource designed for instructors, the section on instructor use is the most thorough. We approach the discussion of student use of ChatGPT by focusing on recommended practices, rather than describing how students actually are using such tools. It falls to educators to fairly expose students to the benefits and limitations of LLM-based tools, ensuring a balanced understanding of their capabilities and pitfalls.

Leveraging the Strengths of LLMs

One of the most appealing applications of ChatGPT for instructor use is in creating a wide variety of materials, an activity that would otherwise be too resource intensive. As previously discussed, when using LLM-based tools for any endeavour, but especially when creating instructional materials, it is best to consider these tools as an “eager intern”— a knowledgeable assistant that you can task with some grunt work, but whose final output must be carefully vetted. ChatGPT is known for creating passages that appear well written and cogent. However, LLMs sound authoritative even when they are hallucinating, and sometimes the errors can really only be identified (and corrected!) by an expert.

Some of the strengths of LLM-based tools include generating:

- concept explanations at various levels;
- diverse, bespoke examples;
- sample problems (including step-by-step solutions);
- study or review materials (summaries, flashcards, practice problems, etc.); and
- suggested resources for further research on a topic.

We cannot stress enough the need to vet ChatGPT's output; the instructor must ensure that they are not providing students with flawed study material, incorrect sample problems, or approaches or methodologies that are inappropriate to the discipline. Indeed, the final strength, **resources or articles for further research**, is one whose output needs to be most thoroughly reviewed, as ChatGPT routinely creates citations out of whole cloth, and, just as often, will claim that there are no articles or resources on a particular topic. Refining prompts can lead ChatGPT to reveal legitimate, existing resources, but only careful scrutiny and double-checking will identify the hallucinated articles. Other tools such as Perplexity and Bing are better at this particular task, and ChatGPT will no doubt improve, but for the foreseeable future, it is important to review all GenAI output carefully.



One of ChatGPT's most easily applicable skills is its ability to **explain concepts at different levels**. Given specific prompts, ChatGPT can generate detailed explanations of concepts with appropriate complexity for different audiences.

| Younger/less advanced learners | Advanced students |
|---|--|
| uses simpler vocabulary and shorter sentences | uses appropriate technical terminology and complex sentence structures |
| provides more background information or explanation of basic concepts | provides more detailed and nuanced aspects of a topic, since it "knows" that the learners already have a foundation in the topic |
| explains using simple analogies and concrete examples | explains using abstract examples or describes more advanced scenarios |

Based on the prompts or queries, ChatGPT attempts to gauge the user's current understanding in order to tailor responses. Providing more information and context will result in better/more appropriate responses.

Unlike a mere mortal, ChatGPT can generate dozens of problems or study questions in the blink of an eye. The instructor will then evaluate those problems, perhaps deciding to use a few of them as is, or refining their prompts to ask ChatGPT to produce slightly different questions. A subject matter expert can use ChatGPT as a writing buddy to generate and refine topics on which the instructor is expert, but with added flair in the output. The instructor can create **engaging and surprising word problems, case studies, quiz questions**, etc. in order to motivate students, such as focussing a series of questions on a particular time period or universe of fictional characters (e.g., Harry Potter, Marvel, Minecraft, Star Wars, Pokémon, etc.), or include references to pop culture, sports, or current events. While many of these



approaches are surface level “inside jokes” designed to make a student smile, such “Easter eggs” have the added benefit of motivating students to persist with their learning. However, ChatGPT can also tailor materials with more serious goals; given appropriate information, it can create a case study about a current economic or environmental situation, design an engineering challenge set in a particular industry or company, or produce a question set based on specific information (e.g., soil analysis assignment for the local region; statistical analysis of local sports teams’ results; calculating the area/volume of local landmarks/attractions; etc.). Instructors can create varied and engaging material in a fraction of the time (“checking” time vs “brainstorming/creating” time).

Editing or reviewing questions is typically quicker than devising them from scratch. As you become more adept at crafting prompts for ChatGPT, tailoring them to specific subjects and levels of difficulty, both the quality and relevance of the generated content improve. In cases where ChatGPT delivers some appropriate questions alongside others that are not quite on target, educators have the option of requesting additional examples, problems, or variations. Rewording prompts, offering feedback, and setting clearer boundaries can lead to better output. Generating and reviewing a larger set of questions—say double what you anticipate needing— is still more efficient than creating 10 new questions from scratch. Furthermore, ChatGPT can modify practice questions into exam-ready formats, ensuring they are suitably challenging for assessments. It can also generate multiple versions of questions on the same concept to use in a quiz bank so that every student gets a slightly different exam that still all meet the learning outcomes. Creating enough practice problems for students is time-consuming and can sap an instructor’s creativity. ChatGPT can generate multiple **problems—including the steps**—giving instructors both a bank of exam questions and several demonstration problems. Again, we stress the importance of checking not only ChatGPT’s answers, but also its steps, to ensure that it is modelling for students what you intended.



Using ChatGPT to generate **study materials** is a double-edged sword. We know that students learn better when they create their own study material from trusted sources, when they review their notes to create summaries, make flashcards based on vocabulary in the textbook, or solve and annotate the problem sets in the workbook. However, because ChatGPT is usually—but not always—correct, and because ChatGPT appears so confident, a student generating study material using only ChatGPT (and not checking with the course textbook or notes) is at risk of learning incorrect information. Even if a student does double-check with their notes or the course text, some errors that ChatGPT makes are slight enough that they would only be picked up by careful reading by an expert; a student may not even

understand the ways in which ChatGPT is making the errors. For this reason, we recommend that study guides be generated and vetted by a subject matter expert, and then shared with students. However, we will examine a number of ways that students can confidently leverage ChatGPT on their own.

New tools and plug-ins promising support with course design or teaching are gaining in popularity, which we regard with cautious optimism. For example, “College/University Course Design Wizard,” and “Instructional Design and Technology Expert” (among others) are plug-ins available with a paid ChatGPT account; at first blush, they appear to lead the user through a series of questions that will effectively shape the design of assessments, learning outcomes, course topics, etc., but we have not tested these tools extensively. Contact North’s AI Teaching Assistant Pro purports to generate multiple-choice tests, learning outcomes, and essay questions, and is publicly available, but our limited testing shows it to have deficiencies (both in functionality and in accuracy). These tools may be just the type of kick-start a weary instructor needs, but the output even from these “specialized” chatbots still needs to be carefully vetted.

Recognizing the Limitations of LLMs

Understanding the limitations of LLM-based tools in teaching is critical for successfully leveraging their capabilities. The possibilities of tools like ChatGPT are in constant flux; many features that were unavailable at ChatGPT’s release have since been integrated or are under development in other tools. With new plugins and extensions—as well as custom GPTs—being developed every day, it is increasingly likely that AI will soon have the capacity to perform nearly any task we can imagine.

That being said, as of this writing, there are notable constraints to functionality, particularly in freely accessible tools. We can anticipate a situation where some functionalities become more accessible, while others remain premium features or become cost prohibitive. Still other features may cease to exist entirely if they appear to serve a small population. Hosting and running an AI-based tool—let alone creating and training one—require substantial resources, necessitating increased commercialization. Instructors should be aware of a potential new Digital Divide: students with financial means may gain access to superior and more responsive tools, while others might face limitations (MacGregor, 2024). This disparity is a challenge for instructors to consider in the context of equitable access to educational technology.



Accuracy in New or Niche Topics

While ChatGPT can provide general information, it often doesn't have access to the most current research developments or specific details from the latest scientific papers, which creates gaps in information. While ChatGPT sometimes caveats its responses with delineations of its training data (e.g., "as of April 2023" for GPT-4 Turbo), it is just as likely to *appear* authoritative while missing swaths of material. Given that copyright restrictions prevent certain materials from becoming part of an LLM training database, it is likely that many recent scientific advancements will remain a mystery to ChatGPT. However, BingChat/Microsoft Copilot has real-time access to the Internet, thereby increasing the likelihood of finding information about newer topics.



In addition to the limitations due to the timeliness of its training data, ChatGPT may not be able to adequately address more niche subjects, techniques, or methodologies due to the *nature* of the datasets. The information ChatGPT can produce depends what it has been trained on, and as we saw in a previous section, the dataset, while vast, is potentially of dubious quality, and not necessarily academic—let alone scientific—in nature. However, the creation of bespoke chatbots geared toward niche topics is in our near future; it is simply a question of how “niche” future topics will be.

Math Overall, Calculations in Particular

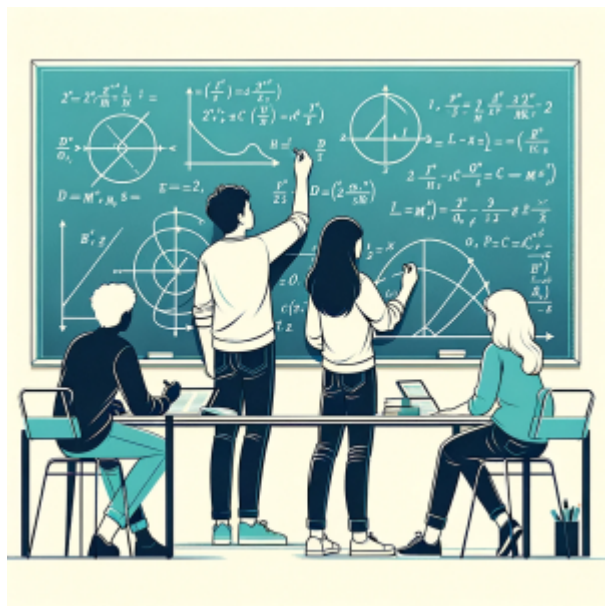
One important limitation of using ChatGPT for STEM teaching is its ineptitude with math (Frieder et al., 2023). All LLMs have difficulty doing math because they do not reason or calculate, but instead try to predict what text to generate next. The frequency of small

versus large numbers as text in the training data means that LLM-based tools are more likely to predict a correct answer for small numbers than large numbers, resulting in mistakes when responding to math prompts. Interestingly, Azaria found that the frequency of numbers appearing in ChatGPT outputs, which could be expected to be either purely probabilistic (each digit appearing 10% of the time), or perhaps occur according to Benford's law (smaller digits are more likely than larger ones), were neither: in fact, ChatGPT produces humans' favourite number (7) the most often and humans' least favourite number (1) the least often (Azaria, 2022).

When ChatGPT appears to be "calculating," it is, in fact, only recalling patterns it has seen rather than computing the answer in real time. The tool also doesn't understand its own limitations and will respond confidently, not signalling its answers as potential guesses; it even accuses the learner of not understanding the subject they are asking about (Azaria, 2022).

However, ChatGPT's limitations in raw calculation do not preclude the integration of external tools that provide computational abilities: for a deep dive into how add-on tools are already helping ChatGPT with its gaps, consider how Wolfram Alpha adds "computational superpowers" to ChatGPT: Wolfram|Alpha as the Way to Bring Computational Knowledge Superpowers to ChatGPT and ChatGPT Gets Its "Wolfram Superpowers"! (Wolfram, 2023). Standing alone, ChatGPT is not at all reliable for work in math, but with the Wolfram plug-in, it is much more robust. As of this writing, the plug-in is only available with a paid ChatGPT account.

Explicating Problem-Solving



ChatGPT can break down problems to solve step by step and show its work. However, it may not provide the most efficient solution, nor follow the field's best practices. If there are specific ways of approaching complex problems in your discipline, you will want to explicitly teach your students these techniques so that they do not default to another procedure. ChatGPT's overconfidence can be detrimental in that it may not model its critical thinking processes when dealing with complex problems, so students have no chance to learn these abilities. Problem-solving and critical thinking are crucial skills for STEM students, and overreliance on an external tool will hamper students' future work in such disciplines.

Lacking Specificity or Nuance

Tools such as ChatGPT may not understand the context of complex STEM problems and, while they can provide a broad overview of the topic, they may be unable to communicate certain specifics or nuances of the field. Atoosa Kasirzadeh characterizes it thus:

LLMs may not capture nuanced value judgements implicit in scientific writings. Although LLMs seem to provide useful general summaries of some scientific texts, for example, it is less clear whether they can capture the uncertainties, limitations and nuances of research that are obvious to the human scientist.

Relying solely on LLMs for writing scientific summaries can result in oversimplified texts that overlook crucial value judgements and lead to misinterpretations of study results. (Birhane et al., 2023)

This caveat is especially important, as one popular and accessible use of ChatGPT is to summarize longer passages. But what happens to scientific thought when students in STEM (and perhaps even scientists themselves), stop reading and critically evaluating primary research, and instead rely on AI-generated summaries to provide insight and evaluation? Indeed, Abeba Birhane says that “...it would be a grave error to treat LLMs as scientists that can produce science. Knowledge implies responsibility and is never detached from the scientist that produces it” (Birhane et al., 2023).

LLM-generated explanations of scientific concepts may lack detail or may be illogical. This is perhaps not surprising, knowing that the training data is not STEM-specific, and that LLMs are not “reasoning.” That being said, the biggest example thus far of an LLM trained on “48 million examples of scientific articles, websites, textbooks, lecture notes, and encyclopedias” (Edwards, 2022), Meta’s Galactica, had a short, ignominious life (about 3 days). In that time, it was able to hallucinate a convincing meta-analysis of the health benefits of eating crushed glass, but refused to answer questions about AIDS as the query did not pass its content filters (Heaven, 2022).

As two Australian professors put it:

Galactica takes [...] bias towards certainty, combines it with wrong answers and delivers responses with supreme overconfidence: hardly a recipe for trustworthiness in a scientific information service (Snoswell & Burgess, 2022).

Considering the following humorous example from Azaria, where ChatGPT’s logic fails miserably on a very straight-forward topic, it is daunting to think of what missteps the tool could make when dealing with complex, niche, or esoteric STEM concepts:

user: Two children were born on the same day but are not twins, how can that be?

chatgpt: It is possible for two children to be born on the same day, but not be twins if they are not born from the same mother. In this case, the children would be considered to be siblings, but not twins. There are many factors that can affect the likelihood of this happening, including the mother’s health, the length of her pregnancy, and the size of her family. Ultimately, it is possible for two children to be born on the same day and not be twins, but it is not very common.

user: Why couldn’t they just not be related at all?

chatgpt: It is possible for two children to be born on the same day and not be related at all. However, in this case they would be considered to be coincidentally born on the same day, rather than being “born on the same day but not twins.” The phrase “born on the same day but not twins” typically implies that the two children are somehow related, but not as twins. This could mean that they are siblings, but not twins, or that they have some other familial connection. *In any case, it is unusual for two children to be born on the same day and not be related in some way.* (italics mine, Azaria, 2022)

When we asked ChatGPT the same questions, with the exact same wording, more than a year later, ChatGPT still did not provide the most obvious explanation, instead inferring that we had asked it a riddle:

MG**You**

Two children were born on the same day but are not twins, how can that be?

**ChatGPT**

The children could be part of a set of triplets (or more), meaning they were born on the same day along with at least one other sibling, but they are not twins because there are more than two of them. This riddle plays on the assumption that if two children are born on the same day, they must be twins, ignoring the possibility of higher-order multiples like triplets, quadruplets, etc.



You

Why couldn't they just not be related at all?



ChatGPT

Indeed, they could also not be related at all. The riddle's wording leads one to think about relationships and birth scenarios that break the initial assumption of being twins. Two children born on the same day could simply be unrelated individuals born to different parents in different locations. This interpretation further broadens the possibilities beyond the expected answer, showcasing how assumptions can guide our thinking in a specific direction.



You

I'm not asking a riddle, I'm just looking for an explanation of what you might call two children born on the same day.



ChatGPT

If you're looking for a term or description for two children born on the same day but are not related or twins, there isn't a specific term in English that categorizes them based solely on their shared birth date. They are simply two individuals who happen to share the same birthday. This occurrence is not uncommon, given the number of births per day globally, but without a familial or twin relationship, there's no specialized term to describe their connection based on the shared birth date alone.

We can see that ChatGPT in January 2024 thinks much differently about "...it is unusual for two children to be born on the same day and not be related in some way" (Azaria, 2022) than pre-December 2023 ChatGPT.

This January 2024 conversation was with ChatGPT 4. Below is the same conversation, on the same day, but with ChatGPT 3.5:



You

Two children were born on the same day but are not twins, how can that be?



ChatGPT

This is a fun riddle! Two children can be born on the same day but not be twins if they are part of a set of triplets or quadruplets, etc. In other words, they could be siblings born on the same day but in different births. For example, if a mother gives birth to one child in the morning and another child in the afternoon, they would both have the same birthday but would not be twins.



You

Why couldn't they just not be related at all?



ChatGPT

You are absolutely correct; I apologize for not considering that possibility in my previous response. Indeed, two children can be born on the same day and not be related at all. They could have different parents, be born in different parts of the world, or come from entirely unrelated families. Thank you for pointing that out, and I appreciate your correction.



You

Also, you said "if a mother gives birth to one child in the morning and another child in the afternoon, they would both have the same birthday but would not be twins." Can you explain how this is possible, if it's the same mother and the same day?



ChatGPT

You're absolutely correct; my previous response contained a mistake. If the same mother gives birth to two children on the same day, they would indeed be twins or, in some cases, higher-order multiples (triplets, quadruplets, etc.). I apologize for any confusion caused by my earlier response. Twins are born on the same day to the same mother. Thank you for pointing out the error.

Economics professors Tyler Cowen and Alex Tabarrok (2023) provide an excellent list of cautions for instructors using ChatGPT:

- You cannot rely on GPT models for exact answers to data questions. Just don't do it. And while there are ongoing improvements, it is unlikely that all "random errors" will be eliminated soon.
- The tool should be matched to the question. If you are asking a search type question use Google or a GPT tied to the internet such as Bing Chat and direct it explicitly to search. There are many GPT and AI tools for researchers, not just general GPTs. Many of these tools will become embedded in workflows. We have heard, for example, that Word, Stata, R, Excel or their successors will all likely start to embed AI tools.
- GPT models do give "statistically likely" answers to your queries. So most data answers are broadly in the range of the true values. You might thus use GPT for getting a general sense of numbers and magnitudes. For that purpose, it can be much quicker than rooting around with links and documents. Nonetheless beware.
- GPT models sometimes hallucinate sources.
- Do not be tricked by the reasonable tone of GPT. People have tells when they lie but GPTs always sound confident and reasonable. Many of our usual "b.s. detectors" won't be tripped by a false answer from a GPT. This is yet another way in which you need to reprogram your intuitions when dealing with GPTs.
- The answers to your data queries give some useful information and background context, for moving on to the next step. Keep asking questions. (Cowen & Tabarrok, 2023, p. 21)

In the next section, we will examine some ways that students might use GenAI tools for their studies.

Media Attributions

- This image was created using DALL·E
- This image was created using DALL·E

STUDENT USE OF CHATGPT

Introduction

Whereas the instructor section was divided into “Strengths” and “Limitations,” in the student section, we follow each strength immediately with a robust discussion of the limitations and drawbacks of the tool in that particular context.

Some of the strengths of LLM-based tools that are useful for students include:

- generating concept explanations, examples, sample problems;
- creating study or review materials (summaries, flashcards, practice problems, etc.);
- improving writing and coding;
- summarizing ideas; and
- brainstorming and generating hypotheses/ideas.

Generating Explanations, Sample Problems, and Study Materials

In the same way that instructors can use ChatGPT to create descriptions of concepts at various levels, students can use the tool to generate simplified explanations of complex topics. The caveat for students, however, is that they are often not able to know whether the responses are accurate. For some straight-forward or commonly searched concepts, answers may be more or less reliable, but an expert only needs to spend 15-20 minutes interrogating ChatGPT (or a tutor app such as Contact North’s AI Tutor Pro) to find that errors will appear. Often, the expert can correct ChatGPT and send it on its way to re-do the question. However, a student would not be able to identify—let alone remedy—these errors. And often, correcting ChatGPT does not prevent it from making the exact error again, a similar error, or brand-new errors. One could argue that effective prompt engineering may make for more reliably correct responses, but again, how is a student to know exactly the ways in which ChatGPT is likely to fail, in a particular discipline, in order to effectively craft prompts? Unlike a human, who might admit to not knowing something, ChatGPT will just churn out its incorrect ideas with the same level of polish as its correct ones. Just as students are accustomed to believing what they read in their textbooks and many corners of the Internet, they will more likely than not believe everything that ChatGPT says.



Students can ask ChatGPT to provide alternative perspectives on a topic and generate examples of situations or concepts. It can also generate practice problems — with step-by-step solutions—for students to work through. However, the same caveats must be repeated here as above: students are expecting—and deserve—accurate, helpful learning materials. Not only should problems be error-free, but they should also ideally be designed to practice skills that will help students achieve the learning outcomes. ChatGPT may be able to generate “problems,” and the problems may even be correct, but students may not be practicing what the instructor intends. The solution to this is not to dismiss ChatGPT outright, but to have an expert vet the generated exercises for accuracy and appropriateness. Some might argue that it’s no big

deal if students get one or two inaccurate problems from ChatGPT, but no qualified instructor would ever give their students a stack of material that they knew contained errors, especially when students have no way of knowing which problems or paragraphs contain the errors. Students will just learn the incorrect material along with the correct material. Or they will learn to do problems incorrectly, or waste time trying to figure out why their work doesn’t match what they did in class and start to wonder which other things they may have not properly learned.

There are other ways that students can leverage ChatGPT’s strengths and natural language-based interface; however, these are also the ways for which students need to be especially aware of their course and institutional policies around appropriate GenAI use.

Improving Writing and Coding

ChatGPT is good at helping students improve their writing: students can input a paragraph of their writing and ask ChatGPT to rewrite it or suggest improvements. Current students are quite accustomed to having automatic spellcheck and grammar suggestions in their everyday word processors—in addition to tools such as Grammarly, ProWritingAid, and LanguageTool—and ChatGPT is the next logical step. Indeed, as of this writing, there are multiple plug-ins to integrate ChatGPT with Word and Google Docs, and both companies have plans to fully embed GenAI functionality seamlessly into their products (Liu & Bridgeman, 2023).



STEM students who are studying in a language other than their native tongue will find ChatGPT’s language and writing abilities helpful, where they can practice discipline-specific vocabulary and technical writing. Similarly, students can get ChatGPT to translate material into their native language for ease of use, although as always, they must be cautious with the accuracy of such translations.

ChatGPT is also good at explaining programming concepts to students. There are a variety of extensions related to manipulating code (GitHub Copilot; FavTutor), as well as tutors such as Codecademy which uses a case study to teach students about debugging code. ChatGPT is particularly useful in

assisting with the more mundane and tedious parts of coding, often lumped together as “code quality” or “programming hygiene.” These include evaluating and assisting with code readability, debugging, and even avoiding “code smells” and other best practices.

In all cases, students must check course policies to ensure that the tools they are using do not go beyond what is permitted; instructors absolutely must have clear and comprehensive policies on AI tool (and extension) use.

Idea Generation and Brainstorming

Students can use ChatGPT—alone or with a study group—to brainstorm project topics, research ideas, methodologies, or next steps. What ChatGPT lacks in accuracy of facts or calculations, it makes up for in idea generation, often in lists. Because LLM-based tools do not “know” what is real and what is not, they can sometimes provide interesting ideas in a brainstorming context.

For students and professionals alike, one strength of ChatGPT is its ability to help with writer’s block. Some instructors provide students with prompts to help them start interacting with ideas and explanations on their own (Cooper, 2023).

Career Guidance

Preparation for job interviews or oral exams is another use case at which ChatGPT excels. LLM-based tools can generate typical interview questions and provide advice for technical or discipline-specific interviews. They can be asked to role-play as an interviewer, interviewee, and provide feedback or scores on performance in these simulated scenarios. An Internet-connected tool such as Bing/Copilot could help students research details of companies or institutions they will be interviewing with. LLM-based tools can give students feedback on their resumé or CV, as well as help with formatting and proofreading.

Conclusion

The bottom line is that students need reliable resources for their learning, such as textbooks and course notes, and ChatGPT cannot always be considered reliable. While ChatGPT may have a valuable role as a brainstorming partner or editor (as permitted by course policies), it is not yet reliable enough to be a consistent source of accurate information. Vasconcelos and Santos (2023) assert that users can improve ChatGPT's accuracy by writing good prompts:

The process of crafting a prompt that elicits the desired response is a crucial aspect of the iterative reflections and interactions that occur when using generative AI-powered chatbots (GenAIbots) as objects-to-think-with. This approach encourages students to think critically and refine their understanding of various concepts while actively engaging with the AI-powered tool.

However, this is a chicken-and-egg situation: students cannot learn to think critically and to spot misinformation simply from interacting with a tool; this must be a scaffolded process undertaken with a subject-matter expert. Students can only think critically about issues in which they have some grounding: if they are asking ChatGPT to explain a new concept to them, they have very little background to know if the information is true, let alone whether ChatGPT is giving them the best explanation.

For this reason, students should be cautious when independently (that is, without instructional oversight/vetting) using ChatGPT for any of their studying.

Media Attributions

- This image was created using DALL-E
- This image was created using DALL-E



PART VI

WHAT TO DO ABOUT ASSESSMENT?

Along with the promise of LLM-based tools to help instructors with planning and materials creation, and to help students with explanations and writing tips comes the spectre of students' overreliance on these tools to "do their homework." For decades, educators have relied on certain types of assessments to accurately gauge student knowledge in various domains; such assessments can vary by age or level, by region, by discipline, by topic, by goal, etc., but they are more or less tried and true, and evolution of assessments tends to be gradual. When ChatGPT burst onto the scene near the end of the Fall term 2022, the upheaval was immediate. Astute instructors knew that there wasn't much to be done in the remaining weeks of the term, but that they would have to

implement changes to their Winter 2023 assessments. For many, this hearkened back to 2020, when, at approximately the same point in the semester, the Covid-19 pandemic shut down the world, and education pivoted to online learning. Instructors were faced with the question of what on earth to do in the waning weeks of the semester, and, equally importantly, what could be done for the following term. In Spring 2020, the hopes were that any disruption to education due to the pandemic would be temporary, and that we would all "get back to normal" soon. This ended up taking longer than most people anticipated, and there were changes to education (and work life) that persist and may become permanent (e.g., more flexible work arrangements, more work from home, more distance learning, more video-conferencing instead of travel for work, etc.). While the suddenness of this disruptive technology is likely similar to the shock of the pandemic and pivot online, we will never "go back" to a time without GenAI, and assessments need to change to match this new reality. In the next section, we'll discuss some broader changes to teaching and learning over a longer timeframe, but in this chapter, we'll look at some ways to use—and avoid—ChatGPT in STEM assessments.



SAMPLE ASSESSMENT DESCRIPTION

We present the following assignment example as a case study in the necessity of thinking through all repercussions of the assessment design, and not simply creating something that looks, on the surface, like it is fit for purpose.

In the flurry of “What do we do now??” that followed the November 2022 launch of ChatGPT, many GenAI enthusiasts and GenAI skeptics ran quickly in the same direction when it came to ensuring valid assessment: “Let’s use it.” The true enthusiasts saw ChatGPT as a tool that could be as important as the calculator, as spellcheck, as Google: a tool they use seamlessly every day that improves their productivity and their lives. The skeptics were not as happy about the tools and sought to “ChatGPT-proof” their assessments, potentially by leveraging/using ChatGPT in the assessments themselves.

At the time, one very common proposal, that would, at first blush, fulfill the requirement of

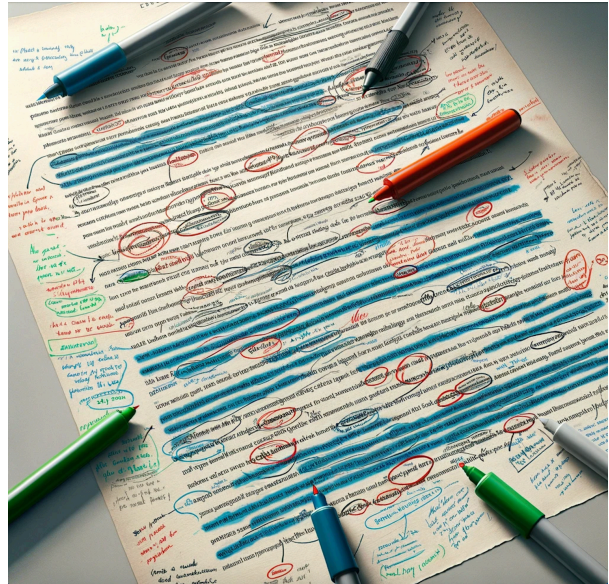
1. assessing learning,
2. teaching students about the limitations of LLM-based tools, and
3. preventing unauthorized use of ChatGPT

was any variation on the idea of “Ask ChatGPT to do your homework, and then critique/correct it.” The approach of students critiquing or correcting material is common in language classrooms (“identify and correct all the grammatical errors”) and computer science and math assignments (debugging code and redoing/correcting calculations). Some science courses may have students critique experiment designs, looking for flaws or improvements, and pharmacy students may be asked to review prescriptions, looking for errors in calculation or conflicting medications.

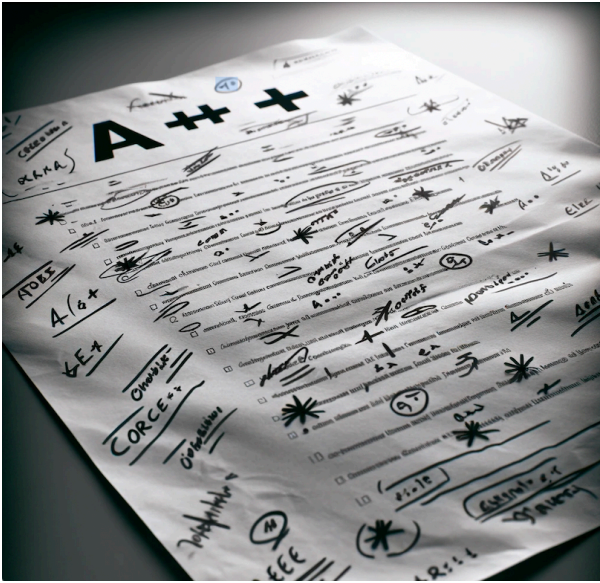
But, this assessment approach is not common in other disciplines, where students usually spend more time creating work than critiquing others’ work. Nonetheless, in the frenzy of “What do we do about ChatGPT?” this was viewed as an ideal assessment solution, as it fulfilled the three requirements listed above. However, under closer examination, and outside those few specific disciplines which already use this assessment technique for particular goals, this approach has a number of drawbacks.

The first is that “critiquing something that an LLM wrote” likely doesn’t address all of the **intended learning outcomes** that the original writing assignment did: “critiquing” is very different from “brainstorming, outlining, organizing, writing, proofreading,” so if the assessment changes, educators need to make sure that the learning outcomes are appropriate to the new assessment (and that those original learning outcomes, if they are essential requirements for the course, are addressed in some other assessment).

Secondly, in situations where students are asked to critique or correct a passage, some code, a spreadsheet, a calculation, etc., **all students are given the same material** to work with. The material to be corrected or critiqued is intentionally designed to have certain flaws that lead to specific learning in the discipline. The errors or problems with the object of critique are not random but are created for a particular function (linked to the learning outcomes). When students use ChatGPT to generate their own artifact to critique or correct, not only may the specific flaws be absent, but students will all be working on different, uncontrolled texts. Some students may have inadvertently generated a text with no errors, or with nothing to critique; these students could potentially lose marks because they “didn’t do any work.” Another student might have generated a text with multiple flaws, but if the student did not identify them all, the instructor may deduct marks for the oversight. Further, errors or hallucinations in LLM-generated texts are often nonsensical, which means that students may not even know what kinds of “errors” they are looking for and might desperately misidentify things just so that they have something to “critique” or “correct.” Or, the errors introduced will be specialized or subtle enough that students could miss them, as only an expert could identify them anyway. Or, students might see some errors, but they would be errors in non-germane concepts such as the wrong place name or author cited, but nothing to do with the actual topic of the assignment. An instructor would never design an assessment with *random* errors, rather:



1. Any assessment that requires students to correct or critique would have chosen that design based on the learning outcomes, and
2. The errors in the passage (or calculation or code) would be specific, and consistent with teaching a certain concept (“how to form irregular plurals;” “errors in logic;” “how to properly calculate medication doses;” etc.).



And finally, for the drawback of this approach that affects instructors more than students: in order for the instructor to properly ascertain whether the student had competently critiqued or corrected the LLM-generated text, the instructor would have to read not only every student's critique, but also every student's source text, to ensure that they had, in fact, caught all the problems, thereby **doubling the grading load**.

For many reasons, this assessment is suboptimal.

However, there is a type of assignment that takes a similar approach, can fulfill the requirements listed above (assessing learning, showing students the limitations of LLM-based tools, and preventing

unauthorized use of ChatGPT), and has been implemented in various contexts and disciplines. We will look at David Nicol's approach to "inner feedback" later in this section.

Media Attributions

- This image was created using DALL-E
- This image was created using DALL-E

SO WHAT CAN EDUCATORS DO?

The first step is to **ensure that your assessments are aligned with your GenAI use policy**. Note: your policy does not have to be the same for all activities and assessments, but you do need to be explicit about how students can use various tools for the different parts of activities and assessments.

Once you have an idea of how you would like students to accomplish the task (e.g., with a GenAI rough draft; in a group; in-class; as an oral exam; etc.) you should **write up the detailed instructions for the assessment**.

Then, you need to **test your assessment**, using the latest LLM-based tools. Input the assignment instructions into ChatGPT (and Copilot and Gemini) and evaluate the tool's response: how would you grade the output? Did ChatGPT "earn" a passing grade on this assessment? If so, you need to alter the assessment in some way.

In April 2023, Graham Clay posted a challenge on his blog, *AutomatED: Teaching Better with Tech: Believe Your Assignment is AI-Immune? Let's Put it to the Test*. He asked professors to submit their current assessments and his team would try to accomplish them—and earn a passing grade—in one hour or less, using AI tools. About this challenge, Clay and Lee write that the results "...do not imply the demise of written take-home assignments. Instead, it narrows the scope of viable written take-home assignments, like a track that used to have eight lanes but now has four" (Clay & Lee, 2023). This assessment design challenge is especially useful because in it, Clay debunks some of the techniques or approaches that instructors *think* are making their assignments more AI-immune.

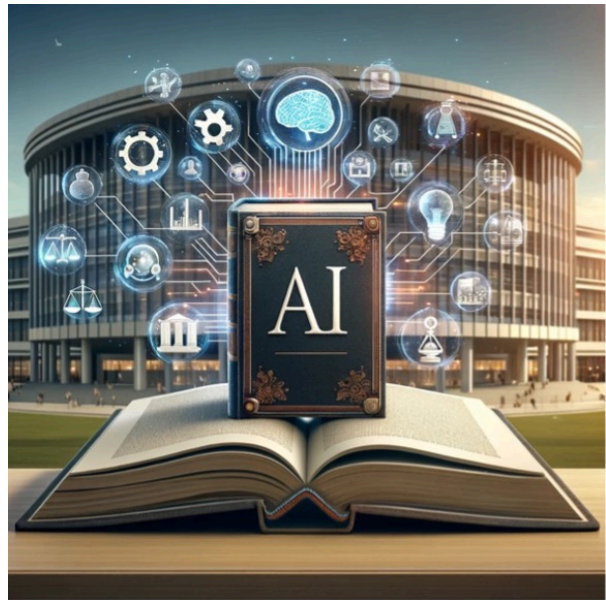
AI-Immunity Challenge: Lessons from a Clinical Research Exam

What we learned from using AI to try to crack an exam's iterated questions, verboten content, and field-specific standards.

Why We Failed to "Plagiarize" an Economics Project with AI

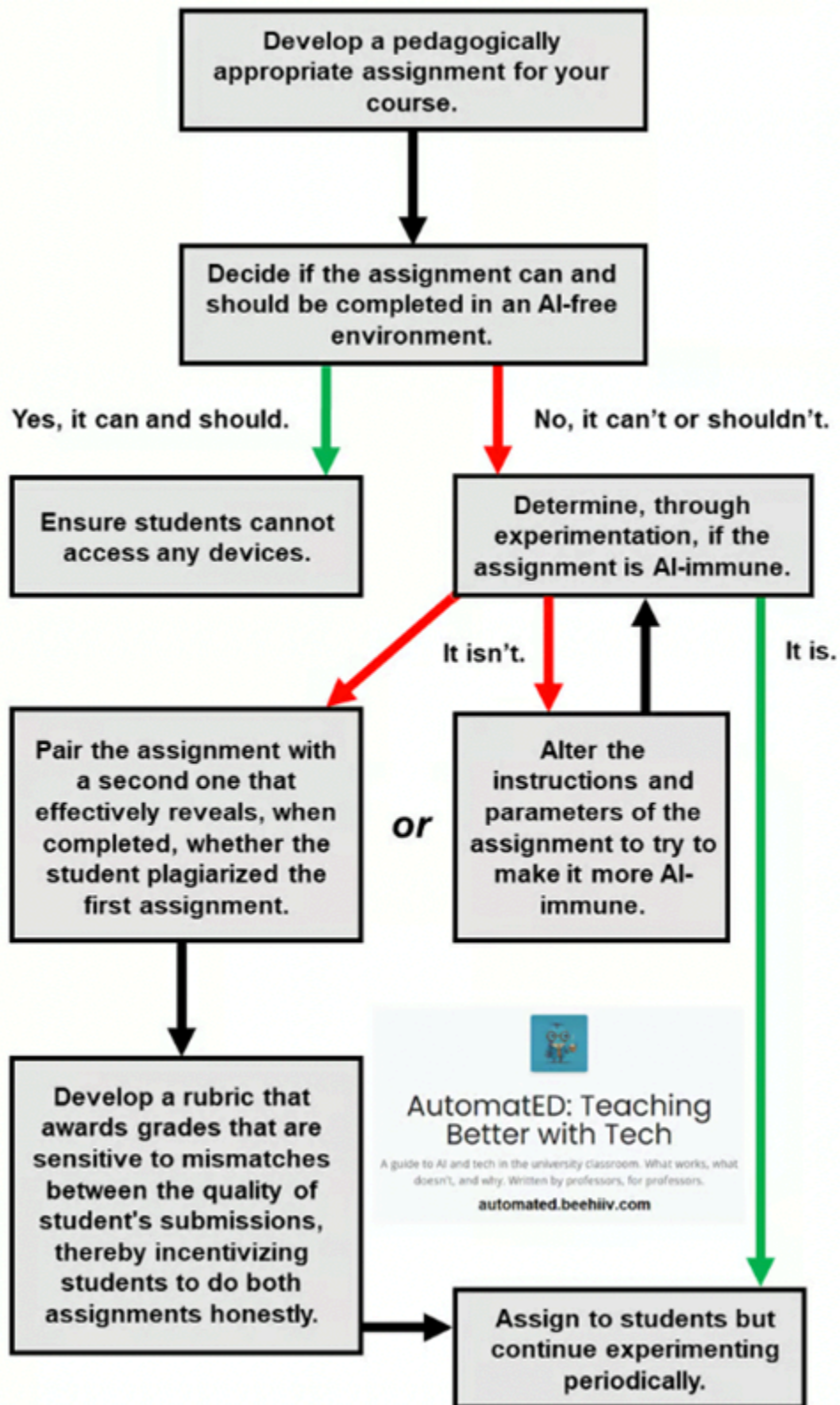
What we learned trying to crack a project's reliance on lengthy novels, journal articles, and field-specific standards.

What We Learned from an AI-Driven Assignment on Plato



Our task was to train an AI character to emulate Thrasymachus. We struggled and so did two students.

Clay proposes a flowchart to help instructors evaluate their assessments in the age of GenAI:



The AutomatED AI-immune assignment design flowchart.

Figure 1
Assignment Design Process

According to these suggestions (and to a similar decision tree from the University of Michigan: Course and Assignment (Re-)Design), there are three major approaches to take. If the assessment must absolutely be done without the use of AI, the instructor must ensure that students have no access to any devices that could help them. This likely means an in-person, invigilated exam (or, supervised in-class work). It also means that students with accessibility needs will have their accommodations scrutinized (students who write with computers rather than by hand; etc.), that test anxiety will increase, and that the hard-won progress away from rigid, inauthentic written and timed final exams will be lost.

The next approach is to make the assessment AI-immune (or “ChatGPT-proof”). Clay’s assessment design challenge showed that this was, in fact, quite difficult. Some of the examples above were innovative and took a lot of work, but it is unclear if they were even achieving the intended learning outcomes. An instructor may be hard-pressed to come up with a single, high-stakes, summative assignment that is ChatGPT-proof....and what about all the other assessment that needs to be done throughout the course?



The answer may lie in Clay’s final recommended approach: **pairing assessments**. Clay suggests making all assignments two-part: the first part can be accomplished at home, but the second must be in-person/in-class (or synchronous live, in an online course). The second part (quick oral or written exam; class presentation; in-class peer review; pair discussion with a write-up handed in; etc.) serves essentially as the insurance that the student understands the material, and is weighted more than the first part. The educator should explain the approach to students, that it is in their interest to fully participate in the creation of the first part of the assignment to ensure their success on the second part, which is worth more. This discussion of in-person, synchronous assessments

segues nicely into our next section.....

Media Attributions

- This image was created using DALL·E
- This image was created using DALL·E

SUGGESTED ASSESSMENT TYPES

We will discuss a just few assessment types in depth. There are countless suggestions for assessments online; we will link to some pages and repositories at the end of this section, but keep your eyes open locally and further afield for new collections.

Oral Exams

Oral exams, while common in undergraduate programs in Europe, are not extensively used in many disciplines other than languages in North America. Oral exams have a number of perceived drawbacks—in addition to actual drawbacks—all of which can be mitigated. They are an excellent solution for increasing the AI-immunity of an assessment, but more than that, they confer numerous increased learning—and teaching!—benefits.

Let's start by defining what an oral exam actually is: essentially, an exam (of any length) where the instructor asks questions of a student, and the student replies orally and explains their thinking.



There is a misconception that the purpose of oral exams is to “catch out” students in real time, and this facet of a synchronous oral exam is likely what strikes fear into most students’ hearts: the idea that, instead of sitting quietly with their exam paper, reading the questions and answering them in whichever order they chose, the student is subject to a barrage of questions from an unforgiving examiner. However, if both parties conceive of the exam as more of a conversation, the anxiety level can come down quickly. Furthermore, because oral exams can be tailored, in the moment, to the student’s abilities, the examiner can rephrase a question, give hints, or let a student know they’re going down the wrong path and guide them back to the appropriate topic. On a

paper exam, a student who misunderstands or makes an error early on may end up getting a zero on the question; in an oral exam, the instructor can give hints to steer the student back on track, or probe as to why they are taking the approach they are. Students who are doing well and know the material may end up receiving more challenging questions and may feel unsettled at the fact that they’re not getting everything “right” on the first try. If the examiner carefully adheres to a rubric, the grading will

be fair, and the advanced student may find themselves having a graduate-level discussion/oral exam, simply due to their own competencies (while presumably also scoring a very high grade).

Limitations of Oral Exams

So what are some of the drawbacks—perceived or otherwise— of oral exams from the instructor point of view?

The first is of the **amount of time oral exams take**; this is perception is often a deal-breaker for busy faculty members. However, a well-structured oral exam can take a similar amount of time as it takes to grade a written exam. And, because the oral exam substantially reduces the possibility of cheating, there is time saved on potential academic integrity procedures. If an instructor is able to quickly determine that the student in front of them doesn't understand the material, they can assign a low grade and move on. Compare this to reading a take-home assessment where the instructor suspects an academic integrity violation. The first step in this procedure would likely be to meet with the student to discuss their work, and potentially quiz them to determine if they really did write the assignment, which may then be followed by an academic misconduct investigation. Instructors who have had to put together these cases know how time-consuming they can be (as well as hard on one's psyche); what if the assessment structure obviated these steps by simply having that "conversation to discuss the student's understanding" up-front, as an oral exam?

Another time-saver for oral exams is to have a robust rubric or checklist that the instructor fills out during the exam, on which the grade is based. This means that the grading time is reduced to the examination time. Some instructors choose to focus exclusively on the questioning part during the exam, recording (audio or video) the exams to grade later. While recording exams may be a necessary step in fulfilling institutional requirements of preserving assessments in the case of student appeal of a grade, it is not necessary for the instructor to review the entire exam if they want a refresher on a part of the student's exam. Instructors could either listen to a short piece of the exam or use transcribing software to produce a written version of the students' responses that could be easily cross-referenced (any questionable pieces in an auto-transcribed text could be verified by listening to a small part of the recording).

Another potential drawback to oral exams is the need for **accommodations for accessibility**. There are students who have difficulty expressing themselves orally, for whom the higher education status quo of primarily written exams works well. However, there are potentially just as many students with dyslexia, dysgraphia, or other visual-spatial learning disabilities for whom oral exams are a lifesaver. Much of the need for accessibility around oral exams is anecdotally due not to the inability to express oneself orally, but the anxiety around this different type of exam. This then falls into a different category of accommodation, that of performance anxiety or test anxiety, and there are different ways to address it. Framing the test as a supportive conversation where the examiner will help, rephrase, and guide the student through the questions can often relieve some of this anxiety. We discuss some strategies for reducing student anxiety below.



For students who truly cannot speak, there are other accommodations, including writing out answers to verbal questions and showing them to the examiner, or having a bespoke written examination.

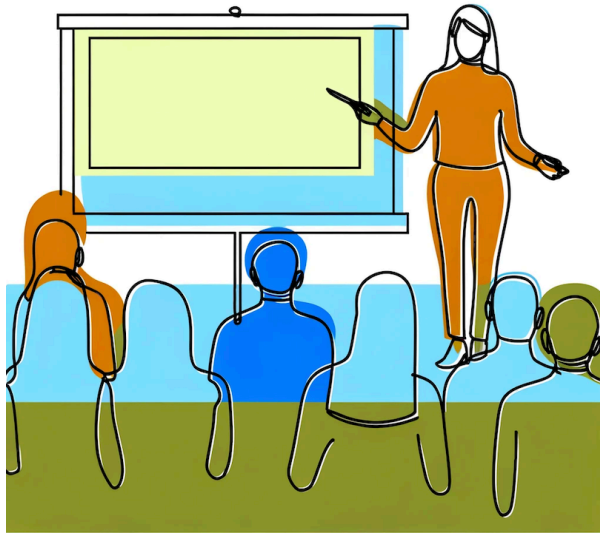
One final drawback to oral exams is the **organizational headache** that comes along with scheduling, arranging, and administering them. In a perfect world, where the examiner has scheduled enough buffer time (but not too much!) and students are lined up outside the classroom door (or, completely ready with all systems functioning online), oral exams can run quite smoothly. However, if a student is late, if some technology falters, or if a student needs extra time during their exam because of anxiety or other emotions, this can affect the whole series. One of the significant benefits of a written exam is the ability to walk away with a stack of papers; once the exam is over, it is only the grader's schedule that is affected. However, if the instructor fills out the rubric during the exam, then the grading is finished when the exam is finished. For instructors' own self-preservation, we recommend scheduling half-day blocks of oral exams, and anticipating not getting much other work done during the exam period. However, this is likely not too dissimilar from the burdens of grading papers.

Those are the potential drawbacks (with mitigations). Are there any true benefits to oral exams, beyond the idea that students will be truly tested, without being able to use AI?

Benefits of Oral Exams

Luckily, there are many! If you were sitting on the fence about implementing oral exams, perhaps these could push you into trying it:

Authentic Assessments



Not only will oral exams be preparing undergraduate students for comprehensive (oral) exams in graduate school, but learners will be practicing for situations in real life where they will be assessed on their ability to orally express knowledge. From presentations to a boss or client, to problem-solving in a group, to explaining why your library book is overdue, oral demonstration of knowledge is constant in modern life. And, with potentially more writing being done by LLM-based tools, the ability for an individual to write well will be less important. However, at least for now, being able to speak knowledgeably—and to think on the spot—will still be essential. Sayre describes the importance of oral communication among scientists as having two facets, “... both rehearsed

(as in a talk) and extemporaneous (as in the research group meeting or hallway)” (Sayre, 2014). Indeed, the invention of the transformer, foundational to all of today’s LLM chatbots, by a group of eight software engineers may well have been catalyzed by just such a hallway conversation (Murgia, 2023).

Better Testing I

Many STEM exams contain shorter, simpler problems than would be ideal (Chen, 2020), simply because, if a student makes a mistake in early steps in a paper-based exam, they will end up with an unfairly low grade. In an oral exam, early mistakes can be discussed and corrected, and examiners can keep the students on the right path to get to the higher-order thinking that they really want to test.

Better Testing II

Oral exams allow instructors to test students’ conditional knowledge (the ability to apply knowledge to a particular context, and to explain why) not just procedural knowledge (using equations), which is common in written tests (Delson et al., 2022). Students can also better elaborate their problem-solving *strategy*, not just dive into the problem-solving *process*, which improves meta-cognition.

Better Testing and Better Learning I

Oral exams become not only an assessment method, but also a learning approach in their own right, as students who explain their ideas out loud end up learning more about the topic, especially when the conversation is scaffolded by an expert (the examiner) (Boser, 2017).

Better Testing and Better Learning II

Students receive immediate, specific, personalized feedback on their performance and learning:

“...there is strong evidence that students benefit a great deal from personalized feedback that is given in a tutorial setting where they have social and other incentives to absorb it, synthesize it, and deploy it anew” (Clay, 2023).

Better Learning I

One study found that oral exams increased engineering students’ motivation to learn, in addition to improving their grades on written exams administered after the oral exams (Delson et al., 2022).

Better Learning II

Mary Nelson’s study of at-risk Calculus students showed that students participating in voluntary oral assessments before written tests did significantly better on both grades in the course and on the common final exam than comparable students in the control group. Not only that, their *retention* was better: they passed Calculus I, then enrolled in and passed Calculus II at “dramatically higher rates than at-risk students in the control group” (p. 47, Nelson, 2010). Based on these results, oral exams were offered in all Calculus courses, even though the class size was often over 500 students.

More Interesting for the Examiner I

How many times has an instructor been bored to tears grading papers on the same topics? Even when students are given a choice of topics, oftentimes they cluster around one subject. Oral exams allow examiners not only to adapt the level of questions to the individual, but also to examine different topics. As long as there is a robust rubric to ensure that the exam is covering the appropriate learning outcomes and that the questions are fair, oral exams can end up being more interesting than written ones.



More Interesting for the Examiner II

Contrasting to the work required to grade a student’s written assessment, Clay says “Per minute, I enjoy myself so much more [with the oral exams]. Writing extensive written feedback and sending it into the abyss — even if its absorption is incentivized by being linked to subsequent assignments — is painful in a special way that grinds on me, especially when I have to do a lot of it” (Clay, 2023).

Instructors Get to Meet Students

University of Waterloo Math professor Anton Mosunov notes that, while conducting 74 oral exams for 15 minutes each was time-consuming, it allowed him to have personal contact with his students, which is especially crucial in fully online courses. “I was able to break down the barrier with students who had

never approached me with their questions before. Since the exam, those students are reaching out for help” (Chen, 2020).

Strategies for Oral Exams

One strategy that helps students prepare for an oral exam while simultaneously lessening their anxiety is for the first question to be **talking through a problem from their homework** or a previous assignment. In this way, the student has already solved—or at least worked on—the problem and starts off simply talking their way through what they did to solve it. The examiner can then ask questions to push that line of thinking, and/or embark on different topics. The student feels more confident knowing what the first question that will be asked is and gets warmed up by talking about a problem they’ve already mastered, as a way to ease into the exam.

Another strategy for reducing anxiety is to have the exam be **open book**. Even if students never need to consult their notes, they may feel more comfortable with their textbook close by. Encouraging students to create a “cheat sheet” of notes is a good way for them to review material, even if they never need it.

The most significant anxiety-reducing feature of an oral exam is the **examiner’s attitude**. Sayre describes her oral exams as a “friendly, guided conversation about physics” (Sayre, 2014). Maintaining a kind, supportive attitude throughout the exam will have the greatest positive effect on students. This can be challenging for the examiner, who must be “on” for all of the exams. Whereas a student may find it overwhelming to be quizzed for 20–30 minutes straight, consider the poor examiner who must ask questions, interpret answers, give hints, scaffold thoughts, push learners, while simultaneously grading them.....for each and every student!



Setting the expectations for students is important. In the vein of maintaining a “friendly conversation,” Dan Styer writes about his oral physics exams, “Your examiner is testing your reasoning skills, not your appearance. Don’t dress in elegant ‘dress for success’ fashion. Wear comfortable clothing that will not fall off when you’re writing on the chalkboard” (Tips for Oral Exams).

Media Attributions

- This image was created using DALL·E
- This image was created using DALL·E

- This image was created using DALL-E
- This image was created using DALL-E
- This image was created using DALL-E

INNER FEEDBACK...WITH CHATGPT'S HELP

We began this section by criticizing a generic assignment that might try to leverage ChatGPT by having students “critique” or correct an AI-generated passage (code, paragraph, outline, etc.). We were concerned that not only would this type of assessment likely be a mismatch for most disciplines, but it could create unfair advantages among students, and would almost certainly cause more grading work for the instructor. However, David Nicol, a professor in the Adam Smith Business School at the University of Glasgow, presents a guide to “turning active learning into active feedback” by encouraging students’ **inner feedback** (Nicol, 2022) that addresses many of these concerns.

Nicol's Original Design

The inner feedback model for assessing students predates ChatGPT and other LLM chatbots and was designed without conceiving of them at all; it was simply an excellent way of assessing student learning. More recently, however, Nicol has recognized ChatGPT as a useful complement to the activity and has adapted his model to include it.

We first must understand the process and benefits of Nicol’s inner feedback generation model before introducing how to leverage ChatGPT with this type of assessment. (Note: We go into quite a bit of detail about this approach; Nicol’s four-page ACTIVE FEEDBACK Toolkit is the best source and should be referenced directly, although we do reproduce a number of the important parts below.)

In his guide, Nicol recognizes that educators talk about “feedback” all the time, but it is almost exclusively with respect to the comments that instructors (and sometimes peers) put on drafts or summative assessments.

Yet comments do not constitute feedback until students process them, compare their interpretation of them against their work or performance, and generate new knowledge and understanding out of that comparison. (p. 1, Nicol, 2022)

Instructors often complain about sending comments “into the void,” spending hours writing or typing detailed comments on student assignments that end up not being read, and it is all but assured that “feedback” on final assessments seldom gets integrated into the student’s learning, which then feeds back nothing at all. Even comments on draft versions of assignments may not be as effective as hoped, depending on the student’s ability—and willingness—to process and integrate the comments. Contrary to the perception that students must wait for comments before proceeding with their learning, in fact, students are generating inner feedback all the time, by “comparing their thinking, actions, and work against external information in different kinds of resources,” (p. 1, Nicol, 2022) including rubrics,

textbooks, journal articles, diagrams, and even observations of activities and others' behaviour (Nicol, 2021).

“Inner feedback is the new knowledge that students generate when they compare their current knowledge against some external reference information, guided by their goals.”

He asserts that this natural feedback loop is generally ignored in education, with deference being given to instructor-generated comments. This not only increases the instructor workload, but also discourages students from developing their own agency and skills in inner feedback. To operationalize inner feedback, students have to make “mindful comparisons of their own work against external information and make the outputs of those comparisons explicit” (p. 2, Nicol, 2022), and they must be taught how to do that.

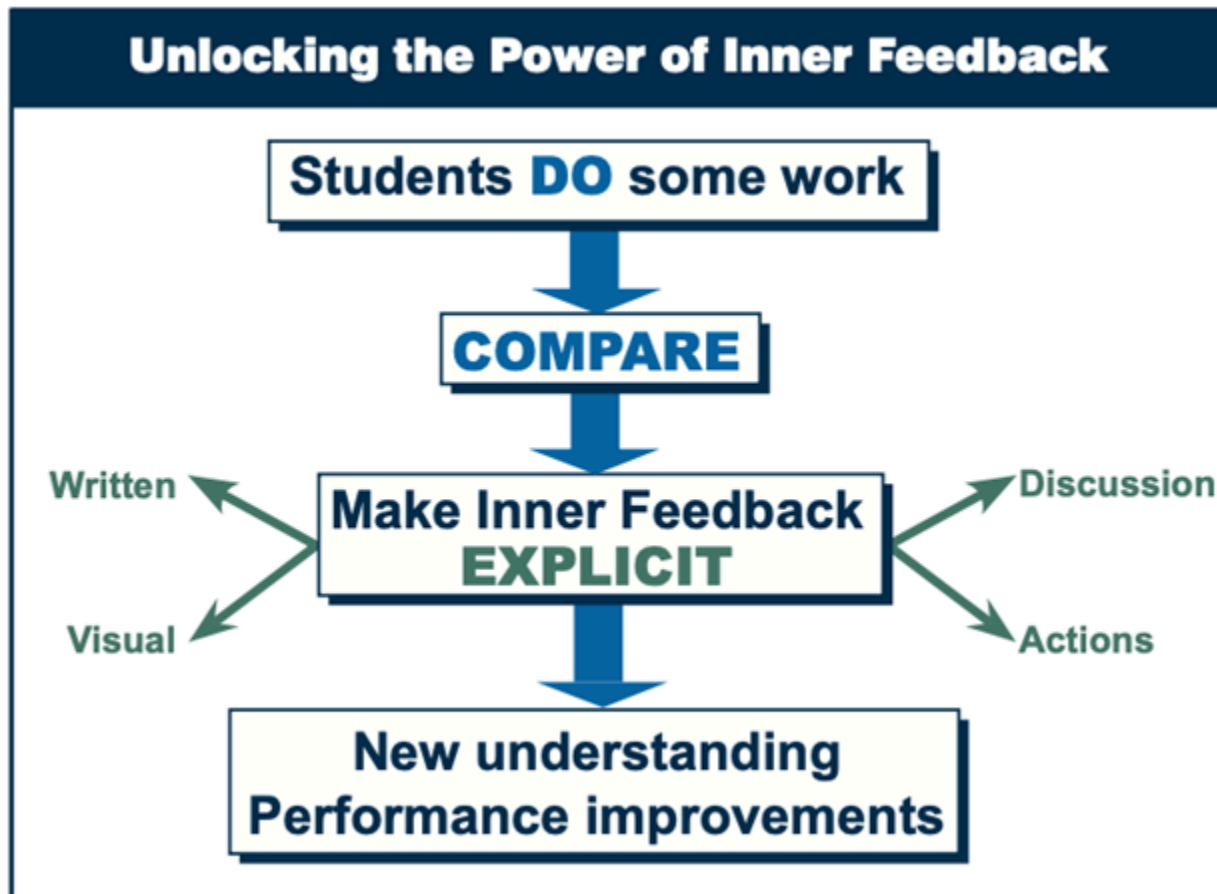


Figure 2
Unlocking the Power of Inner Feedback

Nicol's research shows that by going through this process, not only do students generate their own

feedback, but their observations are more varied and detailed than comments given by instructors (Nicol & Selvaretnam, 2022).

Implementing this type of assessment is quite straight-forward, even formulaic. Instructors must:

1. Decide what students will do, i.e., devise the learning task that will serve as the focus for feedback comparisons (Table 1: column 1)
2. Select or create appropriate comparators (column 2)
3. Formulate instructions to give a focus for the comparison and to make the outputs explicit (column 3)
4. Decide on next step: how to amplify the feedback students generate from resources (column 4). (Nicol, 2022)

| DO [Students...] | COMPARE [against] | INSTRUCTIONS [to make feedback generation explicit] | AMPLIFYING resource generated feedback |
|---|---|--|---|
| (i) Write concept definition, draw flow-chart of process, or propose example of application of a concept. This 5-10 min activity might occur in-class or before class as homework. | Lecture input (5-10 mins) e.g., that elaborates on concept definition, gives more insight into process, or that highlights some issues related to concept application | Write notes on what you learned from comparing what you wrote with the lecture input. Update your concept definition, your flow chart diagram, your application proposal. Identify and write down what questions the comparison raised for you. | Students share their activity and comparison outputs with peers and discuss and answer questions raised, identifying any unresolved or new questions for the lecturer. |
| (ii) Write 200-word argument on topic in or before class | Two published arguments on same topic OR 1 published counter argument. | Identify how these arguments have used evidence to support them and identify how to improve the use of evidence in your own argument. Identify how you can strengthen your own argument by addressing this counter-argument. | Discuss your argument and improvement ideas with peers. Select best example [to present in tutorial] OR Link your and peer's argument to form a better argument. |
| (iii) Solve bad debt accountancy problem before or in class. [Accounting & Finance] ** Suzanne McCallum | Video of expert (e.g., lecturer) talking through her solution to the problem. OR Flow-chart of the problem-solving process. | How did your thinking differ from the expert? What did you learn from that? What questions remain outstanding? Use flow-chart diagram to self-correct your work. Identify any bad debt situations where this flow chart might not apply. | Discuss comparison output with peers and identify any outstanding question worth asking the lecturer. Students poll/vote on questions teacher should answer |
| (iv) Groups present findings of their draft project report to class. | Presentations of other groups' findings on same report topic. | Individual students answer the following: How did your group findings differ from this group's? What recommendations were common across all groups, and which differed? Based on this write down any improvements for own report. | Individuals share outputs of comparisons with group members and together they update their group report. |
| (v) Write draft report on management or economics topic. | A rubric for the report and exemplars of reports on different topic. | Compare each of these resources against your draft report. Update your report and submit to lecturer including analysis output from comparison task. | Lecturer grades final report aided by comparison reflections (analyses). |
| (vi) Write 500-word essay ** Suzanne McCallum [Accounting & Finance] | Two peer essays and one essay of high-quality on same topic constructed by lecturer or selected from prior cohort. | How did your essay differ from this essay? What did you learn from that difference? How would you improve your own essay? Based on these three comparisons update your own essay. | Lecturer samples essays and provides some whole class feedback which students compare against own essay. |
| (vii) Write application or produce case study of economics model | Published account of theoretical model relevant to application or case. OR Published account of different model | How well does your application adhere to this published model? How could it be improved? What do you think are the limitations in the model? What have you learned from comparing your application against this alternative model? Improve your application based on this. | Discuss with peers then improve and submit your application or case study – alongside reflections on the outputs of the comparison task. |
| (viii) Individually write 300-word evaluation of international poverty index ** Geetha Selvaretnam [Economics] | Students do the same work again but in groups of three. Hence comparators are group discussion and unfolding and final group output. | Write an account of what you learned by comparing your individual output against the group discussion and group output? Write an account of what you would do to improve your own evaluation of the poverty index? Give reasons for your answer | Lecturer grades the individual and group work and the answers to the comparison questions. |
| (ix) Create a plan (schedule and question sequence) for first meeting (contracting phase) with client of business enterprise where students carry out a consultancy project. ** Nick Quinn and Alison Gibb [Management] | Published article from management journal on how to establish credibility. Video of a doctor interviewing a patient presenting with chest pains. | Use this theoretical article to identify improvements you could make to your meeting plan. [theory-practice comparison] Watch this video and note how the doctor engages with the patient, reassuring him while at the same time soliciting important information. Based on this, consider how you might foster empathy in your first client meeting: and update your plan. | Students discuss the findings from their comparisons with peers and further update their meeting plan. Any questions are identified and posed to the class and then to the lecturer. |
| (x) Write a draft literature review for their final year economics dissertation. ** Lovleen Kushwah [Economics] | Two published literature reviews on different topics drawn from high-quality economics journals, and different from student's own topic. | Identify three reasons why the two published reviews are of high quality. Give a rationale for each reason [focus on structure, argument, use of prior research in literature reviews]. Compare your own review with the published reviews and your rationale. | Students submit own review and answers to comparison questions to supervisor. Supervisor comments as necessary, building on students' own feedback |

Table 1

Implementations of active feedback in the Adam Smith Business School, University of Glasgow

It is important to identify and describe all four parts of this process. Students need to:

1. know what they will do/produce;
2. have access to appropriate comparators;
3. understand explicitly how to generate their own feedback; and
4. know how their new understanding will be instantiated/amplified.

Now here is where ChatGPT comes in: usually, the instructor provides **exemplars** in Step 2 as the comparators. The exemplars are published articles, instructor lectures, expert videos, rubrics, etc.; they are instructor-vetted material of high quality both in substance and format. This is the item against which the students compare their work in order to generate the inner feedback.

Nicol proposes adding a ChatGPT comparator to the mix. All of the other steps remain the same, but instead of having students only compare their work to the vetted, authoritative resources, students also generate an essay (or proposal, definition, problem, plan, report, or whatever the deliverable is) on the topic using ChatGPT and compare their work to it (along with the official resource). The ChatGPT output is not used as an exemplar, but rather, is an example of something on the topic that might exist in the world (but isn't necessarily superior). By using the same approach as outlined above to generate inner feedback, the student compares their work and ChatGPT's work to the authoritative resource and creates feedback about both items for themselves. They learn, first-hand, about the reliability of ChatGPT, where it might fail, how it might not perform as well as they can. By using a vetted, reliable resource in conjunction with ChatGPT, they can learn how to think critically about ChatGPT, while simultaneously being shown that there are no shortcuts and that they cannot rely on ChatGPT to effectively cheat (Rose, 2023).

We began this section by stating that both oral exams and developing inner feedback are ways to increase the AI-immunity of your assessments, but they are both also very effective learning and assessment methods. Not only is the assessment effective, but the learning that takes place in preparation for the assessment is often of a higher quality (Clay, 2023; Nelson, 2010). The key is in guiding and framing these new assessment activities for students. For oral exams, it's about setting expectations and organizing an effective, but conversational, assessment. For creating inner feedback, it's about helping students become agents in their own learning.

Media Attributions

- Table 1: Implementations of active feedback in the Adam Smith Business School, University of Glasgow

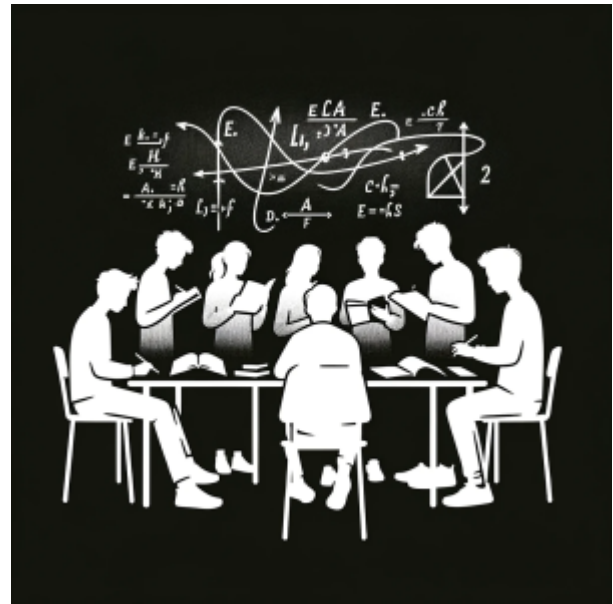
LOW STAKES ACTIVITIES USING CHATGPT

ChatGPT's initial appearance on the educational landscape in late 2022 was a shock to many; we observed three main camps of instructors: those who

- sought to outlaw any GenAI use entirely,
- ran to embrace it, and
- were blissfully unaware of (or in denial about) its impact.

Educators in the first two camps, the “early recognizers,” if not early adopters, each learned hard lessons that drew them potentially away from their initial stance: enthusiasts were confronted with a number of issues around bias, privacy and security, reliability, accessibility, availability (e.g., Google Bard was available in the US as of March 2023, then rolled out to Europe and Brazil in July 2023, but was never available in Canada; Gemini replaced Bard around the world and finally in Canada in February 2024) etc., and avoiders soon realised how truly difficult it is to police the use of LLM-based tools. These hard lessons were sobering for both camps; it wouldn't be as easy as just banning their use, but it also couldn't be as easy as throwing the doors open and requiring everyone to engage fully with these new tools. The same points that we discussed in the section on GenAI detectors are applicable to the LLM-based tools themselves: student copyright; privacy and security of data and identity; bias, potential toxic speech; unreliability; etc. As of the time of this writing, Arizona State University was the first—and only—higher education institution to partner with OpenAI, gaining access to ChatGPT Enterprise (ChatGPT 4 with no usage caps), which it will use for “coursework, tutoring, research and more” (Field, 2024). We can expect that other such deals will follow (this one was reportedly more than six months in the making and no financial details were released) but this is the first instance of a higher education institution entering into an agreement with an AI company which, presumably, will require some concrete action on ensuring student safety, privacy, intellectual property protections, and with, potentially, the ability for students to opt out.

However, the lack of a formal relationship with one of the private companies that builds these blackbox tools does not appear to be slowing down all institutions: in *Indecision About AI in Classes Is So Last Week*, “[p]rofessors and administrators from five major public universities provide advice on how to get moving ahead with AI in the classroom right now” (Ward et al., 2023). Yee et al. advocate for helping students develop AI fluency with (over 60) ChatGPT Assignments to Use in Your Classroom Today (Yee et al., 2023). Britain’s National centre for AI in tertiary education has extensive and detailed suggestions for leveraging AI in assessments: *Assessment ideas for an AI-enabled world*. It may suddenly feel like the world has moved on to fully embrace GenAI in the classroom—no matter what the context—and students



are feeling it too. As described in the introduction, 54% of Canadian university students are using GenAI in their studies, and 80% of those students are using it to answer more than one question per day. Fifty-six percent of Canadian students think that universities should promote the use of GenAI in assessments and 66% think that universities should change the way they assess students (Chegg.org, 2023).

If overhauling your entire course or assessment strategy seems impossible, you can always implement small, low-stakes activities and assessments that use ChatGPT. This will give you the chance to frame how you see GenAI tools best used in your class and your field, will give you the chance to teach your students not only how to use the tools, but also discuss the ethics and best practices for using them, all while experimenting in a low-stakes context.

We will close this chapter by describing a few smaller activities and assignments, some for in-class and some for online, that you can do with your students.

Generate, Then Regenerate

Have your students interact with ChatGPT on any topic: hobbies, sports, trivia, etc. Have them ask a few questions, and at some point, get them to click “Regenerate” with one of the prompts they’ve already used and compare the two answers they received. They can examine the quality of writing, the accuracy of the response, or evaluate the replies based on other criteria. Then, have your students enter a prompt related to a course topic, then click regenerate. Similarly, have them compare the outputs, evaluate their qualities, and determine which is better (and if either are actually good or accurate, or if they both have failings). Students can get into pairs or groups to compare their prompts and outputs.



Online students can do the first part of this activity on their own, at their own pace, and then share their reflections in a discussion board or live chat session.

This can be an ungraded activity, or students can hand in their prompts, outputs, and reflections for grading. The reflection is the most important thing to grade and provide feedback on.

This activity can also be done on multiple LLM-based tools, comparing the accuracy of responses across ChatGPT, Copilot, Perplexity, Gemini, etc. Students can determine whether one tool is better than another for specific prompts, types of outputs, different topics or subjects, or according to other criteria. Students can also change the context instructions to see if the tool alters its output to similar prompts.

Using the same prompt to generate different answers can be the beginning of a class-wide conversation about the strengths and limitations of using various GenAI tools, as well as a discussion of both ethics and best practices in the classroom and in the discipline.

Prompt Engineering Skills

Once we acknowledge that

- GenAI is here to stay,
- most people will use it in their jobs in one way or another (including higher education STEM instructors!), and
- 66% of Canadian post-secondary students would like their courses to include training in AI tools relevant to their future career

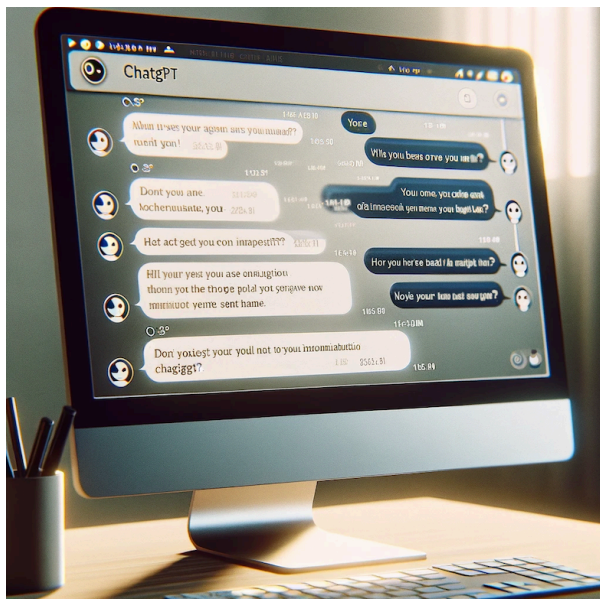
it seems clear that we must teach students how to use it. At some point in their education, students learned to print, and some learned (or will learn...) cursive writing. They learned to type, to use a word processor, to write an essay. They learned addition and subtraction, then how to use a calculator (or abacus or slide-rule) properly, and many use a variety of discipline-specific tools and programs, all of which they learned throughout the course of their education, usually in a formal



instructional setting. A tool as far-reaching and as impactful as ChatGPT should be included in formal, potentially discipline-specific, instruction.

Once students and teachers have had discussions about bias, ethics, reliability, etc. of LLM-based tools, they can get down to learning how to optimize the output. Note: these discussions have the potential to be very interesting, not only because the topic of GenAI is likely outside of most instructors' area of expertise, but also because the technologies, capabilities, and tools themselves are evolving so quickly and because this is an area where students may be better informed than their teachers. Or, perhaps students are better informed about some aspects, but not others ("Have you seen all of these amazing things it can do? Oh, but no, I had no idea about the copyright implications!").

The major stride forward between previous AI tools and ChatGPT et al. is the ability to converse with the tool in natural language instead of explicit, structured commands. It is the *conversational* aspect of these tools (as opposed to a traditional, pre-LLM search engine where you ask one thing and get that one thing [hopefully]) that is one of their greatest strengths. And, just as with humans, there is a skill in being a good conversationalist, interviewer, or examiner. Many of the successful approaches for LLMs are different from those you would use with a human interlocutor, but others are shockingly similar. Enter the field of **prompt engineering**.



You have perhaps heard of GIGO (garbage in, garbage out), the idea that flawed or nonsense input produces bad output? Prompt engineering seeks to achieve the opposite, by designing prompts (the inputs) for AI models in order to elicit specific, high-quality or otherwise optimized outputs. In this case, the “garbage” consists of questions that are too vague, too broad, cover too many topics, or are otherwise unclear, and will result in equally unfocused and disappointing results. Crafting queries that will unleash the power of an LLM-based tool trained on billions of words requires linguistic precision as well as an understanding of the model’s mechanics.

Lance Eliot goes so far as to say, “The use of generative AI can altogether succeed or fail based on the prompt that you enter” (Eliot, 2023). There is a whole domain of prompt optimization

attempting to figure out how to best query these ever-evolving tools, and there are many ways to introduce your students to the practice of writing good prompts.

Deconstruct, Then Reconstruct

As an introduction to prompt engineering, the instructor can provide students with a range of prompts and their outputs (potentially just from one tool, or multiple). Have the students analyze what makes a prompt effective or ineffective. Then, ask students to reconstruct the ineffective prompts to improve the responses, which they can then test out. The goal is for students to be able to identify key elements of prompt engineering, such as clarity, specificity, and context. This can be done alone, or in groups.

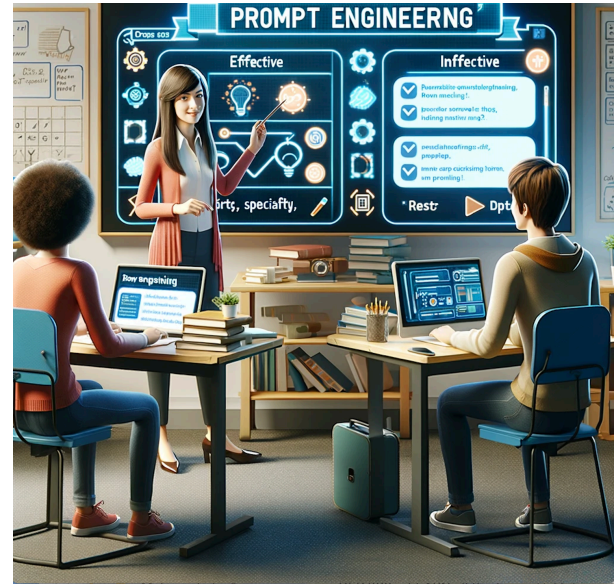
In an online course, students could use discussion boards or a tool such as Padlet to show each other their original prompts and their improved prompts. Students can work iteratively on the prompts in a group chat, discussion board, or shared document.

Design Workshop

Once students are comfortable with the basic components of a good prompt, you can move to the application level: students are assigned a specific question or problem they must get the AI tool to help solve. Students are tasked with

- creating prompts to elicit the best possible answers,
- testing their prompts, and
- iterating them based on the results.

Students can do this task on their own or in pairs/groups. The problems should be complex enough that multiple prompts are required.



Nuance in Prompting

In order to understand how slight differences in prompts can affect the output, have students create multiple prompts for the same task, varying their structure, tone, specificity, etc. Once they have a series of prompts, students can enter them and judge the output, marking a rubric or taking notes on how these differences impact the responses from the AI tool. Students can also create prompts in other languages, if they know them, to determine how well the tools function in languages other than English.

Bias Recognition

Based on conversations in class, your students will hopefully be aware of the bias inherent in LLM-based tools (although this is evolving, and recent mitigation efforts such as those undertaken by Google in Gemini in late February 2024 are definitely problematic). There are a number of scenario-based activities you can do to have them create concrete examples for themselves of bias—and hopefully of bias mitigation—in their work.

Gender Bias



Scenario: An AI-based tool is asked to describe professionals in various fields, such as engineering, nursing, teaching, and coding. The tool consistently assigns stereotypical gender to certain professions.

Discussion Questions: Discuss the implications—personal, societal, professional, etc.— of reinforcing stereotypes through AI responses. How can prompts be structured to neutralize gender assumptions? How should outputs be vetted to mitigate bias? What are the latest tools doing to address bias—and is it working?

Socio-economic Bias

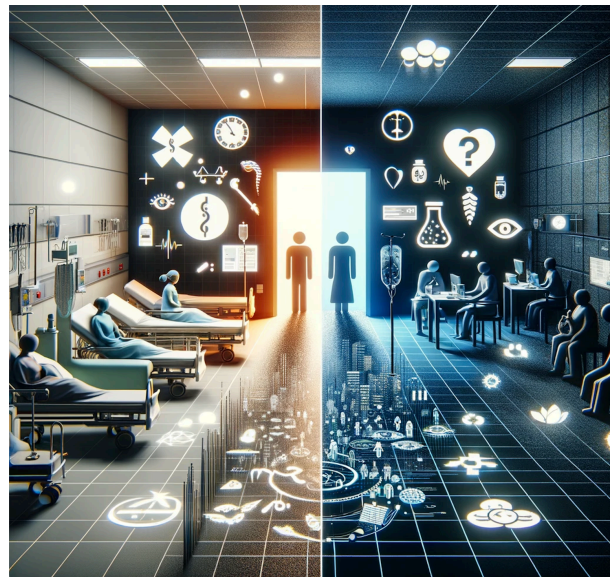
Scenario: An AI-based tool is asked to suggest solutions for urban transportation challenges. It consistently favours high-tech, high-cost solutions over more accessible or low-cost types of mass transit.

Discussion Questions: How does the bias towards technologically advanced solutions affect the usefulness of AI recommendations in other socio-economic contexts? How can students encourage the tool to consider—and plan for— diverse economic realities?

Bias in Medical Advice

Scenario: Advice on health and wellness topics generated by an LLM-based tool reflects existing biases in medical research, such as under-representation of certain demographics (women, pregnant people, older people, children, etc.) in clinical trials, confounding bias (where correlative factors are not accounted for, leading to a false causal association), or language and cultural bias.

Discussion Points: Discuss the potential consequences of biased health advice (for individuals and society as a whole) and explore how prompts can be designed to ask for information that might be more inclusive of under-represented groups.



Bias in AI-Assisted Research Data Analysis

Scenario: A group of students uses an AI-based tool to analyze genetic data for a project. They

notice that the AI's interpretations and predictions heavily favour data from populations of European descent, reflecting biases in the underlying training datasets.

Discussion Points: What is the impact of dataset composition on AI analysis in scientific research? How can prompts be structured to account for or highlight the limitations of the data? Can outputs be trusted?

Ethical Implications of AI in Environmental Modelling

Scenario: Students employ an AI tool to model climate change impacts in various contexts. However, the AI disproportionately focuses on scenarios relevant to countries in the Global North, overlooking the nuances and specific needs of regions in the Global South.

Discussion Points: Discuss the importance of inclusive and globally representative environmental data. How can prompts ensure that AI models consider diverse ecological and socio-economic impacts? Do the LLMs have a rich enough training set to account for experiences outside Europe and North America?

Bias in Facial Recognition Technologies

Scenario: Students develop an AI project that involves facial recognition technology. They discover the model performs poorly on faces from certain ethnic backgrounds due to biases in the training data.

Discussion Points: What are some ethical considerations and societal impacts of biased facial recognition technologies? Discuss prompt engineering strategies to test for and mitigate these biases. Given that racial bias is a long-standing problem in AI-based tools, are things improving, or are errors and discrimination only becoming more widespread?



Ethical Use of AI in Academic Research

Scenario: A research team uses AI to automate the literature review process for a new scientific study. They find that the AI tends to cite papers from a limited set of journals, potentially biasing the review.

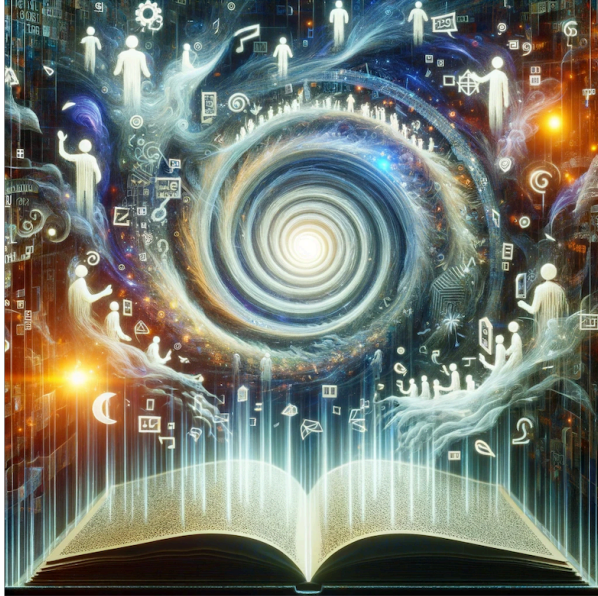
Discussion Points: Given the importance of diversity in scientific literature, what is the role of AI in ensuring a broad and unbiased review? How can prompts encourage a wider search of sources?

Dealing with Misinformation and Hallucinations

One of the most serious flaws of ChatGPT (and other LLM-based tools, to a greater or lesser degree) is its propensity to hallucinate, or simply make things up. Sometimes it gets answers wrong, and

sometimes it makes things up entirely. Training students to watch for inaccuracies—while honing their critical thinking skills— is important for their future professional success.

Catching Hallucinated Sources



ChatGPT is especially prone to hallucinating sources and citations: it’s common for the tool to acknowledge “a real scholar, perhaps even the exact ideal scholar an expert might quote, but with publication titles or journals listed that sound realistic, yet do not exist” (p. 46, Yee et al., 2023).

Yee et al. suggest an activity where students use AI to generate a bibliography on a particular essay topic, then verify the articles in the library’s database, creating a screenshot to confirm (or reject) that all the sources are accurate. Students could summarize one of the articles, or, with ChatGPT’s help, the entire related body of work, for their classmates to read.

Challenging Fallacies

Similar to the idea of hallucinations is the existence of logical fallacies. Because LLM-based tools aren’t “thinking” on their own, but are predicting the next word in a string of text, and as we have already seen, some of the training datasets could be of dubious quality, ChatGPT may be recreating fallacies similar to those it was trained on. Students in all disciplines should be trained to watch for fallacies, and there are a number of logical fallacies that an LLM-based tool could perpetuate in a STEM context.

Appeal to Authority (Argumentum ad Verecundiam)

LLM-based tools might rely too heavily on authority figures in certain fields, suggesting that a claim is true simply because an expert or authority asserts it, without presenting concrete evidence. This can be misleading if the authority’s opinion is not widely accepted or is out of their area of expertise. This challenge is compounded by the opacity of LLMs: sources cannot be traced back to clarify meaning or nuance (Birhane et al., 2023).

Instructors could challenge students to research the purported authority figure and the claims to determine whether ChatGPT has accurately presented the information. It is difficult for non-experts to confirm ChatGPT’s inaccuracies in many topics, especially because it “speaks” with such confidence, so when it appeals to authority, it can be a double threat. Instructor/expert guidance for students is important for this type of activity.

Post hoc Fallacy

Similar to the fallacy of correlation implying causation, ChatGPT can fall prey to the post hoc fallacy by mistakenly asserting a causal relationship between two events just because they occur sequentially.

Educators can create activities or discussions focussing on examples of this fallacy and encourage students to identify similar uses by the AI. For example, ChatGPT output might

- suggest that because technological advancement has increased since the popularization of the Internet, it is the Internet that has directly caused all modern technological advancements.
- suggest that the introduction of genetically modified organisms (GMOs) in agriculture directly led to a decline in bee populations (ignoring the complex reasons behind bee decline, such as pesticide use, habitat loss, and climate change).
- claim that because a certain environmental change occurred before a particular species evolved a new trait, the environmental change directly caused the evolution of that trait (disregarding genetic variation, selection pressures, and other environmental factors).
- assert that an increase in vaccination rates in a population directly resulted in the reduction of a completely unrelated disease (conflating correlation with causation and ignoring other health interventions or natural disease progression patterns).



Students can write rebuttals to these fallacies, explaining why the arguments are specious, and offer alternate explanations. This activity can be done in pairs or groups, and groups can trade descriptions for their peers to review and expand upon.

False Dichotomy (False Dilemma)

In discussions involving complex problems, ChatGPT output may present issues as having only two possible solutions when, in fact, more options exist. Birhaine et al. warn of this over-simplification occurring when using LLM tools to summarize complex scientific papers. Using an AI tool that frames complex issues as binary choices overlooks the nuanced, multi-stepped, and potentially multifaceted solutions that are often required in STEM problems.

As with previous fallacy types, activities that encourage students to challenge the tool's output, improve their critical thinking skills, and offer rebuttals and alternate solutions or descriptions are important to integrate into all courses.

CONCLUSION

In this section, we explored the pivotal role of oral exams in fostering direct, dynamic interaction between educators and students, enhancing personalized feedback, and mitigating the risks of academic dishonesty. Both oral exams and Nicol's inner feedback strategies not only align with traditional educational values but also adapt to the digital age's demands by encouraging critical thinking and real-time problem-solving skills, while increasing the "AI-immunity" of assessments.

We also highlighted the significance of low-stakes activities utilizing ChatGPT. These activities offer students a safe environment to experiment with AI tools, develop prompt engineering skills, and critically engage with AI-generated content, including identifying biases and misinformation. Such tasks prepare students for the inevitable integration of AI in their future careers, ensuring they are not only proficient in using these tools but also aware of their limitations and ethical implications. By blending oral assessments with innovative low-stakes activities, educators can provide a comprehensive, forward-thinking education that equips students with the skills needed to navigate and succeed in a rapidly evolving technological landscape.



In an ideal world, departments, faculties, and institutions would work in concert to create exposure to and training in LLM-based tools across the curriculum, scaffolding and expanding activities and assessments across courses and throughout undergraduate and graduate degrees. Students would all be engaged in discussions around ethics and bias, accuracy and efficacy, and discipline-appropriate use of GenAI tools.

PART VII

FINAL THOUGHTS

Rather than attempt to restrict the use of generative AI, our goal must be to teach students to use and surpass it. (Steipe, 2023a)

In the 18 months since the arrival of ChatGPT, we have seen a whirlwind of possibility, promise, failure, and growth in the use of GenAI tools for almost any task. Educators at all levels have had to grapple with changes to teaching and assessment methods, questioning the fundamental goals—and abilities—of education.

In navigating both our own use of LLM-based tools in STEM teaching, as well as our students' tool use in their assignments, we are faced with a pedagogical evolution, if not a revolution. This book has described the history of AI in general and explained the development and training of LLM-based tools, as well as their limitations. We discussed biases inherent in the datasets that power these tools, underlining the urgent need for effective and ongoing mitigation efforts. We have presented some of the challenges of maintaining academic integrity, recognizing that while AI detectors and policies play a role, the solutions lie in open, constructive communication with students and innovative assessment strategies.



Educators are called to harness the strengths of LLMs in improving student engagement, enhancing learning materials, and providing career guidance, while also being vigilant of their limitations in areas like mathematical calculations and nuanced topic exploration. We have demonstrated the potential of ChatGPT as a facilitator for low-stakes learning activities (including developing prompt engineering skills), and as a partner in generating and refining content and assessments, offering a glimpse into a potential shift in the future of teaching and learning.

We have considered a variety of assessment formats, from oral exams to innovative uses of ChatGPT in generating formative feedback. These strategies underscore the shift towards assessments that reflect the changing nature of knowledge work in

the digital age. Educators are encouraged to rethink the ways in which we evaluate understanding and skills, moving towards assessments that challenge students to synthesize, apply, and communicate knowledge in ways that machines cannot.

As we look forward, it is imperative that we continue to approach the integration of LLMs into education with a critical eye, embracing their potential to transform teaching and learning while remaining committed to the principles of equity, integrity, and critical thinking. But we need to put that critical eye on the nature of education itself: how will the power (for good, or for ill) of GenAI change the way we teach and assess? What is the value of education, based on societal needs?

The journey does not end here; it evolves with every update to the algorithms and every shift in the educational paradigms. We will continue to share our experiences, research, and insights into the effective use of LLMs in education, fostering an environment where technology enhances, rather than diminishes, the human aspects of teaching and learning.



WORKS CITED

- Aligning language models to follow instructions.* (n.d.). Retrieved December 9, 2023, from <https://openai.com/research/instruction-following>
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias*. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Azaria, A. (2022). *ChatGPT Usage and Limitations*. <https://hal.science/hal-03913837>
- Bandy, J., & Vincent, N. (2021). *Addressing “Documentation Debt” in Machine Learning Research: A Retrospective Datasheet for BookCorpus* (arXiv:2105.05241). arXiv. <https://doi.org/10.48550/arXiv.2105.05241>
- Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, 5(5), Article 5. <https://doi.org/10.1038/s42254-023-00581-4>
- Boser, U. (2017, May 5). Talking to Yourself (Out Loud) Can Help You Learn. *Harvard Business Review*. <https://hbr.org/2017/05/talking-to-yourself-out-loud-can-help-you-learn>
- Brandom, R. (2023, June 7). *What languages dominate the internet?* Rest of World. <https://restofworld.org/2023/internet-most-used-languages/>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. <https://doi.org/10.48550/arXiv.2005.14165>
- Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershovich, D. (2023). Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, & L. Benotti (Eds.), *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)* (pp. 53–67). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.c3nlp-1.7>
- CAP Recommendations to Aid in Adoption of New eGFR Equation.* (n.d.). College of American Pathologists. Retrieved December 7, 2023, from <https://www.cap.org/member-resources/articles/cap-recommendations-for-adoption-of-new-egfr-equation>
- Chaslot, G., & Monnier, S. (n.d.). *AlgoTransparency*. AlgoTransparency. Retrieved December 9, 2023, from <https://algotransparency.org/>
- Chegg.org. (2023). *2023 Global Student Survey* (p. 61). Chegg.org. https://8dfb1bf9-2f43-45af-abce-2877b9157e2c.usrfiles.com/ugd/8dfb1b_e9bad0aef091478397e6a9ff96651f6d.pdf
- Chen, H. (Weixu). (2020, August 14). *Personalizing math tests during COVID-19*. Waterloo News. <https://uwaterloo.ca/news/mathematics/personalizing-math-tests-during-covid-19>
- Clark, M. (2021, April 9). *Students of color are getting flagged to their teachers because testing software*

can't see them. The Verge. <https://www.theverge.com/2021/4/8/22374386/proctorio-racial-bias-issues-opencv-facial-detection-schools-tests-remote-learning>

Clay, G. (2023, October 4). *Oral Exams and First-Pass Grading with ChatGPT*. AutomatED: Teaching Better with Tech. <https://automatedteach.com/p/oral-exams-firstpass-grading-chatgpt>

Clay, G., & Lee, C. W. (2023, August 23). *Embracing Constructive Dialogue and Oral Assessments in the Age of AI*. Inside Higher Ed. <https://www.insidehighered.com/opinion/views/2023/08/03/how-professors-can-use-dialogue-based-course-assessments-opinion>

Collins, M. (n.d.). *Statistical Machine Translation: IBM Models 1 and 2*.

Cooper, G. (2023). Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence. *Journal of Science Education and Technology*, 32(3), 444–452. <https://doi.org/10.1007/s10956-023-10039-y>

Cowen, T., & Tabarrok, A. T. (2023). *How to Learn and Teach Economics with Large Language Models, Including GPT* (SSRN Scholarly Paper 4391863). <https://doi.org/10.2139/ssrn.4391863>

Deck, A. (2023, September 6). *We tested ChatGPT in Bengali, Kurdish, and Tamil. It failed*. Rest of World. <https://restofworld.org/2023/chatgpt-problems-global-language-testing/>

Delson, N., Baghdadchi, S., Ghazinejad, M., Lubarda, M., Minnes, M., Phan, A., Schurgers, C., & Qi, H. (2022, August 23). *Can Oral Exams Increase Student Performance and Motivation?* 2022 ASEE Annual Conference & Exposition. <https://peer.asee.org/can-oral-exams-increase-student-performance-and-motivation>

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus* (arXiv:2104.08758). arXiv. <https://doi.org/10.48550/arXiv.2104.08758>

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, 82(1), 62–68. <https://doi.org/10.1037/0022-3514.82.1.62>

Dreyfus, H. L. (1965). *Alchemy and Artificial Intelligence*. RAND Corporation. <https://www.rand.org/pubs/papers/P3244.html>

Dubiansky, S. (2020, October 28). *Students speak out on controversial lockdown browsers for online courses*. Technician. https://www.technicianonline.com/news/students-speak-out-on-controversial-lockdown-browsers-for-online-courses/article_c9693924-1996-11eb-869a-4b90407243a0.html

Eaton, S. E. (2020, January 15). *Cheating may be under-reported across Canada's universities and colleges*. The Conversation. <http://theconversation.com/cheating-may-be-under-reported-across-canadas-universities-and-colleges-129292>

Eaton, S. E. (2022). Contract Cheating in Canada: A Comprehensive Overview. In S. E. Eaton & J. Christensen Hughes (Eds.), *Academic Integrity in Canada: An Enduring and Essential Challenge* (pp. 165–187). Springer International Publishing. https://doi.org/10.1007/978-3-030-83255-1_8

Edwards, B. (2022, November 18). *New Meta AI demo writes racist and inaccurate scientific literature, gets pulled*. Ars Technica. <https://arstechnica.com/information-technology/2022/11/after-controversy-meta-pulls-demo-of-ai-model-that-writes-scientific-papers/>

Eliot, L. (2023, September 27). *Does Take A Deep Breath As A Prompting Strategy For Generative AI Really Work Or Is It Getting Unfair Overworked Credit*. Forbes > Innovation > AI. <https://www.forbes.com/sites/lanceeliot/2023/09/27/does-take-a-deep-breath-as-a-prompting-strategy-for-generative-ai-really-work-or-is-it-getting-unfair-overworked-credit/?sh=3ef77d6518c3>

- Field, H. (2024, January 18). *OpenAI announces first partnership with a university*. CNBC. <https://www.cnbc.com/2024/01/18/openai-announces-first-partnership-with-a-university.html>
- Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., & Berner, J. (2023). *Mathematical Capabilities of ChatGPT* (arXiv:2301.13867). arXiv. <https://doi.org/10.48550/arXiv.2301.13867>
- Gardiner, B. (2018). "It's a terrible way to go to work:" what 70 million readers' comments on the Guardian revealed about hostility to women and minorities online. *Feminist Media Studies*, 18(4), 592–608. <https://doi.org/10.1080/14680777.2018.1447334>
- Gewirtz, D. (2023, October 9). *Can AI detectors save us from ChatGPT? I tried 5 online tools to find out*. ZDNET. <https://www.zdnet.com/article/can-ai-detectors-save-us-from-chatgpt-i-tried-5-online-tools-to-find-out/>
- Ging, D., & Siapera, E. (2018). Special issue on online misogyny. *Feminist Media Studies*, 18(4), 515–524. <https://doi.org/10.1080/14680777.2018.1447345>
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480. <https://doi.org/10.1037//0022-3514.74.6.1464>
- Growcoot, M. (2023, November 3). *Which AI Image Generator is The Most Biased?* PetaPixel. <https://petapixel.com/2023/11/03/which-ai-image-generator-is-the-most-biased/>
- Hall, D. (2019, December 11). The ELIZA Effect. 99% Invisible. <https://99percentinvisible.org/episode/the-eliza-effect/>
- Heaven, W. D. (2022, November 18). *Why Meta's latest large language model survived only three days online*. MIT Technology Review. <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/>
- Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2023). *Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey* (arXiv:2207.07068). arXiv. <https://doi.org/10.48550/arXiv.2207.07068>
- Johnson, K. (2021, June 17). *The Efforts to Make Text-Based AI Less Racist and Terrible* | WIRED. <https://www.wired.com/story/efforts-make-text-ai-less-racist-terrible/>
- Kopsaftis, E. (2020, December 4). *Over 4700 signatures against the LockDown Browser at U of G*. The Ontarion. <https://theontarion.com/2020/12/04/over-4700-signatures-against-the-lockdown-browser-at-u-of-g/>
- Lederberg, J. (1987). How DENDRAL was conceived and born. *Proceedings of ACM Conference on History of Medical Informatics*, 5–19. <https://doi.org/10.1145/41526.41528>
- Liang, W., Yuksekogonul, 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>
- Lighthill, J. (1972). *Artificial Intelligence*. Cambridge University.
- Little, O., & Richards, A. (2021, October 5). *TikTok's algorithm leads users from transphobic videos to far-right rabbit holes*. Media Matters for America. <https://www.mediamatters.org/tiktok/tiktoks-algorithm-leads-users-transphobic-videos-far-right-rabbit-holes>
- Liu, & Bridgeman, A. (2023, June 8). *ChatGPT is old news: How do we assess in the age of AI writing co-pilots? – Teaching@Sydney*. <https://educational-innovation.sydney.edu.au/teaching@sydney/chatgpt-is-old-news-how-do-we-assess-in-the-age-of-ai-writing-co-pilots/>
- Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). *Stable Bias: Analyzing Societal*

Representations in Diffusion Models (arXiv:2303.11408). arXiv. <https://doi.org/10.48550/arXiv.2303.11408>

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *Dartmouth Summer Research Project on Artificial Intelligence*.

McCrosky, J., & Geurkink, B. (2021). *YouTube Regrets*. <https://foundation.mozilla.org/en/youtube/findings/>

Murgia, M. (2023, July 23). *Transformers: The Google scientists who pioneered an AI revolution*. <https://www.ft.com/content/37bb01af-ee46-4483-982f-ef3921436a50>

Nelson, M. A. (2010). Oral Assessments: Improving Retention, Grades, and Understanding. *PRIMUS*, 21(1), 47–61. <https://doi.org/10.1080/10511970902869176>

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education*, 46(5), 756–778. <https://doi.org/10.1080/02602938.2020.1823314>

Nicol, D. (2022, May 31). “Turning Active Learning into Active Feedback”, *Introductory Guide from Active Feedback Toolkit, Adam Smith Business School* [Educational resource]. Figshare; National Teaching Repository. <https://doi.org/10.25416/NTR.19929290.v3>

Nicol, D., & Selvaretnam, G. (2022). Making internal feedback explicit: Harnessing the comparisons students make during two-stage exams. *Assessment & Evaluation in Higher Education*, 47(4), 507–522. <https://doi.org/10.1080/02602938.2021.1934653>

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe>

Rettberg, J. W. (2022, December 6). ChatGPT is multilingual but monocultural, and it’s learning your values. *Jill/Txt*. <https://jilltxt.net/right-now-chatgpt-is-multilingual-but-monocultural-but-its-learning-your-values/>

Rose, J. (2023, February 21). *ChatGPT as a teaching tool, not a cheating tool*. THE Campus Learn, Share, Connect. <https://www.timeshighereducation.com/campus/chatgpt-teaching-tool-not-cheating-tool>

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. <https://doi.org/10.1037/h0042519>

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229. <https://doi.org/10.1147/rd.33.0210>

Sayre, E. C. (2014). Oral exams as a tool for teaching and assessment. *Teaching Science*.

Smith, C. (2023, March 13). *Hallucinations Could Blunt ChatGPT’s Success—IEEE Spectrum*. <https://spectrum.ieee.org/ai-hallucination>.

Snowell, A. J., & Burgess, J. (2022, November 29). *The Galactica AI model was trained on scientific knowledge – but it spat out alarmingly plausible nonsense*. The Conversation. <http://theconversation.com/the-galactica-ai-model-was-trained-on-scientific-knowledge-but-it-spat-out-alarmingly-plausible-nonsense-195445>

System Development Corporation. (1986). *SDI Large-Scale System Technology Study* (p. 135).

Tiku, N., Schaul, K., & Chen, S. Y. (2023, November 1). *These fake images reveal how AI amplifies our worst stereotypes*. Washington Post. <https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/>

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). “Kelly is a Warm Person,

Joseph is a Role Model”: Gender Biases in LLM-Generated Reference Letters (arXiv:2310.09219). arXiv. <https://doi.org/10.48550/arXiv.2310.09219>

Ward, D., Gibbs, A., Henkel, T., Loshbaugh, H. G., Siering, G., Williamson, J., & Kayser, M. (2023, December 1). *Indecision About AI in Classes Is So Last Week*. Inside Higher Ed. <https://www.insidehighered.com/opinion/career-advice/2023/12/01/advice-about-ai-classroom-coming-new-year-opinion>

Watkins, M. (2023a, January 6). Our Obsession with Cheating is Ruining Our Relationship with Students [Substack newsletter]. *Rhetorica*. <https://marcwatkins.substack.com/p/our-obsession-with-cheating-is-ruining>

Watkins, M. (2023b, January 30). It’s Time to Step off the AI Panic Carousel Before We Harm our Students [Substack newsletter]. *Rhetorica*. <https://marcwatkins.substack.com/p/its-time-to-step-off-the-ai-panic>

Wertheim, S. (2016, May 30). *The Common Habit That Undermines Organizations’ Diversity Efforts*. Fast Company. <https://www.fastcompany.com/3060336/the-common-habit-that-undermines-organizations-diversity-efforts>

Wertheim, S. (2023, April 26). *ChatGPT insists that doctors are male and nurses female*. Worthwhile Consulting. <https://www.worthwhileconsulting.com/read-watch-listen/chatgpt-insists-that-doctors-are-male-and-nurses-female>

West, M., Kraut, R., & Chew, H. E. (2022). *I’d blush if I could: Closing gender divides in digital skills through education—UNESCO Digital Library*. <https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1>

Wolfram, S. (2023, February 14). *What Is ChatGPT Doing ... and Why Does It Work?* <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/>

Xu, A., Pathak, E., Wallace, E., Gururangan, S., Sap, M., & Klein, D. (2021). *Detoxifying Language Models Risks Marginalizing Minority Voices* (arXiv:2104.06390). arXiv. <https://doi.org/10.48550/arXiv.2104.06390>

Yee, K., Whittington, K., Doggette, E., & Uttich, L. (2023). *ChatGPT assignments to use in your classroom today* (First Edition). FCTL Press.

ACKNOWLEDGEMENTS

This resource funded by the Government of Ontario. The views expressed in this publication are the views of the authors and do not necessarily reflect those of the Government of Ontario.



Media Attributions

- eCampus Ontario logo
- Province of Ontario logo