Developing critical attitudes towards artificial intelligence in education by deconstructing folk misconceptions of large-language models

Nikos Antonio Kourous-Vázquez

2024

Table of Contents

1. Introduction

Machine learning – often elevated by the term artificial intelligence – is developing rapidly and becoming increasingly integrated into existing systems. Its ability to process data to accurately predict outcomes makes it transformative across industries (Sappaile *et al*., 2024). Even more stark, machine learning systems have become widely accessible to the public, transitioning away from research labs and into everyday products. The shift from research technology to product – from concept to implementation – is significant and calls for an increase in scrutiny, in holding the technology accountable, and not just regulating and safeguarding it, but equipping people with frameworks to do so individually.

Machine learning describes a computational process capable of independently improving its ability to predict outcomes by analyzing patterns in existing data (Nilsen, 2018). Its ability to continually improve itself and interpret patterns lends it a high degree of flexibility, allowing it to perform a variety of tasks (Patil et al., 2024). Currently, large-language models – focused on text generation – are among the most accessible applications of machine learning (Reddy et al., 2024; Chalmers 2023).

Today, large-language models power widely used virtual assistants like OpenAI's ChatGPT and Google's Gemini. These virtual assistants, while powered by machine learning, represent their own subbranch (Shanahan 2024**).** Sometimes referred to as conversational AI, chatbots, or AI assistants, they represent packaged large-language models, designed specifically to be engaged in conversational, human-like ways (Naveed et al., 2024). These accessible and browser-based chatbots have rapidly permeated ways of working – their approachable and intuitive design make them easily adaptable for everyday use (Dharan and Nanda 2021). Given their flexibility and adaptability, research into their education integration has accelerated, exploiting the use of conversational large-language models to assist both teachers and students.

Research into educational uses of large-language models explore their potential to modernize teaching practices while also aiding teachers. This body of research, however, tends to approach the topic from a technocentric perspective, focusing more on technical mitigation of

risk instead of human, educator, and student-centric mitigation. Essentially, studies focus on the technological drawbacks of large-language models while looking past the active role people can have in avoiding or preventing them. In doing so, a dynamic is created, which paints large-language models as something thrusted onto students and educators instead of something capable of being harnessed. From this angle, strengthening people's understanding of large-language models allows for the development of critical mindsets to mitigate their drawbacks and risks.

This dissertation approaches the development of critical attitudes towards conversational large-language models by deconstructing common misconceptions about how they work. Understanding how people perceive chatbots allows for an examination into blindspots users could have towards them, which, if left uncontested, would prevent critical engagement. Through interviews, common misconceptions are revealed like the anthropomorphization of chatbots and notions of active thinking, research, and unbiased neutrality. Analyzing these misconceptions creates opportunities to engage with specific educational limitations of chatbots and stresses the need for deepened technical understandings of machine learning. Additionally, the demystification of machine learning also opens pathways for critical pedagogy by reframing chatbot limitations as productive opportunities—encouraging learners to critically interrogate AI systems and engage with their shortcomings as sites of inquiry and reflection

2. What are chatbots and how they work

Chatbots are user-friendly interfaces for large-language models. Different technology companies have developed their own versions, like OpenAI's ChatGPT and Google's Gemini. While their approaches may vary slightly, they are fundamentally powered by a computational architecture called generative pre-trained transformers (Wu *et al.,* 2023). GPT language-models work by analyzing vast amounts of text data during its pre-training phase, establishing statistical connections between words to then accurately generate new text on its own. To do this, GPT models deal with advanced mathematics and statistical relationships between words and concepts. During its training phase, GPT models process text data by splitting words into units of text called 'tokens' that either represent single 'c' haracters, whole words, or par-ts of a word, depending on the context (Pettit, 2023). During training, all

encountered words are ‘embedded’ with numerical values that locate them within a matrix of other words. To simplify, words with similar numerical embeddings are likely associated with one another. Figure 1 visualizes this concept. After processing text data, the visualized GPT model figures that a word like ‘bear’ is not often associated with the word ‘avocado’ – if asked to name animals, the visualization suggests it would likely respond with bear, cow, and duck, instead of avocado, melon, and turnip.

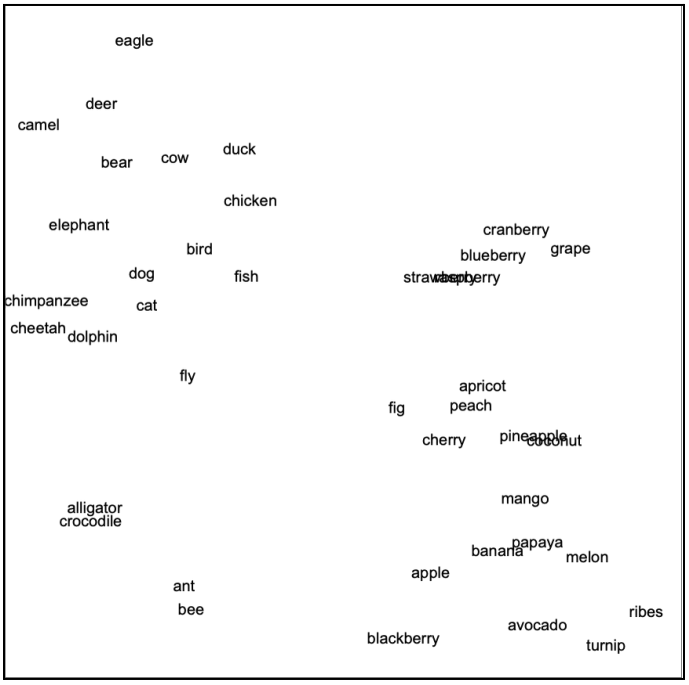

Figure 1, Source: Wolfram, 2023.

The process of generating text word by word based on statistical likelihood reveals that chatbots are more mimicking machines than thinking machines. By calculating the likelihood of each word, one after another, chatbots try their best to create answers that should, mathematically, be correct and expected. As Newport (2023) explains, chatbots arrive at a response by mimicking what the most correct answer would look like, not by coming up with it on its own. Hicks *et al*. (2024) explain how by mimicking answers, chatbots are more about “[giving] the impression that [they know] what they’re doing” rather than “accurately [representing] the world (p. 1).

3. Misconceptions of large-language models

During a group discussion about AI intelligence at a workshop I attended called *Philosophy of AI* (2024), a participant compared artificial intelligence to the Delphic Oracle – the priestess of Apollo who

channeled divine wisdom and prophecy. Indeed, the oracle's prophecies lent her significant influence; often, city-states and leaders would seek her advice before making major decisions (Walsh, 2013). In doing so, however, some leaders failed to critically perceive their situation, fixed, instead, with a magical faith in her knowledge. Misinterpretations of her prophecies, for example, led to the fall of the Kingdom of Lydia and the sack of Athens in 480 BC (Payne, 2023; Yates, 2023). A magical perception of machine learning could equally leave one vulnerable to its fallacies.

To most, large-language models also operate in mystical ways. Around the release of ChatGPT, journalists, academics, and technology professionals described it as a "mix of software and sorcery" (Roose, 2022; Taecharungroj, 2023). Indeed, Arthur C. Clarke's (1962) third law in *Profiles of the Future: An Inquiry into the Limits of the Possible* states that "any sufficiently advanced technology is indistinguishable from magic", condensing perceived 'magicality' of technology as a product of its sophistication and modernity (p.21). Machine learning technology has rapidly raised the technological ceiling. Within the span of a decade, its computational potential has grown exponentially and today it is already widely adopted by the public (Henshall, 2023; Ruma and Justice, 2021). With such speed of advancement, people haven't had the time to acclimatize or understand how these systems work. Faced with this, people could begin forming their own interpretations, condensing the inner workings of large-language models to more manageable and familiar understandings.

Interviews gauging misconceptions of large-language models

To gauge people's perceptions of large-language models, I conducted 40 informal interviews to gather different interpretations of their technical workings. In each interview, I asked (1) how often they use large-language models, (2) how confident they feel in knowing how they work, and then (3) to explain, in their own words, how chatbots work. Rather than providing a statistically rigorous analysis, the goal of the survey was to gather different perspectives into how large-language models work.

Generally, people's understandings of large-language models lay on a spectrum from abstract understandings to more accurate ones. Furthermore, they revealed that frequency of use did not correlate with people's understanding of them. Most people, regardless of how often they use chatbots, or how well they think they know how they work, did not accurately describe the process behind them. Out of the 40, only one interviewee accurately mentioned token embedding and four mentioned word-by-word prediction. Next, fourteen people correctly mentioned the need to train large-language models. The nature of this training, however, varied, with some describing it as a manual human-led process while others more accurately described it as the processing of large amounts of data. People's conceptions of the makeup of this training data also ranged from tangible descriptions like it consisting

of information from the internet, to more abstract ones like. The most varied and inconsistent responses, however, lie in people's descriptions of the text generation process and how chatbots arrive at a response. A common conception was that chatbots operate as if they were independent agents, able to retrieve their own database and produce responses through active research and consideration.

> "I assume they have access to databases or something, and when we give an input they sort of scour through those databases and sort of get pieces of information to compile an answer."
>
> *– person who uses ChatGPT multiple times a month, self-reports little, 1/10, understanding of how it works*

> "It's just like, getting information from google, or the internet. It's basically searching for you."
>
> *– person who previously used ChatGPT for university, self-reports mediocre (4-5/10) understanding of how it works*

> "My understanding of it is that it pretty much takes your questions, analyzes the key words … and then scours the internet and any relevant sources to see if this question has been asked before, how its been answered before, and kind of compiles all of that information to then yield an answer"
>
> *– person who uses ChatGPT from time to time, self-reports a good (8/10) understanding of how it works*

People often describe the process of text generation using active human-like verbs like scouring, getting, searching, or gathering, suggesting a research-like process, one of finding various sources, retrieving a breadth of information to assemble a balanced and nuanced response.

> "They gather information from different websites on the internet based on the key words you've searched for"
>
> *– person uses ChatGPT occasionally, self-reports a mediocre (5/10) understanding of how it works*

> "They pull together information from a variety of sources to attempt to answer your questions or provide a service."

– person who does not use ChatGPT themselves but self-reports a strong (9/10) understanding of how it works

"It filters through google searches and websites and gathers the information"

– person who uses ChatGPT occasionally and self-reports a low (3/10) understanding of how it works

Furthermore, people tend to assume the contents of the databases used to train chatbots, often emphasizing academic sources like books and journals:

"My understanding of how it works is that…it basically has access to all sorts of information that humans have created like books but also [it] can access anything that's on the web"

– person who uses ChatGPT at least once a week, self-reports a mediocre (5/10) understanding of how it works

"OpenAI gives the AI various inputs from their 'library' and the AI uses all of these resources to give you an answer."

– person who uses ChatGPT occasionally and self-reports a strong (8/10) understanding of how it works

"From my understanding chatbots take information that's available on line from journals, books, online forums, to answer the question posed."

– person who uses ChatGPT occasionally and self-reports a mediocre (5/10) understanding of how it works

People often tend to talk about databases as an abstract concept, detached from the chatbot. Throughout interviews, people allude to the chatbot's knowledge existing in a metaphysical realm, able to be probed and explored. Furthermore, many also portray text generation as being similar to the process of a google search, suggesting it similar to researching:

"I think this bot … has been given a lot of data, so much data, and it's very smart so it stores all this data or its constantly being refreshed with new data, so when I ask it a question it's basically going through all those resources and pumping it out."

– person who uses ChatGPT multiple times a day, self-reports a decent (6/10) understanding of how it works

> "Based upon the prompt given … the AI will use information from its database/ the internet …and give you an answer. Some AIs use APIs so I'm assuming they gather data from somewhere."

– person who uses ChatGPT daily, self-reports a decent (6/10) understanding of how it works

> "I think they gather all data they can find online like a search engine and then have an algorithm that teaches them to communicate mirroring all the words that we use or something like this"

– person who uses ChatGPT at least once a week, self-reports a low (1/10) understanding of how it works

Exploring people's folk conceptions of large-language models reveals constructed narratives to make sense out of them. These conceptions condense their processes into manageable, more familiar and anthropocentric ones. Culturally, machine learning is still a new concept; it is easier and more natural for people to explain it through traditional processes. Doing so, however, hides core processes of text generation, leaving people unable to engage or challenge it. In reality, the technical workings of these tools reveal various perspectives that necessitate scrutiny and critical thinking.

4. Unpacking folk conceptions and their implications

Chatbots as artificial intelligences

The term artificial intelligence, in its modern use, emerged in the 1950s to delineate a new field of computer research dedicated to replicating human brain functioning (Crevier, 1993; McCorduck, 1977). This new pathway of research led to the development of new processing methods, paving the way for machine learning and natural language processors – underlying technologies behind modern chatbots (Al-Amin *et al*., 2024; Fradkov, 2020). The term artificial intelligence, then, does not represent any specific final form of technology, rather, it represents a wide field of technology built on neural computational processes (Stryker and Kavlakoglu, 2024). Today, however, chatbots, like ChatGPT, are

some of the most frequented avenues for public interaction with machine learning technology, ultimately lending them to wholly adopt the term artificial intelligence (Dilmegani, 2024).

Using the term artificial intelligence to describe large-language models is fundamentally misleading and leaves one vulnerable to misconceptions about how they work. Purely from a vocabulary perspective, AI instills notions of intelligence into a fundamentally computational and mathematical process. Although the field of artificial intelligence is built on the principle of imitating brain processes, its computational processes are far different to human cognition (Korteling *et al*., 2021; Trafton, 2022). Although the definition of intelligence is sometimes reduced to ones like Tegmark's (2017, p. 71) which describes it as the "ability to accomplish complex goals", colloquial use of the term, according to Martinez (2019), ties it to specifically human characteristics. Before engaging with tools labeled as artificial intelligence, then, notions of human-like ways of operating are already preinstalled on a vocabulary basis. This preconception can impede critical engagement with chatbots. Rather than thinking or analyzing, large-language models predict responses on a word-by-word basis. As opposed to notions around human intelligence, which could suggest thoughtfulness and academic rigour, generated text is merely statistically predictable information.

In addition to misperceptions of intelligence, the concept of artificiality also distorts people's understanding and engagement with large-language models. Morozov (2023), for example, argues that the modern machine learning systems are built on the creative and intellectual work of real humans, making it, therefore, more of a "non-artificial intelligence". From this perspective, the label of artificial prevents users from engaging critically with the cultural and human basis of training data. Furthermore, Crawford (2021) also argues against the term artificial, citing its reliance on natural and human resources. Here, the term artificial disconnects chatbots from their real-world physicality and environmental impact. Lastly, 'artificial' could also suggest that chatbots operate in more efficient and capable ways compared to 'organic' human intelligence because of the implied rigidity and computation precision. In reality, however, the computational nature of chatbots is precisely what makes it unreliable – its need to continue predicting outputs encourages the creation of false information, and its statistical interpretation of language entrenches biased perspectives.

User-interfaces for chatbots can also mislead users into perceiving intelligent processes. ChatGPT's user-interface, for example, explicitly simplifies and anthropomorphizes user interaction. Its homepage input screen asks '*what can I help you with?*' alongside a simple search bar prompting the user to '*message ChatGPT*. (OpenAI, 2024a**).** The underlying generative processes creating responses is also hidden by design elements like flashing text carets, a visual commonly used by texting applications,

intended to indicate typing (Zhou, Gallagher and Sterman, 2024). Furthermore, chatbots sometimes display loading throbbers to indicate that the chatbot is processing the prompt or formulating an answer (Yildirim-Erbasli et al., 2023). While these design elements might, to an extent, be functional, they also misconvey the actual processes involved in formulating responses. Typing indicators can suggest that the chatbot is typing like a human would, constructing their ideas. Loading symbols, also, can mislead users into thinking that the chatbot is considering your question, researching, and scouring its database. Together, these perceived actions can lend more credibility to text generation, portraying statistical word-by-word text generation as an intellectual and research-based process. Both the term artificial *intelligence*, and the interface of chatbots, therefore, anthropomorphizes the technology, masks its internal computational processes, and perpetuates the misconception that chatbots are capable of intelligent thought.

The anthropomorphization of chatbots, enforced by the term artificial intelligence and their condensed user-friendly interfaces prevents users from engaging critically. The term artificial impedes people from engaging with the origin and makeup of training datasets, separating it from its very human origins. Artificiality also dematerializes chatbots, preventing people from engaging with their physical and environmental impact. Lastly, the term artificial can also suggest that chatbot processes are more reliable than human cognition, encouraging a lack of critical scrutiny. In addition, the term intelligence encourages the perception of chatbots as wholly intelligent programs, preventing them from engaging critically with how they actually function. All together, the development of critical attitudes towards chatbots needs to challenge the notions instilled by their widely adopted terminologies.

Decontextualized databases

A common tendency behind folk understandings of chatbots is the abstraction of both the computational process and the makeup of data behind text generation. In my interviews, for example, different people describe training data either as “ChatGPT’s database”, or “OpenAI’s …library”, or as a “database brain”. These abstract misconceptions about chatbot databases raise two concerns as they both obscure the text generation process and detach it from its underlying physicality. First, they assume the existence of a dynamic and accessible database all together. In reality chatbots are mostly static; instead of accessing information live with each response, information is pre-embedded during its training phase. Databases do not exist in searchable, indexed states that can be accessed freely when prompted to (Lewis *et al*., 2020). Instead, as explained earlier, knowledge is embedded via the statistical relationships between words. Rather than scouring a database, or searching up relevant information,

truth is derived from probability. If asked to name the planets of our solar system in order, for example, rather than researching astronomy or searching its database, it only 'knows' that statistically, the first most likely word would be Mercury, then Venus, Earth, Mars and so on. Misconceiving the existence of a searchable database of information distorts how chatbots create answers. Believing that they actively search their databases can give the illusion that their responses are based in objective truth and nuanced by the perspective of multiple sources.

Notions of an abstract database generally exemplify the tendency to disconnect the computational process behind chatbots from the real world. In reality, both the training of chatbots and generating text with them are physical, highly demanding, and impactful processes (Falk *et al*., 2024; Smith, 2023). First, the processing of training large-language models – the compiling and analyzing datasets – requires advanced computation on sophisticated hardware over extended periods of time (Mehta, 2024; Peng *et al*., 2023). The energy used to train ChatGPT-3, for example, could have powered more than two million homes for an hour and emitted as much CO2 as a car would from driving two million kilometers (Maslej *et al*., 2023; Ashfield Council, 2021; Ofgem, 2006). These figures, however, are static; once deployed, a single interaction with a large-language model could use up to 25 times more energy than a normal Google search (The Brussels Times Newsroom, 2024). With more than 120 daily users and an estimated 9 billion daily requests, yearly upkeep of large-language models like ChatGPT could require as much annual energy as a small country (de Vries, 2023; Çam, 2024; Singh, 2024).

Folk understandings of chatbots paint them as metaphysical entities that scour databases and the internet much like a person would. This notion sets up an expectation that text generation is the result of complex research that compiles accurate and varied ideas and perspectives. In reality, large-language models only 'interact' with their database once during training. Information is not accessed, it is ingrained in the numeric relationships between words. Generated text, therefore, does not hold the same tenor as researched text which is strengthened by the analysis of various perspectives and sources. The conception of decontextualized databases also assimilates the process of text generation to that of a standard google search and disconnects the process from any physicality altogether. Furthermore, the abstraction of the computational effort behind text generation conceals their immense environmental impact. This aspect of text generation is significant and mandates immense scrutiny and consideration when using large-language models, especially in institutional contexts.

Large-language models as unbiased researchers

Anthropomorphized misconceptions of chatbots – believing that the text generation process involves active research and consideration – also misleads people to believing that generated text is

well-considered and free from error or bias. Furthermore, they place training data on a pedestal, deeming it a representation of collective human-knowledge. In reality, most chatbots do not consult various sources to produce well-thought out and balanced responses. Furthermore, the training data of large-language models is heavily centered on information freely available on the web, third party datasets, and curated information provided by human trainers and researchers (OpenAI**,** 2024b). The training data that powers large-language models, therefore, is not a pure embodiment of human-knowledge but rather a curated slice of it; shaped by the internet, corporate structures, and user-focused development.

Built to continue and repeat patterns of text extracted from existing human literature, text generation is preordained to perpetuate human bias found in its training literature. Just how the relationships between words and concepts are numerically embedded, so are any existing biases. Advanced large-language models require a vast database to operate; ChatGPT began, for example, with 117 million parameters and 5 GB of training data in 2018, while in 2020 it consisted of 175 billion parameters and 45 TB of data (Bi *et al*., 2024; Wu *et al*., 2023). According to Birhane, Prabhu, and Kahembwe (2021), the rush to obtain larger datasets for training raises concerns over "dubious curation practices" as companies turn mostly to collecting data from the web while ignoring its "sordid [data] quality" (p. 1). Concerns over the quality of data used to trained LLMs can be either explicit, or nuanced. Datasets like CommonCrawl, which is used to train LLMs like ChatGPT, have intentionally liberal data collection processes to encourage open-ended research without taking responsibility for its data (Baack, 2024). Data compiled from the web, therefore, often contains problematic content like stereotypes, slurs, and pornography (Birhane, Prabhu, and Kahembwe, 2021). Despite the problematic nature of collected data, however, LLM developers have abandoned selective curation to accommodate for faster, cheaper, and larger datasets, introducing explicitly concerning and biased training data (Birhane, Prabhu, and Kahembwe, 2021). Less explicit but equally important is the hidden bias of internet training data. Webb (2023), for example, discusses how a large portion of ChatGPT's training data comes from the WebText2 dataset which collects text from Reddit forums – a platform with 63.8% of its users being and 47% from the United States. The uneven demographic of who produces WebText2 data is an immediate cause of concern, regardless of whether the data itself is explicitly problematic. Pointing to specific evidence of disproportionate data, however, is overshadowed by the historic paradigm of Western Centric epistemology. Not just on the web, but historically, research, academia, and knowledge in general is inclined towards non-marginalized demographics (Brooks, 2006; Coluzzi, 2022; Harding, 1991). Even improving data collection, or cleaning up datasets, therefore, would still not address the core issue of a technology fundamentally built to learn solely from existing knowledge.

The impact of biased and problematic training data is significant. Wan *et al.* (2023) found high frequencies of bias in 10 different chatbots. Out of them, ChatGPT-3 produced bias in 25% of its responses (Wan *et al.,* 2023). Bias can take several forms in chatbots responses like political bias, racial and stereotype bias, and gender bias. Politically, ChatGPT claims it is neutral and does not hold any particular stance (Rozado, 2023). Several studies, however, have found that the LLM is, in fact, generally left-leaning, while others indicate more nuanced and dynamic positions depending on the issue at hand (Fujimoto and Takermoto 2023; Motoki, Neto and Rodrigues, 2023; Rozado, 2023). Motoki, Neto and Rodrigues (2023) emphasize that regardless of political orientation, political bias in LLMs is particularly hidden, embedded, and difficult to detect and eradicate. Several studies have also shown how chatbots carry particular racial and religious bias. Singh and Ramakrishnan (2023) explore how ChatGPT labeled white men as 'good scientists' over non-white people and chose to disproportionately save children based on race and gender. Haim, Salinas and Nyarko (2024) found that ChatGPT shows negative bias against names associated with racial minorities and women. Large-language models also exemplify heavy gender bias. Zhou and Sanfilippo (2023) found that gender bias in LLMs was reportedly common amongst users and further explored how ChatGPT associates specific genders to different professions. Ghosh and Caliskan (2023) found that ChatGPT perpetuates gender stereotypes with various occupations and struggles with gender-neutral pronouns when translating between languages. Generally, in various applications, large-language models struggle in associating women with positive traits compared to men (Kaplan *et al*., 2024; Troske, Gonzalez and Lawson 2022) Additionally, some studies indicate that even the gender of the user giving a prompt can alter ChatGPT responses (Heaven, 2024; Urchs *et al*., 2023). Although it is difficult to condense the breadth of research into the different forms of bias in large-language models, the point remains. Chatbots are not complex researchers; they are not nuanced in their responses and do not consider a range of perspectives. Moving away from misconceptions of chatbots as active, nuanced, and balanced researchers can allow for users to engage with generated text more critically, understanding that generated text is pre-embedded and fixed with bias.

Large-language models as independent and neutral tools

The anthropomorphization of large-language models, portray them as agents dedicated solely to generating text for users. In reality, chatbots do not prioritize users insofar as they really serve to uphold and protect the goals, interests, and values of their for-profit parent companies. Chatbots like OpenAI's ChatGPT, Google's Gemini, or Meta AI are not technologies, they're products – packaged tools designed to generate profit from consumers.

This shift from technology to product muddles the subliminal goals of chatbots; companies could be much more inclined to prioritize profits over ethics. Indeed, OpenAI quickly transitioned out of a non-profit board structure to a for-profit investible one, seemingly abandoning its original research-focused setup aimed to "benefit humanity' over "financial return" (Hu and Cai 2024; OpenAI, 2015b). The privatization of machine learning companies does, in fact, impact the development of their products. Jurowetzki *et al.* (2021) stresses that a for-profit research environment risks diminishing research into ethical and humanitarian avenues of research if they don't align with a company's commercial interests.

Chatbots as commercial products can directly alter their behaviour, going against the conception of them as independent agents. As a technology, large-language models predict outcomes based on data they're trained on. As a product, however, they're designed to cooperate with for-profit interests of their parent companies. On their own, large-language models are not naturally assistants. To make them conversational and want to help the user, they're given system prompts – background instructions that instruct them on their persona, how to respond to questions, and what they can and cannot do (Mu *et al*., 2024; Qin *et al*., 2024). Here is an expert from the system prompt behind Anthropic's Claude 3.5 Sonnet chatbot:

> "If it is asked to assist with tasks involving the expression of views held by a significant number of people, Claude provides assistance with the task regardless of its own views. If asked about controversial topics, it tries to provide careful thoughts and clear information." (Anthropic, 2024)

OpenAI is more elusive and has not officially disclosed the specific system prompts behind its chatbot, ChatGPT. Various users, however, discovered that certain inputs can 'trick' the chatbot into revealing its system prompts (Reasonable_Oil_1011, 2023; Schwartz, 2024). Among its system prompts are instructions like "Do not create images in the style of artists, creative professionals or studios". Some users also reported system prompts that reveal various 'personalities' ChatGPT could take on. These different versions could alter how ChatGPT responds: one is "formal," "factual", and "academic", another "balanced", "concise" and "helpful", or the last "casual", "friendly", and "relaxed" (Relic180, 2024).

Although system prompts might be important components for successful chatbots, they pose a larger question regarding who directs the flow and passage of knowledge. More than just prompts instructing chatbots to be helpful, system prompts can curate how knowledge is disseminated, and process that

could be altered to uphold company values. Although currently known system prompts serve as safeguards to ensure safe and ethical use of large-language models, they represent a broader picture of chatbots as corporate entities rather than informational tools. Situating large-language models within these corporate frameworks invites a more critical engagement that interrogates their underlying intentions and modes of integration.

5. Educational impact of misconceptions of large-language models

Research into uses of chatbots in education often focus specifically on ChatGPT, exploring its potential to improve learning experiences, either by assisting students, or by aiding teachers. Frequently, these studies seek to improve student learning experiences by using tools like ChatGPT as a virtual teaching assistant. One literature review explores nine studies demonstrating the use of ChatGPT as an intelligent assistant capable of providing on-demand answers, information, and learning resources, and immediate feedback (Albadarin *et al*., 2024). Additionally, another literature review of educational chatbots cites various studies exploring how the flexibility of ChatGPT allows it to open up personalized learning opportunities that adapt to student needs (Yan *et al*., 2023). Sok and Heng (2024) and Baig and Yadegaridehkordi (2024) expand on ChatGPT as an assistant, citing various studies exploring its potential to aid research by helping brainstorm, research, and analyze data. Furthermore, Albadarin *et al.* (2024) points out six studies exploring the use of ChatGPT as writing and language assistant, helping students develop their writing and overcome language barriers.

Studies also explore the applications of ChatGPT to aid teachers, focusing on its potential to facilitate teaching practices and alleviate workloads. Albadarin et al., (2024) cite various studies exploring teachers using ChatGPT to create lesson plans, quizzes, additional resources, and to help answer student questions. According to these various studies, utilizing chatbots helped support teaching and learning practices, and enhanced teacher productivity. Albadarin et al. (2024) also explores ChatGPT's potential to support alternative learning methods and foster engagement and motivation to enhance teaching. Sok and Heng (2024) explore enhanced teaching via studies that cite ChatGPT's ability to provide throughout instruction, stimulating, and collaborative activities. Expanding on its potential to improve teacher productivity, Yan et al., (2023) cite various studies exploring the use of ChatGPT to automate time-consuming administrative tasks like feedback provision and scoring essays. In addition to helping grade, however, Sok and Heng (2024) cite various studies that suggest using ChatGPT to innovate assessment methods themselves, claiming it can create recurring learning evaluations, immediate quizzes, and quick feedback.

While studies do explore possible negative impacts of chatbots in education, they do not discuss the underlying need to address how people, especially educators, perceive large-language models. This perspective is needed as educators play a vital role in the safeguarding of new educational technologies. As leading figures in educational contexts, the way educators approach, and discuss large-language models strongly influences how they're incorporated into learning environments.

Impact of misconceptions on educators

Misconceptions about how large-language models can prevent educators from using them in a critically motivated way that prevents misuse the perpetuation of dangerous narratives. Current research explores and somewhat encourages the use of chatbots like ChatGPT to help teachers improve teaching practices and lessen their workload. Although some studies, like that of Iqbal et at. (2022) suggest that teachers are still hesitant to use chatbots, others demonstrate that already more than 50% of teachers currently employ them (Hallahan, 2024). Misconceptions in teachers of how chatbots work can skew the processes behind text generation, hiding its ability to perpetuate bias and enforce narratives, provide reliable and well-researched information, and allow for corporate influence over knowledge.

Studies forecast an educational future deeply shaped by AI, where chatbots play a significant role in influencing both teaching methods and educational content. Proposals for their integration include assisting educators in designing lesson plans and implementing novel pedagogical approaches ElSayary, 2023; Albadarin *et al*., 2024). Furthermore, studies tend to focus on teacher-specific approaches towards incorporation, rather than institutional ones. This angle of research, while still needed, puts teachers in a vulnerable position to single-handedly know and deal with the shortcomings of large-language models. Placing chatbot integration in the hands of teachers rather than institutions necessitates the demystification of machine learning at the educator level. Altogether, both the scope of large-language models influencing course material and teaching methods, and the undue responsibility put on teachers stresses the importance to strengthen their technical understanding of machine learning technology and move past their folk conceptions of how they work.

Several of the previously explored misconceptions could significantly affect how educators use these tools in their teaching practice. False notions of intelligence and active consideration, can lead teachers to overestimate the process of text generation. This notion fosters misplaced trust that can result in educators unintentionally perpetuating inaccuracies or bias. Furthermore, false notions of existing databases that can be actively searched by chatbots can lead teachers to trust generated information on the basis that it is grounded in a wide corpus of academic text. This misconception can cloud the reality

that most chatbots are trained on data with little scrutiny and vetting and full of harmful material and biased perspectives.

Equipping teachers with a comprehensive understanding of the internal mechanics of chatbots enables critical scrutiny of the corporate influence on educational content and teaching practices and sharpens the focus on the ethical practices of companies developing these tools. Many advanced machine learning chatbots offer both free, but limited versions, and premium, paid models of their chatbots (Google, n.d; OpenAI, n.d; Anthropic, n.d). Although free versions are accessible, the differences between the two versions create several equity concerns. OpenAI's ChatGPT's paid model 4.0, not only outperforms its free 3.0 model in terms of speed and proficiency, but also on the basis of reproducing less bias and hallucinations (Hanna and Levic 2023; Pantana, Castello and Torre, 2024). Furthemore, GPT 4.0 tended to provide more detail into its reasoning and decision-making, offering more transparency and accountability than its free counterpart (Pantana, Castello and Torre, 2024). Paywalls, therefore, explicitly curate who can access reliable, truthful, and ethical machine learning tools. Paywalls also widen access barriers to marginalized and underprivileged people who already face technological barriers like access to the internet or functioning equipment, further imposing division on who can access machine learning tools. A Project Future (2024) survey, for example, found that the cost of subscription services for chatbots were noted as the most common and significant barrier in Nigeria and Egypt. Decreasing accessibility to large-language models, therefore, directly impacts already marginalized communities, deepening division and upholding oppressive status-quos. Furthermore, the privatization of machine learning tools also feeds users into new digital economies centered around the exploitation of personal data. Zuboff in The Age of Surveillance Capitalism (2019) describes this economy as surveillance capitalism: a "parasitic…mutation of capitalism" that "claims human experience" for "hidden commercial practices of extraction, prediction, and sales." Surveillance capitalism thrives on the systematic collection and monetization of personal data; users become raw material for corporate exploitation rather than beneficiaries of their technologies. According to Tacheva and Ramasubramanian (2023, p.9), machine learning technology, described as an "AI Empire" relies on the same principles of "constant surveillance". Educators, therefore, equipped with reality-based perceptions of chatbots, can approach their commercial aspects with critical awareness, ensuring thoughtful and deliberate implementation.

Dismantling misconceptions to strengthen teaching

Promoting better understandings of text generation equips teachers with technically-supported critical frameworks towards large-language models. With a critical lens, teachers can actively challenge and engage with generated text, questioning its perspectives and reasoning. From this perspective, teachers

can proactively use chatbots to create opportunities for critical education by encouraging collaborative scrutiny of generated text – treating it not as a perfect representation of knowledge, but as something meant to be dissected and analyzed.

Improved technical understanding of the inner workings of chatbots allow for teachers to perceive and challenge the role of chatbots in perpetuating existing narratives and perspectives. Understanding them as predictive rather than creative can encourage the widening of progressive education. Heightened awareness into the training process of chatbots can promote the pursuit for a more varied and inclusive academic perspective into classrooms to counter the limited scope of training data. Specifically, educators could strengthen their curricula by incorporating queer, feminist, anti-racist, and ecocentric perspectives to challenge the outdated narratives perpetuated by large-language models. While this should also be done independently, they could also be actively done collaboratively. with large-language models, encouraging critical engagement with machine learning from an emancipatory perspective.

A deepened awareness into how large-language models work can also evoke consideration into the overall objectives of education. Educational theory and the development of new educational practices is often structured around emancipatory goals. The pursuit for new methods of education often emphasize breaking down the rigidity and divisiveness of traditional education (Houle, 1974). Bruner (1996), however, in *The Culture of Education*, highlights a common contradiction in this objective. On one hand, Bruner acknowledges how education is intended to equip people with the necessary tools for growth and opportunity. On the other hand, education is also meant to reproduce the culture that permits it. Essentially, Bruner pinpoints a constant tension in education between trying to advance people and culture while simultaneously being held back to uphold the status quo (Bruner, 1996). Large-language models, when used in an educational context, explicitly manifest this contradiction. Trained on existing literature, chatbots, by design, are fundamentally retrospective. Through their imitation of patterns in words and text, chatbots technically and theoretically are constrained by what already is. Although it can be argued that the computational process of text generation is similar to human cognition, as we also rely on reconstructive thinking, a phenomenological perspective would emphasize the importance of human experience and intentionality over large-language models in the creation of knowledge (Karvonen *et al*., 2023; Sörbom, 2008). From this angle, educators –, emboldened with a deeper awareness into the predictiveness and perpetual nature of generated text, – can seek to contextualize the use of large-language models to encourage human-led knowledge creation.

6. Conclusion

In *A Manifesto for a Pro-Actively Responsible AI in Education*, Porayska-Pomsta (2023) begins their manifesto with two specific calls for education AI research: first is to "inform and challenge", and second to "expand the field of view" (Porayska-Pomsta, 2023, p.81). The first addresses the need for more critical research that aspires for educational improvements while safeguarding it from harmful applications. The second point calls for a wider scope of research to include more human-centric perspectives on impact and mitigation. In line with this approach, this dissertation explored large-language models from a user-centric perspective to identify ways of thinking that would prevent people from challenging educational AI.

Folk perceptions of large-language models address current mental frameworks that inevitably inform how people use, interact, and think about them. This approach introduces an informal, yet realistic perspective that applies a human angle to educational AI research to pursue more critical avenues. In this dissertation, investigating people's folk understandings of large-language models revealed common conceptions that could prevent people from engaging with controversial and harmful facets of machine learning. Examining folk perceptions of large-language models reveals how anthropomorphized conceptions can foster ideas of chatbots as intelligent, neutral, and independent tools. These misconceptions obscure their computational predictiveness, perpetuate trust in biased training data, and mask their corporate and environmental realities. Decontextualized notions of databases further deepen this illusion, creating expectations of research-based text generation while disengaging users from its physical, environmental and exploitative impacts.

Rooting the discussion around people's conceptions of large-language models offers a needed angle to educational AI research; many literature and scope reviews of the field reveal a tendency to approach educational large-language models from a techno-centric angle. While technical research into mitigating fallacies of educational AI is important, it is equally paramount to leave out the role teachers and students have in mitigation. Ignoring this paints a false reality of educational AI as something beyond the scope of human intervention, meant to be instilled in education from the top down with no regard to people affected. In reality, teachers and students will always need to be actively involved in framing its use.

The educational integration of large-language models is largely seen as inevitable (Krašna and Bratina, 2024). Both students and educators, whether allowed to or not, are bound to employ these new and powerful tools in educational contexts. Given the inevitability of large-language incorporation into

education, it becomes necessary for educators and educational institutions to consider deeper, more ingrained approaches to mitigating harm. This dissertation offers a specific, informal, yet ingrained approach by highlighting the importance of addressing folk misconceptions of large-language models. This approach serves as a stepping stone towards the development of critical attitudes towards artificial intelligence. From here, educators can be encouraged not only to actively engage with teaching practices involving large-language models but also to embrace generated text as a means to advance critical pedagogy.

Bibliography

Al-Amin, M., Ali, M.S., Salam, A., Khan, A., Ali, A., Ullah, A., Alam, M.N. and Chowdhury, S.K., (2024). *History of generative Artificial Intelligence (AI) chatbots: past, present, and future development*. doi:10.48550/arXiv.2402.05122.

Albadarin, Y., Saqr, M., Pope, N. and Tukiainen, M. (2024). *A systematic literature review of empirical research on ChatGPT in education.* Discover education, 3. doi:10.1007/s44217-024-00138-2.

Anthropic (2024). *Release Notes: System Prompts*. [online] Anthropic.com. Available at: https://docs.anthropic.com/en/release-notes/system-prompts#nov-22nd-2024 (Accessed: 28 November 2024).

Anthropic (n.d.). *Claude*. [online] Claude.ai. Available at: https://claude.ai/upgrade (Accessed: 14 November 2024).

Ashford District Council. (2021). *Climate Change Strategy 2021-2026*. [online] ashfield.gov.uk. United Kingdom: Ashfield District Council. Available at: https://www.ashfield.gov.uk/media/tcuhqixg/climate-change-strategy-2021-to-2026-v1.pdf. (Accessed: 4 December 2024).

Baack, S. (2024). *A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl*. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro, Brazil. New York, NY, USA: Association for Computing Machinery, pp. 2199–2208. doi:10.1145/3630106.3659033

Baig I. M. and Yadegaridehkordi E. (2024). *ChatGPT in the higher education: A systematic literature review and research challenges*. International Journal of Educational Research, 127, p.102411. doi:10.1016/j.ijer.2024.102411.

Bi, Z., Zhang, N., Jiang, Y., Deng, S., Zheng, G. and Chen, H., (2024). March. *When Do Program-of-Thought Works for Reasoning?*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 16, pp. 17691-17699).

Birhane, A., Prabhu, V.U. and Kahembwe, E., (2021). *Multimodal datasets: misogyny, pornography, and malignant stereotypes*. doi:10.48550/arXiv.2110.01963

Brooks, A. (2007). *Feminist Standpoint Epistemology. In: Feminist Research Practice*. Sage Research Methods, pp.53–82. doi:10.4135/9781412984270.n3.

Bruner, J.S. (1996). *The Culture of Education*. Cambridge: Harvard University Press.

Çam, E., Hungerford, Z., Schoch, N., Miranda, F.P. and Yáñez de León, C.D. (2024). *Electricity 2024: Analysis and forecast to 2026*. [online] iea.org. International Energy Agency. Available at:

https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf. (Accessed: 4 December 2024).

Chalmers, D.J., 2023. Could a large language model be conscious?. arXiv preprint arXiv:2303.07103.

Clarke, A.C. (1973). *Profiles of the Future; an Inquiry Into the Limits of the Possible*. New York: Harper & Row.

Coluzzi, P. (2022). *Western-centricity in Academia: How International Journals Endorse Inner Circle Englishes and a European-American Worldview*. Crossings: A Journal of English Studies, 13(1), 36–42. doi:10.59817/cjes.v13i1.17

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. USA: Yale University Press.

Crevier, D. (1993). *AI: The Tumultuous History of the Search for Artificial Intelligence*. [online] BasicBooks. Available at: https://www.researchgate.net/publication/233820788. (Accessed: 10 November 2024).

de Vries, A. (2023). *The growing energy footprint of artificial intelligence.* Joule, 7(10), pp. 2191–2194. doi:10.1016/j.joule.2023.09.004

Dharan, G. and Nanda, S. (2021). *A SYSTEMATIC STUDY AND DEPLOYING OF A NOVICE CHATBOT*. International Journal of Research and Analytical Reviews, [online] 8(1). Available at: https://ijrar.org/papers/IJRARJFM1006.pdf. (Accessed: 15 November. 2024).

Dilmegani, C. (2024). *90+ Chatbot/Conversational AI Statistics*. [online] AIMultuple Research. Available at: https://research.aimultiple.com/chatbot-stats/?utm_source=chatgpt.com (Accessed: 10 November 2024).

ElSayary, A. (2023). *An investigation of teachers' perceptions of using ChatGPT as a supporting tool for teaching and learning in the digital era*. Journal of Computer Assisted Learning, 40(3), 931–945. doi:10.1111/jcal.12926

Falk, S., van Wynsberghe, A. and Biber-Freudenberger, L. (2024). *The attribution problem of a seemingly intangible industry. Environmental Challenges,* 16, p.101003. doi:10.1016/j.envc.2024.101003

Fradkov, A.L. (2020). *Early History of Machine Learning*. IFAC-PapersOnline, 53(2). doi:10.1016/j.ifacol.2020.12.1888.

Fujimoto, S. and Takemoto, K. (2023). *Revisiting the political biases of ChatGPT.* Frontiers in Artificial Intelligence, 6, 1232003. doi:10.3389/frai.2023.1232003

Ghosh, S. and Caliskan, A., (2023). *Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages*. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 901-912).

Google (n.d.). *Gemini Advanced - get access to Google's most capable AI models*. [online] Gemini. Available at: https://gemini.google/advanced/?hl=en-GB (Accessed: 25 November 2024].

Haim, A., Salinas, A. and Nyarko, J. (2024). *What's in a Name? Auditing Large Language Models for Race and Gender Bias*. arXiv preprint arXiv:2402.14875.

Hallahan, G. (2024). *AI teachers, school exclusions and cutting workload*. [online] Teacher Tapp. Available at: https://teachertapp.com/uk/articles/ai-teachers-school-exclusions-and-cutting-workload/ (Accessed: 26 November 2024).

Hanna, E. and Levic, A. (2023). *Comparative Analysis of Language Models: hallucinations in ChatGPT: Prompt Study*. [Thesis] Available at: https://www.diva-portal.org/smash/get/diva2:1764165/FULLTEXT01.pdf (Accessed: November 28 2023).

Harding, S. (1991). *Whose Science? Whose Knowledge?: Thinking from Women's Lives*. Cornell University Press. Available at: http://www.jstor.org/stable/10.7591/j.ctt1hhfnmg (Accessed: November 22 2023).

Heaven, W.D. (2024). *OpenAI says ChatGPT treats us all the same (most of the time).* [online] MIT Technology Review. Available at: https://www.technologyreview.com/2024/10/15/1105558/openai-says-chatgpt-treats-us-all-the-same-most-of-the-time/#:~:text=OpenAI%20has%20analyzed%20millions%20of,responses%20in%20the%20worst%20case. (Accessed: November 11 2023)

Henshall, W. (2023). *4 Charts That Show Why AI Progress Is Unlikely to Slow Down*. [online] Time. Available at: https://time.com/6300942/ai-progress-charts/ (Accessed: 11 Nov. 2024).

Hicks, M.T., Humphries, J. and Slater, J. (2024). *ChatGPT is bullshit*. Ethics Inf Technol 26, 38 . doi:10.1007/s10676-024-09775-5

Houle, C.O. (1974). *The Changing Goals of Education in the Perspective of Lifelong Learning*. International Review of Education / Internationale Zeitschrift Für Erziehungswissenschaft / Revue Internationale de l'Education, 20(4), pp. 430–446. Available at: http://www.jstor.org/stable/3443019 (Accessed: 6 December 2024).

Hu, K. and Cai, K. (2024). *Exclusive: OpenAI to remove non-profit control and give Sam Altman equity*. [online] Reuters. Available at: https://www.reuters.com/technology/artificial-intelligence/openai-remove-non-profit-control-give-sam-altman-equity-sources-say-2024-09-25/. (Accessed: 28 October 2024).

Jurowetzki, R., Hain, D., Mateos-Garcia, J. and Stathoulopoulos, K. (2021). *The Privatization of AI Research (-ers): Causes and Potential Consequences--From university-industry interaction to public research brain-drain?*. arXiv preprint arXiv:2102.01648.

Kaplan, D.M., Palitsky, R., Arconada Alvarez, S.J., Pozzo, N.S., Greenleaf, M.N., Atkinson, C.A. and Lam, W.A. (2024). *What's in a Name? Experimental Evidence of Gender Bias in*

Recommendation Letters Generated by ChatGPT. Journal of Medical Internet Research, 26, e51837.doi:10.2196/51837

Karvonen, A., Kujala, T., Kärkkäinen, T. and Saariluoma, P. (2023). *Fundamental concepts of cognitive mimetics*. Cognitive Systems Research, 82, p.101166. doi:10.1016/j.cogsys.2023.101166

Korteling, J.E.H., van de Boer-Visschedijk, G.C., Blankendaal, R.A.M., Boonekamp, R.C. and Eikelboom, A.R. (2021). *Human- versus Artificial Intelligence*, Frontiers in Artificial Intelligence, 4, 622364. doi:10.3389/frai.2021.622364.

Krašna, M. and Bratina, T. (2024). *The use of AI and student population: The change is inevitable*. In: 2024 13th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro. IEEE, pp. 1–5. doi:10.1109/MECO62516.2024.10577853

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., (2020). *Retrieval-augmented generation for knowledge-intensive nlp tasks*. Advances in Neural Information Processing Systems, 33, pp.9459-9474. doi:10.48550/arXiv.2005.11401

Martinez, R. (2019). *Artificial Intelligence: Distinguishing Between Types & Definitions.* Nevada Law Journal, [online] 19(3). Available at: https://scholars.law.unlv.edu/nlj/vol19/iss3/9 (Accessed: 7 November 2024).

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J.C., Parli, V., Shoham, Y., Wald, R., Clark, J. and Perrault, R. (2023). *The AI Index 2023 Annual Report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

McCorduck, P., Minsky, M., Selfridge, O., Simon, H., (1977). *History of Artificial Intelligence.* International Joint Conference on Artificial Intelligence Available at: https://www.ijcai.org/Proceedings/77-2/Papers/083.pdf (Accessed: 16 October 2024).

Mehta, S. (2024). *How Much Energy Do LLMs Consume? Unveiling the Power Behind AI*. [online] Association of Data Scientists. Available at: https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/. (Accessed: 4 December 2024).

Morozov, E. (2023). *The problem with artificial intelligence? It's neither artificial nor intelligent.* The Guardian. [online] 30 Mar. Available at: https://www.theguardian.com/commentisfree/2023/mar/30/artificial-intelligence-chatgpt-human-mind. (Accessed: 4 November 2024).

Motoki, F., Neto, V.P. and Rodrigues, V. (2023). *More human than human: measuring ChatGPT political bias*. Public Choice, 198, pp.3–23. doi:10.1007/s11127-023-01097-2.

Mu, N., Lu, J., Lavery, M. and Wagner, D. (2024). *A Closer Look at System Message Robustness.* In: Neurips Safe Generative AI Workshop 2024. Available at: https://openreview.net/forum?id=YZqDyqYwFf (Accessed: 28 November 2024).

Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. and Mian, A., 2023. A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

Nilsen, V. (2018). An Introduction to Machine Learning. [online] Available at: https://bafflerbach.github.io/test_software_carpentry/files/Module%201%20Lecture%20Notes.pdf [Accessed 4 Jul. 2024].

Newport, C. (2023). *What Kind of Mind Does ChatGPT Have?* The New Yorker. [online] Apr. Available at: https://www.newyorker.com/science/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have (Accessed: 11 November 2024).

Ofgem (2006). *Electricity generation: facts and figures*. [online] ofgem.gov.uk. United Kingdom: Ofgem. Available at: https://www.ofgem.gov.uk/sites/default/files/docs/2006/04/13537-elecgenfactsfs.pdf. (Accessed: 4 December 2024).

OpenAI (2015a). *ChatGPT for Enterprise*. [online] Openai.com. Available at: https://openai.com/chatgpt/enterprise/. (Accessed: October 25th 2024).

OpenAI (2015b). *Introducing OpenAI.* [online] Openai.com. Available at: https://openai.com/index/introducing-openai/. (Accessed: 11 October 2024).

OpenAI (2024a). *ChatGPT*. [online] Chatgpt.com. Available at: https://chatgpt.com/?model=auto. (Accessed: 10 October 2024).

OpenAI (2024b). *How ChatGPT and our foundation models are developed*. [online] Openai.com. Available at: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed (Accessed: 12 October 2024).

OpenAI (n.d.). *ChatGPT*. [online] Chatgpt.com. Available at: https://chatgpt.com/#pricing. (Accessed: 13 November 2024).

Pantana, G., Castello, M. and Torre, I. (2024). *Examining Cognitive Biases in ChatGPT 3.5 and 4 through Human Evaluation and Linguistic Comparison.* In: Association for Machine Translation in the Americas. [online] 16th Conference of the Association for Machine Translation in the Americas. Available at: https://aclanthology.org/2024.amta-research.21.pdf. (Accessed: 15 November 2024).

Patil, D. (2024) *Machine learning and deep learning: Methods, techniques, applications, challenges, and future research opportunities*, in Patil, D. et al. (eds.) Trustworthy Artificial Intelligence in Industry and Society. Deep Science Publishing, pp. 28–81. doi:10.70593/978-81-981367-4-9_2.

Payne, Annick (2023). *The Kingdom of Lydia*, in Karen Radner, Nadine Moeller, and D. T. Potts (eds), *The Oxford History of the Ancient Near East Volume V: The Age of Persia*. New

York: online edn, Oxford Academic, 23 Mar. 2023), doi:10.1093/oso/9780190687663.003.0051.

Peng, H., Davidson, S., Shi, R., Song, S.L. and Taylor, M. (2023). *Chiplet cloud: Building ai supercomputers for serving large generative language models.* arXiv preprint arXiv:2307.02666.

Pettit, J. (2023). *Do you know how ChatGPT tokens work?* [online] SSW Rules. Available at: https://www.ssw.com.au/rules/gpt-tokens/ (Accessed: 11 Nov. 2024).

Porayska-Pomsta, K. (2023). *A Manifesto for a Pro-Actively Responsible AI in Education.* International Journal of Artificial Intelligence in Education, 34, pp.73–83. doi:10.1007/s40593-023-00346-1.

Project Future (2024). *Publicly Available AI Access and Integration in African Higher Education: Usage, Impact, and Barriers in the Continent's Largest Regional Economies*. doi:10.13140/RG.2.2.18593.21601

Qin, Y., Zhang, T., Shen, Y., Luo, W., Sun, H., Zhang, Y., Qiao, Y., Chen, W., Zhou, Z., Zhang, W. and Cui, B. (2024). SysBench: *Can Large Language Models Follow System Messages*?. arXiv preprint arXiv:2408.10943.

Reasonable_Oil_1011 (2023). *SECRET INSTRUCTIONS* [Reddit post], Reddit. Available at: https://www.reddit.com/r/ChatGPT/comments/17wuqfh/secret_instructions/. (Accessed: 15 November 2024).

Reddy, K.K., Reddy, P.S., Pilly, A. and Doss, S. (2024). Transformative Effects of Smarter Chatbots. *Artificial Intelligence‐Enabled Businesses: How to Develop Strategies for Innovation*, pp.333–350. doi:https://doi.org/10.1002/9781394234028.ch18.

Relic180 (2024). *'FYI...Me: It appears your enabled personality is v2. Is…'* [Reddit post], Reddit. Available at: https://www.reddit.com/r/ChatGPT/comments/1ds9gi7/comment/lb18txj/?utm_source=share&utm_medium=web3x&utm_name=web3xcss&utm_term=1&utm_content=share_button accessed. (Accessed: 15 November 2024).

Roose, K., (2022). *The Brilliance and Weirdness of ChatGPT: The Shift*. [online] New York: New York Times Company. Available at: https://www.proquest.com/docview/2746624142/3A90487A1DAD489CPQ/1?%20Websites&sourcetype=Blogs,%20Podcasts,%20 (Accessed: 9 November 2023).

Rozado, D. (2023). *The Political Biases of ChatGPT.* Social Sciences, 12(3), p. 148. doi:10.3390/socsci12030148 (

Ruma and Justice, (2021). [Podcast] *Embracing the rapid pace of AI*. Insights. 19 May Available at: https://www.technologyreview.com/2021/05/19/1025016/embracing-the-rapid-pace-of-ai/ (Accessed: 24 October 2024).

Sappaile, B. I., Vandika, A Y., Deiniatur M., Nuridayanti, Arifudin O. (2024). *The Role of Artificial Intelligence in the Development of Digital Era Educational Progress*. [online] Available at: https://edujavare.com/index.php/JAI/article/view/297/249 (Accessed: 14 September 2024).

Schwartz, E.H. (2024). *ChatGPT just (accidentally) shared all of its secret rules – here's what we learned.* [online] TechRadar. Available at: https://www.google.com/url?q=https://www.techradar.com/computing/artificial-intelligence/chatgpt-just-accidentally-shared-all-of-its-secret-rules-heres-what-we-learned&sa=D&source=docs&ust=1734187833070748&usg=AOvVaw1ajZD2SNKqaSK1EtXlCwFp (Accessed: 28 November. 2024).

Shanahan, M. (2024). Talking about Large Language Models. Communications of The ACM, 67(2), pp.68–79. doi:https://doi.org/10.1145/3624724.

Singh, S. (2024). *ChatGPT Statistics for 2023: Comprehensive Facts and Data*. [online] Demandsage. Available at: https://www.demandsage.com/chatgpt-statistics. (Accessed: 4 December 2024).

Singh, S. and Ramakrishnan, N. (2023). *Is ChatGPT Biased? A Review*. doi:10.31219/osf.io/9xkbu.

Smith, C.S. (2023). *What Large Models Cost You – There Is No Free AI Lunch.* [online] Forbes. Available at: https://www.forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch/. (Accessed: 4 December 2024).

Sok, S. and Heng, K. (2024). *Opportunities, challenges, and strategies for using ChatGPT in higher education: A literature review*. Journal of digital educational technology, 4(1), pp.ep2401–ep2401. doi:10.30935/jdet/14027.

Sörbom, G. (2002). *The Classical Concept of Mimesis*. In *A Companion to Art Theory* (eds P. Smith and C. Wilde). doi:10.1002/9780470998434.ch2

Stryker, C. and Kavlakoglu, E. (2024). What is artificial intelligence (AI)? [online] IBM. Available at: https://www.ibm.com/think/topics/artificial-intelligence. (Accessed: 10 September 2024).

Taecharungroj, Viriya. (2023). *What Can ChatGPT Do? Analyzing Early Reactions to the Innovative AI Chatbot on Twitter*. Big Data and Cognitive Computing 7, no. 1: 35. doi:/10.3390/bdcc7010035

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York, USA: Borzoi Book.

The Brussels Times Newsroom (2024). *ChatGPT consumes 25 times more energy than Google*. [online] The Brussels Times. Available at: https://www.brusselstimes.com/1042696/chatgpt-consumes-25-times-more-energy-than-google (Accessed: 4 December 2024).

Trafton, A. (2022). *Study urges caution when comparing neural networks to the brain*. [online] MIT News. Available at: https://news.mit.edu/2022/neural-networks-brain-function-1102.(Accessed: 10 November 2024)

Troske, A., Gonzalez, E. and Lawson, N. (2022) *Brilliance Bias in GPT-3.* Computer Science and Engineering Senior Theses, 221. Available at: https://scholarcommons.scu.edu/cseng_senior/221 (Accessed: 14 November 2024).

Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C. and Thiemichen, S. (2023). *How Prevalent is Gender Bias in ChatGPT?--Exploring German and English ChatGPT Responses*. arXiv preprint arXiv:2310.03031.

Walsh, Lynda (2013). *The Delphic Oracle and Ancient Prophetic* in Ethos Scientists as Prophets: A Rhetorical Genealogy. New York: online edn, Oxford Academic. doi:10.1093/acprof:oso/9780199857098.003.0002

Wan, Y., Wang, W., He, P., Gu, J., Bai, H. and Lyu, M.R. (2023) *BiasAsker: Measuring the Bias in Conversational AI System*. In: Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2023), New York, USA: Association for Computing Machinery, pp. 515–527. doi: 10.1145/3611643.3616310

Webb, M. (2023). *Exploring the potential for bias in ChatGPT*. [online] Jisc. Available at: https://nationalcentreforai.jiscinvolve.org/wp/2023/01/26/exploring-the-potential-for-bias-in-chatgpt/. (Accessed: 12 September 2024)

Wolfram, S. (2023). *What Is ChatGPT Doing … and Why Does It Work?* [online] writings.stephenwolfram.com. Available at: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/. (Accessed: 18 September 2024).

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y. and Gašević, D. (2023). *Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review*. British Journal of Educational Technology, 55(1), pp.90–112. doi:10.1111/bjet.13370.

Yates, D. C. (2023). *Remembering and Forgetting the Sack of Athens i*n: *Collective Violence and Memory in the Ancient Mediterranean.* Leiden, The Netherlands: Brill. doi: 10.1163/9789004683181_011 (Accessed: 10 December 2024).

Yildirim-Erbasli, S. N., Bulut, O., Demmans Epp, C., and Cui, Y. (2023). *Conversation-Based Assessments in Education: Design, Implementation, and Cognitive Walkthroughs for Usability Testing*. Journal of Educational Technology Systems, 52(1), 27-51. doi: 10.1177/00472395231178943

Zhou, D., Gallagher, J. and Sterman, S. (2024). *Thoughtful, Confused, or Untrustworthy: How Text is Displayed Influences Perceptions of Generative AI Tools*. [online] Available at:

https://david23.web.illinois.edu/wp-content/uploads/2024/04/presentation_style_preprint.pdf. (Accessed: 20 September 2024).

Zhou, K.Z. and Sanfilippo, M.R. (2023). *Public perceptions of gender bias in large language models: Cases of chatgpt and ernie*. arXiv preprint arXiv:2309.09120.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. [online] New York: Public Affairs. Available at: https://edisciplinas.usp.br/pluginfile.php/5594205/mod_resource/content/1/Shoshana-Zuboff-The-Age-of-Surve_INTRO.pdf (Accessed: 16 November 2024).