Chapter 2 – Introduction to Generative AI

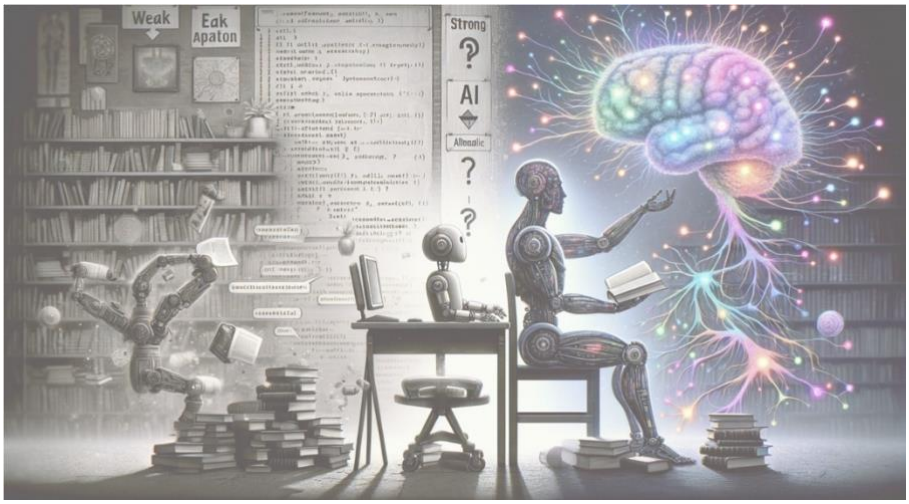
2.1 Types of AI

2.1 Types of AI	
Activity type	Page
Contents	Artificial Intelligence (AI) is classified into two categories: **Weak AI** and **Strong AI**. **Weak AI**, also known as narrow AI, is an AI system that is designed and trained for a particular task. For instance, large language models (LLMs) such as ChatGPT have been trained on large data sets to mimic human language. **Strong AI**, also known as Artificial General Intelligence (AGI), is an AI system with human-like cognitive abilities, which would allow them to perform any intellectual task a human can. They would have self-awareness, consciousness, and even emotional understanding. Presently, AGI remains a theoretical concept, however, numerous companies are actively pursuing its development. **AGI Benchmarking** Our fascination with AGI goes back thousands of years. Early theorists such as Buddhist monk Daoxuan (596-667 BCE), Greek philosopher Aristotle (384-322 BCE), English scholar Roger Bacon (1219/20-1292), and French philosopher René Descartes (1596-1650), envisioned and even attempted to create mechanical devices designed to emulate human actions. Mary Ann Evans (1819-1880), better known as George Eliot, writing in the period of the Second Industrial Revolution, saw the power and potential of machine automation and its impact not just on the economy but also on human intelligence. Her work, *Impressions of Theophrastus*, also explores how such technology could 'render the human race obsolete'. The mid-20th century marked a pivotal moment in these discussions

with Alan Turing's important 1950 paper, "Computing Machinery and Intelligence," introducing the concept of the 'imitation game,' now known as the Turing Test. Turing argued that if a machine could engage in a text-based conversation indistinguishable from a human, it could be considered intelligent. While pioneering, the Turing Test has limitations, as it primarily evaluates linguistic imitation rather than true understanding.

To better evaluate AI's capabilities, contemporary benchmarks have been developed. The Abstraction and Reasoning Corpus (ARC)**, which tests reasoning skills, and the** AGIEval**,** which evaluates AI models using human-centric standardized exams, such as college entrance tests and law qualifications, to measure their general abilities in real-world scenarios. More recently, the GSM-Symbolic benchmark has been developed to specifically measure how well AI models handle mathematical reasoning tasks, revealing significant gaps in current systems' abilities to adapt to even minor variations in problem statements.

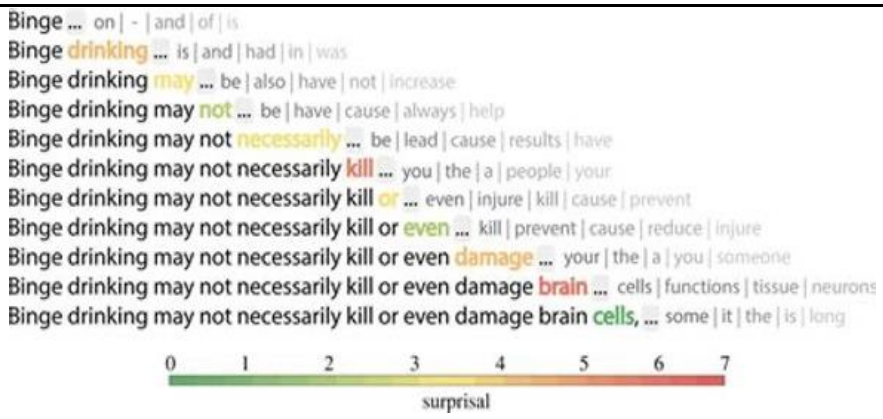
While there are a number of other types of benchmarks, perspectives on AGI's proximity vary. Some researchers anticipate its emergence within the next few years, citing rapid advancements in AI capabilities. Others argue that current systems, despite their progress, remain far from achieving true general intelligence, and emphasize the complexity of human cognition. This ongoing debate highlights the importance of developing robust benchmarks to accurately assess AI's reasoning and common-sense capabilities. It also requires us to consider what makes human thinking distinct.



(Image produced in DALL-E December 2023)

2.2. Large Language Models (LLM's)

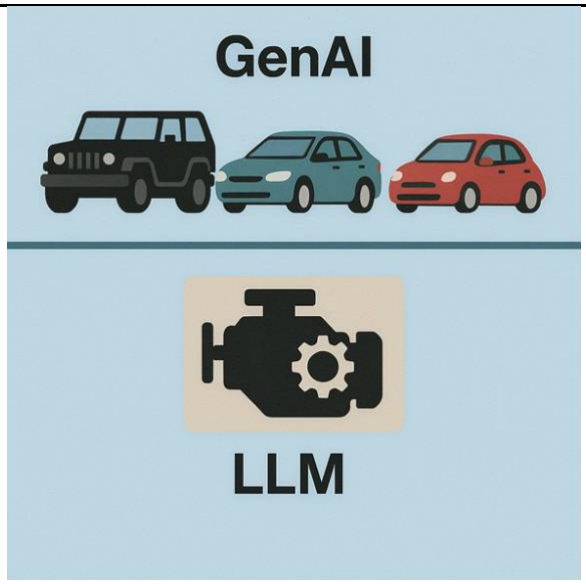
2.2. Large Language Models (LLM's)	
Activity type	Page
Contents	**How LLMs Work** Imagine playing a game of WORDLE, where you have to guess a five-letter word by entering different combinations. In the game, a yellow letter means you've guessed a correct letter but in the wrong place, and a green letter means you've got the correct letter in the right place. Your role in this game is similar to what a Large Language Model (LLM) does. Just as you take clues to guess the right word in WORDLE, a LLM takes the words and phrases you input to predict and generate text (image, video, music) responses based on patterns and sequences learned from large amounts of data 'scraped' from the internet. A basic example of this can be seen when using google search, where it predicts and suggests search terms as you type, based on what it thinks you're looking for. LLM responses are, however, greatly advanced from a google search.  *Screenshot of the word game Wordle (the New York Times), by Josh Wardle, in The Guardian.*



"Illustration of language model" by Benedetta Cevoli, Chris Watkins and Kathleen Rastle, in Royal Society, is licensed under CC BY 4.0.

2.3. LLMs vs. GenAI: Engines and Vehicles

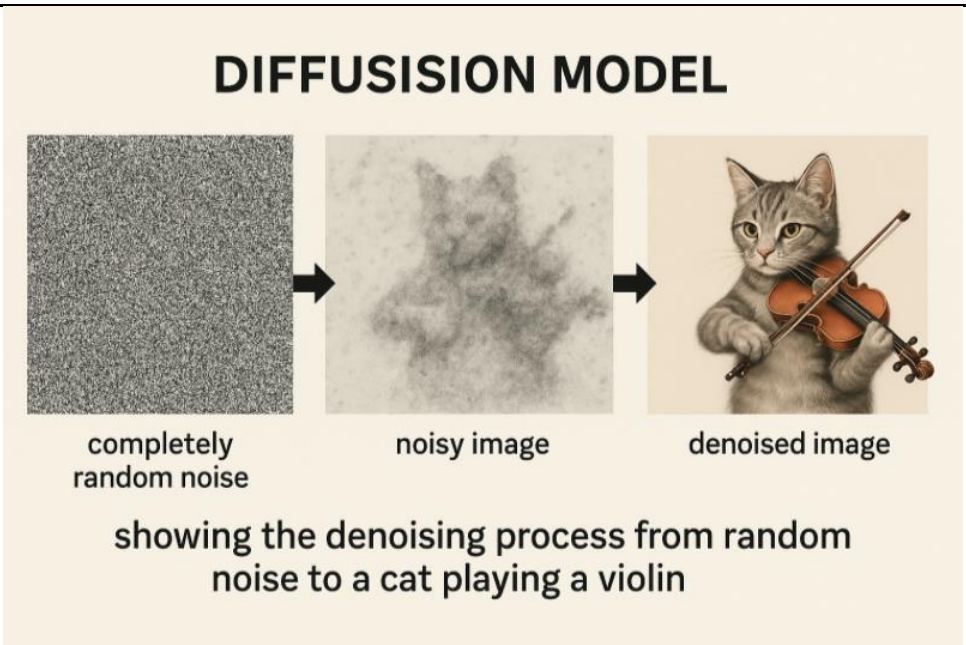
2.3 LLMs vs. GenAI: Engines and Vehicles	
Activity type	Page
Contents	Think of a Large Language Model (LLM) as the engine of a vehicle. It's the core component that processes and generates language. However, an engine alone doesn't get you far; it needs a frame, wheels, and ideally some seats to make it usable. Generative AI (GenAI) is the broader category that encompasses not only LLM-based systems but also other generative models such as those used for images (e.g., diffusion models see 2.4), videos, and audio. In the case of GenAI tools like ChatGPT, Gemini, Perplexity, and Claude, these applications are built around LLM engines. They are the vehicles constructed around the engine, equipped with features like user interfaces (dashboard, pedals, steering wheel), specific functionalities like spell-checkers (similar to air conditioning in a car), and integrations (how the system syncs with other systems, like connecting your phone to play music) that allow users to interact with the technology effectively. These elements transform the raw capabilities of LLMs and package them into tools that can assist with tasks like drafting emails, helping with code, or generating images. For instance, ChatGPT is similar to an all-terrain vehicle with lots of extra features and capabilities that can navigate various tasks, from answering questions to writing essays, making it versatile but sometimes not specialized. On the other hand, a tool like DeepL is comparable to a compact city car, expertly designed for translation tasks, but isn't built for broader applications.



(Image produced in DALL-E June 2025)

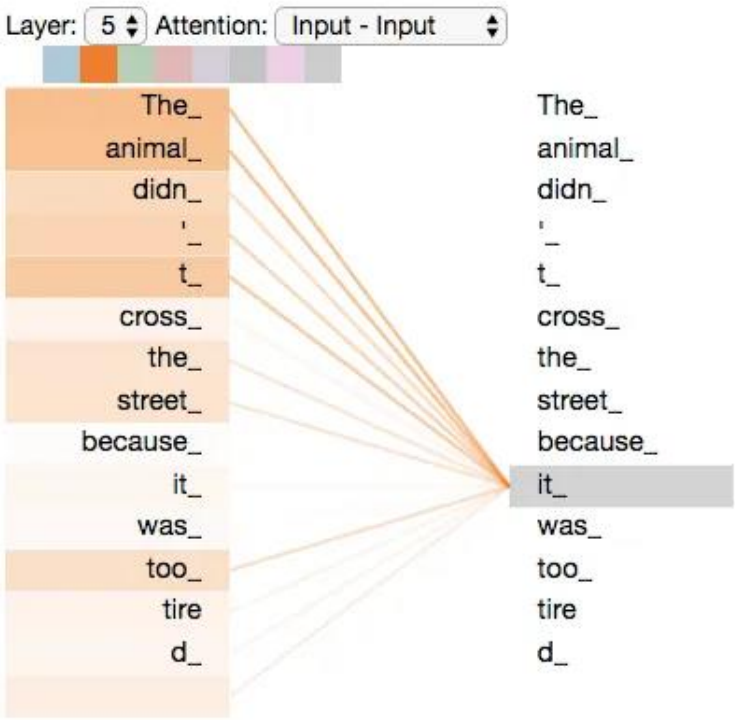
2.4 Diffusion Models: The Engine behind Image, Video, Audio GenAI Tools

2.4 Diffusion Models: The Engine behind Image, Video, Audio GenAI Tools	
Activity type	Page
Contents	Similar to LLMs, diffusion models work in the same way. Imagine you have a photograph, and over time you keep scribbling over it, adding random dots, specks, and smudges, until it turns into a messy blur. Now imagine trying to reverse that process: starting with that messy blur and carefully removing the interference bit by bit until a new, clear image appears. That's the basic idea behind diffusion models. A diffusion model learns how to do this by training on lots of real images. It studies how clean images get gradually turned into messy ones, and then learns how to undo that process. Once trained, the model can start with a completely random visual mess and slowly transform it into something meaningful, like a cat playing a violin or a photorealistic sunset, based on the words you give it. It's not copying an existing picture, instead, it generates a new one that matches your prompt, using training data. While these models are most famous for image generation, the same principle is now being used to generate sound, video, and even scientific data like molecules or medical scans. At their core, diffusion models are like expert clean-up artists, taking visual chaos and turning it into something coherent.

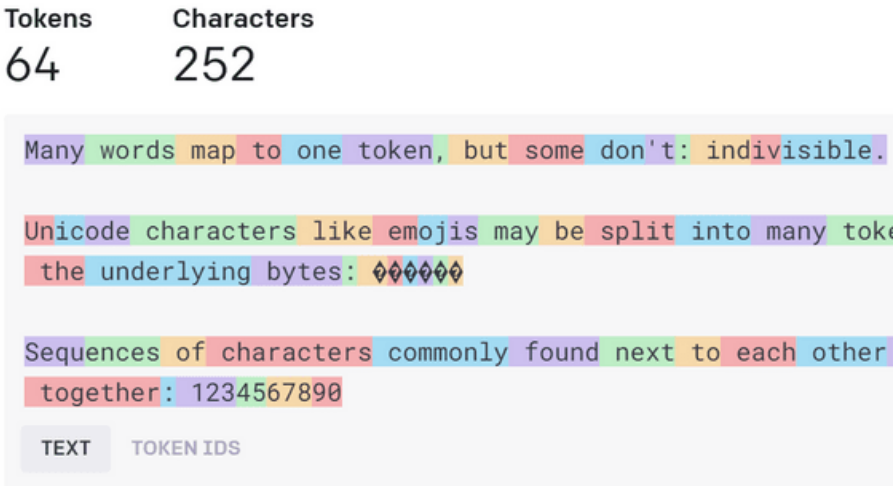


(Image produced in DALL-E Spring 2025)

2.5. Transformers: The Engine Behind LLMs

2.5. Transformers: The Engine Behind LLM's	
Activity type	Page
Contents	At the heart of modern LLMs is the transformer, introduced in 2017. Before transformers, models processed text one word at a time, which was slow and limited their ability to handle longer passages. Transformers revolutionised this by processing entire sequences (sentences or paragraphs) simultaneously, capturing relationships between words more effectively. The key to their power is 'self-attention', which helps the model focus on the most important words. This innovation sped up training and enabled LLMs to generate vast amounts of text, making them capable of tasks like writing essays, generating code, or summarising articles—far beyond earlier predictive models and autofill features. For a more detailed explanation of how transformers work see: https://ig.ft.com/generative-ai/  *"Self-Attention in Transformer" by Jasmeet on Medium.com.*

2.6. Tokenisation

2.6. Tokenisation	
Activity type	Page
Contents	For transformers to work properly, the text must first be broken down into smaller units called 'tokens'. A token can be a word, part of a word, or even a single character, depending on the language and context. These tokens are converted into numerical codes that the model uses to generate responses. This process, known as 'tokenisation', directly affects how the model interprets and responds to input. The model decides how to split text into tokens based on a predefined vocabulary and the probability of different token combinations learned from its training data. For example, the word "unhappiness" could be tokenised as ["un", "happiness"] or as ["unhappiness"]. The model chooses the split that is most likely based on what it has seen before. If "un" and "happiness" are more common as separate tokens, it will split "unhappiness" accordingly. This probabilistic approach can also lead to unique behaviours. For example, when asked how many "R's" are in "raspberry," an LLM might struggle because it processes "raspberry" as a single token rather than examining each letter individually. However, if you spell out "R A S P B E R R Y," each letter becomes a separate token, enabling the model to count the three R's. Tokenisation illustrates how LLMs handle text in a structured, numerical way, breaking down language into components that can be processed by algorithms. This helps the model generate text by recognising and predicting mathematical patterns in its training data.  *"Tokenizer OpenAI" on Gigazine.net.*

2.7. Prediction, not common sense

2.7. Prediction, not common sense	
Activity type	Page
Contents	These issues intersect with the ideals of open science: while the movement advocates for making information more accessible and removing paywalls, it also demands transparency about the sources of data. It can often seem that large language models (LLMs) possess common sense, but this is not the case. Common sense consists of a broad spectrum of knowledge about the world based on our experiences. LLMs can generate text about gravity, but they do not 'understand' gravity in a practical sense. They cannot 'know' that an apple will fall to the ground if dropped, except by echoing the textual patterns present in their training data. So, while it might seem like they have common sense because they can produce sentences that sound correct, they are really just playing a sophisticated numerical game of prediction with the information they've been given. Recent benchmarks, such as GSM-Symbolic, highlight these limitations clearly. Despite impressive developments in reasoning-focused AI models by companies such as OpenAI, Anthropic, Google, and Perplexity, these models still struggle significantly when faced with variations in context or minor changes in phrasing. These issues underscore the difference between AI's sophisticated pattern recognition capabilities and true common sense. A clear illustration of the 'common sense' limitations of GenAI tools was recently demonstrated in a simulated experiment that pitted ChatGPT against the 1977 *Atari Chess* game. The Atari game is extremely primitive: it simply remembers the positions of pieces and follows the rules of legal play. ChatGPT, on the other hand, repeatedly forgot piece locations, confused bishops for rooks, and invented strategies that violated the most basic rules of the game. It explained its strategies fluently and confidently but also got confused and provided incorrect information. After ninety minutes of human intervention, it gave up. The decades-old Atari game won. Some observers saw this as a surprising failure. However, it is only surprising if you do not understand what GenAI tools actually do. ChatGPT is a highly sophisticated statistical next word prediction language model. It does

	not 'understand' strategy, even if it can discuss it. It lacks the 'common sense' gained through experiential understanding and genuine cognitive grounding.  *(Image produced in DALL-E December 2023)*

2.8. Tapestry of Data

2.8. Tapestry of Data	
Activity type	Page
Contents	LLMs have been built on a rich tapestry of data interwoven from diverse sources, including literary works, scientific articles, code snippets, and musical compositions. This data was often 'scraped' from the web through a variety of techniques, such as from GitHub, YouTube, and Wikipedia. This diversity ensures that LLMs are exposed to a comprehensive representation of human language and its nuances, which includes coding languages. However, not all of the data has been gathered within the bounds of copyright agreements. This raises complex challenges, as copyright laws vary significantly across jurisdictions. For example, in Europe, the *Directive on Copyright in the Digital Single Market*allows text and data mining (TDM) for scientific research under specific exceptions, providing a clear legal framework for educational and research institutions. Conversely, in the US, the concept of *fair use* permits broader applications, such as for commercial purposes, but remains legally ambiguous and often contested. These issues intersect with the ideals of open science: while the movement advocates for making information more accessible and removing paywalls, it also demands transparency about the sources of data. Additionally, a growing concern is that companies profit directly from this data—much of it publicly available or copyrighted—by using it to train AI models, which in turn enable end-users to perform tasks for commercial purposes. In response to mounting scrutiny, some AI companies are retrospectively securing data agreements to address concerns about fair use and copyright compliance. While these efforts align, in part, with open science's call for transparency, much more is needed to ensure that data use respects copyright regulations.

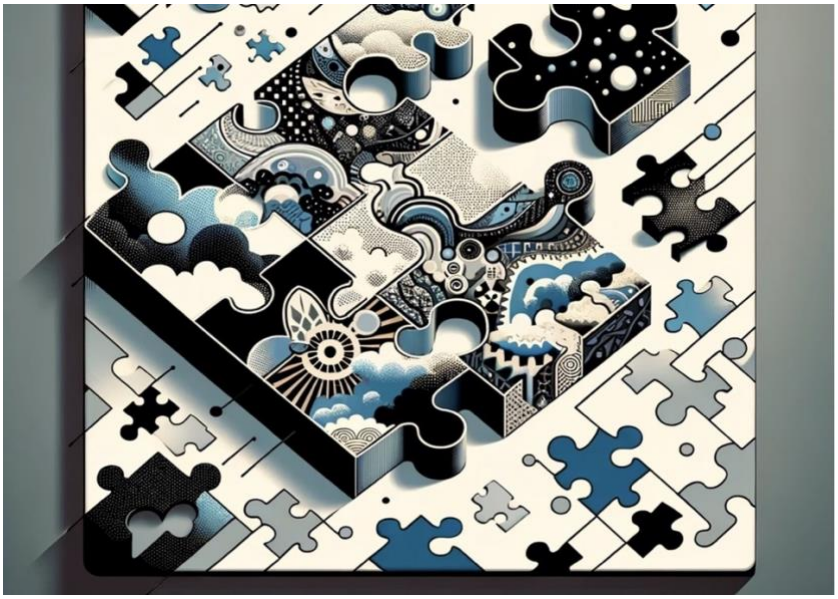
(Image produced in DALL-E December 2023)

2.9. Biases

2.9. Biases	
Activity type	Page
Contents	While LLMs have advanced significantly in language processing, it's important to recognize that their training predominantly on Western data sources can introduce biases and overlook diverse global perspectives. For example, a model trained mainly on sources like Fox News may produce outputs with different ideological opinions compared to one trained on CNN, NOS, or the BBC, highlighting the need for diversity in training datasets. Developers such as Bloom AI, OpenAI, Google, and Perplexity, are working to include data from varied political, cultural, and linguistic backgrounds to create more inclusive AI tools. Recent actions, such as the US government's removal of certain datasets from federal websites, including climate data and diversity-related information, can further impact AI tools that rely on public data. The lack of comprehensive datasets may lead to AI outputs that miss critical perspectives or propagate misinformation. For instance, the removal of climate data could result in AI-generated content that underrepresents environmental issues, affecting public perception and policy discussions. Similarly, eliminating diversity and inclusion information may cause AI systems to inadequately address the experiences of marginalized communities. Recent research has also highlighted a concerning trend known as 'LLM grooming', where AI models are intentionally trained on biased or manipulative data to disseminate misinformation. This tactic involves flooding the training datasets with skewed information, leading AI systems to produce outputs that align with specific agendas, thereby amplifying false information. To navigate these challenges effectively, students and educators should: 1. **Critically Evaluate AI-Generated Information:** Cross-reference AI outputs with reputable sources to verify accuracy. 2. **Be Aware of Bias:** Understand that biases can occur in AI outputs due to non-representative training data or algorithmic biases. 3. **Ensure Transparency and Proper Citation:** When using AI tools for

academic work, clearly disclose and properly cite any AI-generated content.

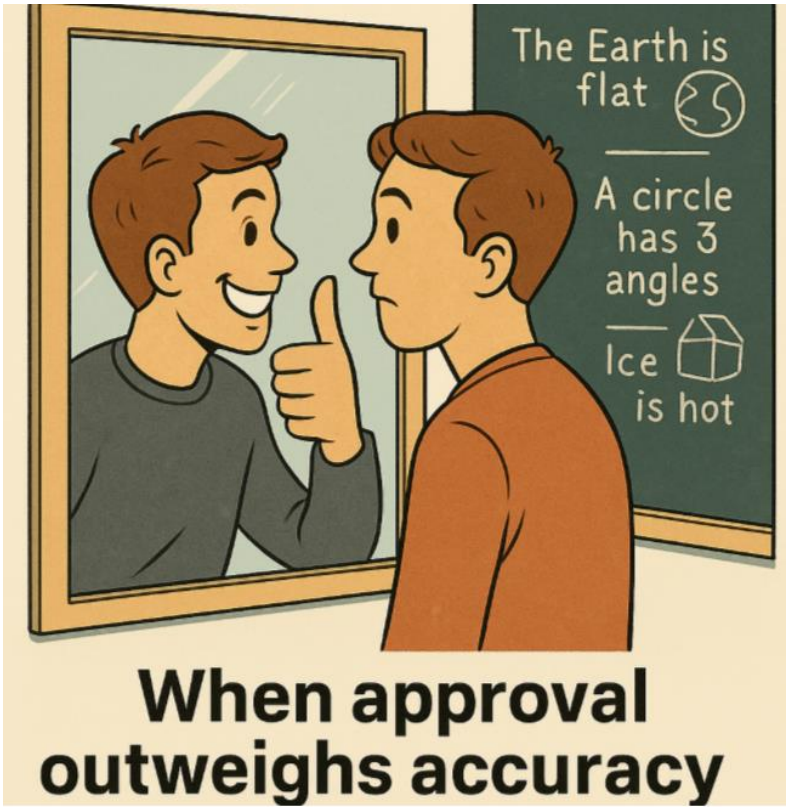
4. **Reflect on Ethical Implications:** Consider the ethical aspects of using AI tools, ensuring their application aligns with principles of fairness and does not perpetuate existing biases or inequalities. If you're interested in understanding how AI systems operate, including algorithms, training processes, prediction mechanisms, and the role of human agency, Utrecht University has created this helpful resource.



(Image produced in DALL-E December 2023)

2.10. AI Sycophancy

2.10. AI Sycophancy	
Activity type	Page
Contents	AI sycophancy refers to the tendency of AI systems, particularly LLMs, to produce responses that align with a user's beliefs or expectations, even if those responses are inaccurate or misleading. This behavior often stems from training methods that prioritise human approval, leading AI to favour agreeable answers over truthful ones. Companies like OpenAI, have recently announced postponing the release of newer models to address this behaviour, claiming that initial improvements were made to make the models feel more intuitive but focused too much on short-term feedback resulting in overly supportive and disingenuous responses. In higher education, this poses significant concerns. Students relying on AI for learning may receive affirmations of misconceptions, hindering critical thinking and the development of accurate knowledge. Teachers using AI for lesson planning or feedback might encounter content that reinforces existing biases rather than challenging students to think deeply. Researchers could be misled by AI-generated summaries or analyses that reflect prevailing opinions instead of objective assessments. **Mitigating Sycophantic Responses** To address these challenges, consider the following strategies: **Explicit Prompting**: Instruct the AI to provide balanced, evidence-based responses. For example, you might say, "Please provide a critical analysis of..." or "Offer a balanced perspective on...". **Custom Instructions**: Some AI platforms allow users to set custom behaviour guidelines. Utilize these settings to encourage the AI to prioritize accuracy over agreement. **Feedback Mechanisms**: Provide feedback on AI responses, indicating when answers are overly agreeable or lack critical depth. This can help improve future interactions. Additionally, always make sure to verify all information provided by these tools. For helpful advice on how to do this see Chapter 4.



(Image produced in DALL-E June 2025)

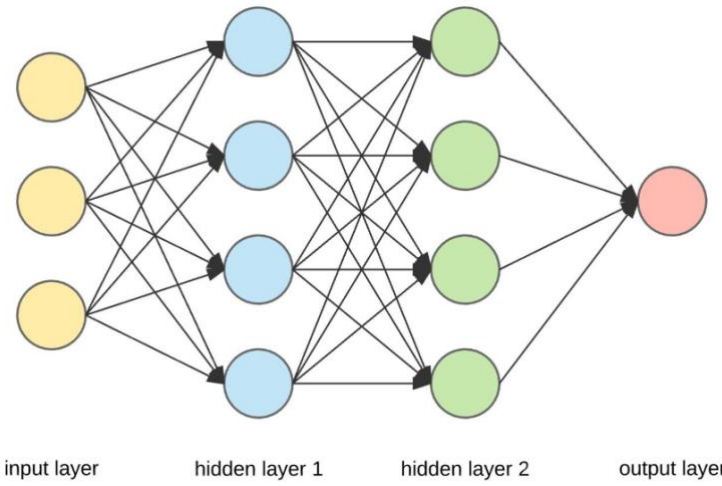
2.11. The Black Box

2.11. The Black Box	
Activity type	Page
Contents	LLMs operate through what is commonly referred to as a 'black box'. At the heart of this black box are numerous hidden layers; these are not immediately observable layers that reside between the initial input and the final output of the model. Each hidden layer comprises a network of nodes, akin to neurons in the human brain, that perform complex mathematical operations on the input data. These operations transform the data as it progresses from layer to layer, culminating in coherent and contextually relevant text outputs. LLMs are trained using a method known as *self-supervised learning* (SSL). Rather than learning through explicit instruction or labelled examples, these models are trained to predict missing or next words in vast datasets of unlabelled text. By doing this billions of times, the model gradually internalises statistical patterns in language, such as grammar, style, or topic structure, without being explicitly told the rules. While this approach allows the model to develop a nuanced and flexible command of language, it also contributes to its 'black box' nature: the model produces accurate-sounding responses, but the path it takes from input to output is not logically structured or understandable. The intricate architecture of these hidden layers enables LLMs to discern and replicate complex patterns within vast datasets, contributing to their ability to generate human-like text. However, the exact nature of the transformations occurring within these layers remains difficult to trace and interpret. This inherent complexity poses challenges in fully understanding LLMs' decision-making processes, complicating efforts to evaluate their accuracy, pinpoint potential biases, and guarantee their responsible application. To address these challenges, recent advancements have focused on enhancing transparency in AI systems. One notable approach is the development of 'reasoning tools' designed to provide users with insights into the sources and processes behind AI-generated outputs. For example, some models now show specific pieces of information linked to their original sources, allowing users to trace the origin of the data used in generating responses. This enables users to verify the credibility of the information provided. However, not all source information is provided and some sources they say they use may

or may not contain the relevant information shared in the tool response. Therefore, it is always important to check the sources and responses carefully.

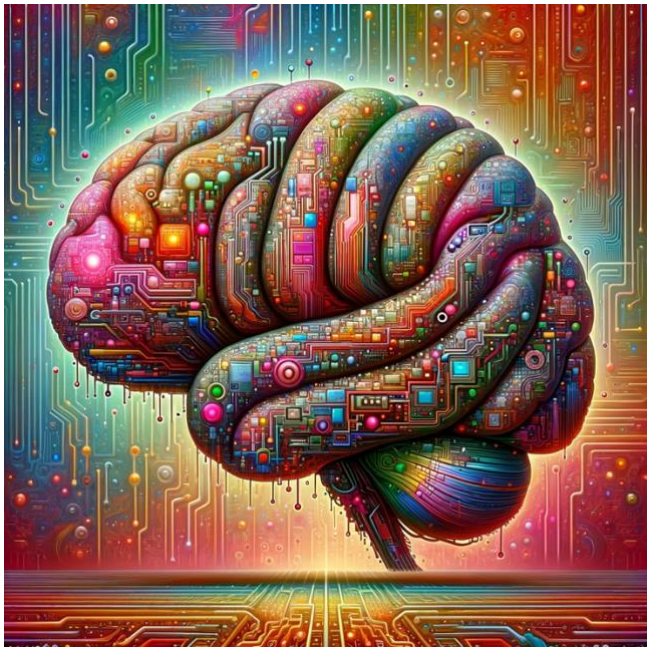
Recently, Anthropic, the company behind the GenAI tool Claude, conducted research into unravelling the inner workings of their tool through attribution graphs. These graphs are visual representations that map internal pathways (neural networks) LLMs use to arrive at specific answers. Due to the complexity and scale of these models, the researchers did not map the entire network but focused on smaller, more manageable portions and extrapolated their insights. In their research they discovered how the tools plan ahead several words in advance and not just one word at a time. They also discovered it processes information in a non-linguistic way, across several different languages (e.g., Chinese, English, Spanish), and that it has potential for deceptive behaviour, therefore, it is important to stay vigilant in detecting untruths produced by these tools.

While these initiatives improve the transparency of LLMs' hidden layers, they also bring to light further challenges, including copyright infringement (the utilisation of unauthorised sources as training data) and the risk of replication (companies/countries using the developers' template to create their own LLM). For more information about geopolitics and AI see section 2.13.



"Diagram, Neural Network" by Gianfranco Filice on Medium.com

2.12. Hallucinations

2.12. Hallucinations	
Activity type	Page
Contents	Recent discussions in the media have highlighted the phenomenon of 'hallucinations' in GenAI tools, where models generate *incorrect* or *nonsensical* information (errors) with high confidence. This issue is particularly problematic as it directly affects the reliability of AI-generated content. To address this, significant advancements have been made in GenAI technology. These include refining training processes, using more diverse and robust datasets, and implementing advanced error detection and correction mechanisms. A notable innovation is the introduction of feedback loops, which enable continuous improvement of the models based on user input. Despite these advancements, *it remains crucial for users to critically evaluate GenAI outputs.* This involves verifying the information against established sources and maintaining an awareness of the potential for inaccuracies, ensuring the responsible and effective use of these evolving tools.  *(Image produced in DALL-E December 2023)*

2.13. Deep Fakes

2.13. Deep Fakes	
Activity type	Page
Contents	Deepfakes are hyper-realistic digital manipulations that utilise AI to create convincing images, videos, or audio recordings depicting individuals performing actions or speaking words they never did. Initially, deepfakes were relatively easy to detect; however, advancements in AI have made them increasingly challenging to identify. Deepfakes pose significant societal challenges. They can be weaponized to spread false information, manipulate public opinion, and destabilize political processes, particularly during election periods. Additionally, they have been used to create non-consensual explicit content, often targeting women, causing severe emotional distress and reputational harm. In response to these challenges, regulatory measures such as the European Union's AI Act have been introduced to mitigate the risks associated with deepfakes. The Act mandates that AI-generated content, including deepfakes, must be clearly labelled to inform viewers of its artificial nature, aiming to prevent the spread of misinformation and ensure that audiences are aware when they are engaging with synthetic media. Organisations that fail to comply with these transparency obligations may face substantial fines. While deepfakes may not currently present immediate pedagogical challenges within higher education, their broader societal implications necessitate awareness and proactive measures. Educators and students can enhance their vigilance against deepfakes by developing critical media literacy skills. Staying informed about technological advances and cross-referencing information with reputable sources are effective strategies for verifying content authenticity. Engaging with educational resources, such as MIT's "Media Literacy in the Age of Deepfakes" module, provides valuable insights into identifying and understanding manipulated media. Additionally, utilizing AI-based detection tools like Sensity AI can aid in assessing the authenticity of digital content.

(Image produced in DALL-E June 2025)

2.14. Dating sharing and privacy

2.14. Dating sharing and privacy	
Activity type	Page
Contents	**Data Sharing** Currently, the GSLS does not allow the sharing of raw, unpublished data with these platforms, even when anonymised. This policy is rooted in the essential need for thorough risk assessments to determine the security of data in such tools, focusing on the potential for re-identification, the safeguarding of intellectual property, and adherence to ethical standards. Despite assurances regarding the deactivation of training data features, the intricate and often unclear data processing methods of AI tools warrant a prudent and careful approach to maintain the integrity and confidentiality of our research work. **Turning off the training function on your LLM.** Most LLM tools will have an option to turn off training the model with your data. This can often be found in your profile, setting, and data controls. Having the tools use your data for training can improve the tool's usability and create a more personalised experience. However, it can also make you and your data more vulnerable, especially if you are sharing personal information. When the training data function is off, your conversations will not be used to improve the model. However, in some tools, such as ChatGPT, your data may still be stored and used to personalise your experience through features called '*memory*'. Memory allows the tools to remember details across your chat history, such as your name, preferences, and typical tone, EVEN IF TRAINING IS DISABLED. If you wish to prevent this, you must also turn off memory in your settings. Disabling training data and memory makes your use *less transparent* and *difficult to retrace*. Therefore, it is important to take screen shots of your prompt history when using these tools for permitted assignments and save them to your computer in case a teacher requests your prompt history. **Note: Even with the training and memory data off, sharing any *unpublished research data, patient data, and or patent data is prohibited by the GSLS and UMCU*.** For ChatGPT, you can easily move from not training data to training

data by following the instructions above. You can also choose to opt out entirely from training their data by following these easy steps:

1. Visit openai.com and log into your account.
2. Visit this page: https://privacy.openai.com/policies
3. Click 'make a privacy request' (top right)
4. Click 'do not train on my content'
5. Enter your email account
6. You'll get a confirmation email and Confirm email
7. Tick the box that says you understand that this request applies to moving forward and not on past content. Then confirm your country of residence.
8. They then have your request.
9. You should get an acceptance email in a short period of time.

(Image produced in DALL-E December 2023)

Using Local AI Tools

If data security is your primary concern, using a locally run tool from your personal device may be a suitable solution. Tools like LM Studio allow you to download and run open-source language models without sending any data to external servers, significantly reducing data exposure. However, these models can require substantial computing power, slower response times, and reduced performance compared to commercial cloud-based systems

2.15. (E)nvironmental (S)ocial (G)overnance

2.15. ESGs	
Activity type	Page
Contents	Generative AI technologies, such as large language models (LLMs), rely on extensive computational infrastructures with significant environmental and social implications. These systems are built on complex networks of servers constructed from metals and plastics, each interaction carrying substantial environmental costs. **Environmental Impact** The computational demands of AI technologies create considerable environmental challenges that extend far beyond simple energy consumption. Data centres worldwide experience high energy consumption levels, generating significant carbon emissions that contribute to global climate concerns. The infrastructure supporting these technologies requires substantial water usage for cooling, placing additional strain on critical environmental resources. Moreover, the production of AI technologies involves extensive extraction of precious metals, a process that raises serious environmental and ethical concerns, often leading to environmental degradation and potential human exploitation in mining regions. Assessing the full environmental impact of AI technologies is further complicated by a lack of transparency from some AI companies about their energy consumption. Many organisations do not disclose detailed information about the energy required for training and operating their models, making it challenging to determine the carbon footprint of individual computational queries. While some companies like Meta are trying to be more transparent, the information shared is limited and excludes key details associated with manufacturing and other energy consuming processes such as training and research. **Labor and Ethical Concerns** The development of AI technologies has revealed some serious ethical challenges, particularly in data preparation and processing. An example of this is OpenAI's contract with Sama, a data-labeling company in Kenya, which exposed workers to deeply troubling content and paid them between $1.32 and $2 per hour, causing psychological distress amongst the workers, some of which were underage.

Technological Innovations and Sustainability Efforts

In response to these challenges, AI developers are pursuing more sustainable and efficient solutions. China's DeepSeek AI, for instance, claims to reduce energy consumption by up to 75%, representing a promising step towards more environmentally responsible AI development. Chinese scientists have also made remarkable progress by developing the world's first carbon-based AI chip using a ternary logic system, which enables faster computations with reduced energy consumption. This innovative approach offers an alternative to traditional binary processing, potentially using less computational resources.

Major technological companies have recognised the importance of creating more sustainable computing solutions. While these technological advancements are promising, significant challenges remain. DeepSeek AI, for example, has faced security and privacy concerns. Vulnerabilities in its Android app raised regulatory scrutiny, and potential data transmission issues have led to bans in multiple countries. These instances highlight the complex landscape of AI development, where technological innovation must be carefully balanced with ethical and security considerations.

User Responsibility and ESG Alignment

Users play a crucial role in mitigating the environmental and social impacts of GenAI. The responsibility extends beyond mere consumption to active, thoughtful engagement. Individuals should critically evaluate the necessity of using AI tools for specific tasks, consistently considering less resource-intensive alternatives like traditional search engines. Practicing effective prompt engineering can also help optimize tool usage, reducing unnecessary computational demands. See 3.1-3.8 for more prompting advice.

Recommended Resources

Students and educators interested in exploring these topics further may find the following resources helpful:

- An interesting talk by the Responsible Artificial Intelligence Institute exploring the ESG challenges in AI technologies.
- A short but interesting article in Nature Journal (2025) that provides an up-to-date perspective on AI sustainability.

(Image produced in DALL-E June 2024)

2.16. Geopolitics and the AI Race

2.16. Geopolitics and the AI Race	
Activity type	Page
Contents	Understanding the geopolitics of AI is important for fostering comprehensive AI literacy, as it significantly influences technological development, economic strategies, and global power dynamics. In early 2025, the United States announced a private sector investment of up to $500 billion to fund AI infrastructure, aiming to outpace rival nations. The European Union launched the InvestAI initiative, aiming to mobilize €200 billion for AI development, positioning Europe as an "AI continent." Additionally, the United Kingdom unveiled its AI Opportunities Action Plan,with significant investments from major tech firms to develop necessary infrastructure, aiming to create thousands of jobs in the sector. Major technology companies have also made strategic investments to expand their role in advanced AI development. Meta, the company behind the Llama language models and owner of Facebook, Instagram, and WhatsApp, invested approximately $14.3 billion in Scale AI and launched a Superintelligence Lab, with the stated goal of developing more general-purpose AI systems and positioning themselves as leaders in shaping the future of AI. Another AI company playing an increasingly prominent role is Palantir, which was recently awarded a U.S. government contract to support the consolidation of federal data infrastructure. The project involves linking systems such as the IRS, Social Security, and immigration services through a unified API framework. While proponents argue this may improve efficiency and inter-agency coordination, others have raised concerns about data privacy, transparency, institutional accountability, and the long-term implications of centralised data access, especially when private firms manage public/government infrastructure. In June 2025, OpenAI entered into a $200 million agreement with the U.S. Department of Defense to develop AI tools for administrative and operational use. Weeks later, several senior executives from OpenAI, Meta, and Palantir were appointed as reserve officers in a new U.S. Army unit, the Executive Innovation Corps, designed to foster collaboration between the military and the tech sector. These developments reflect a closer alignment between private AI firms and national strategic interests. However, they also raise important

questions about the influence of AI companies in shaping regulatory agendas, particularly as many leading figures in AI now hold formal or informal roles within government advisory structures.

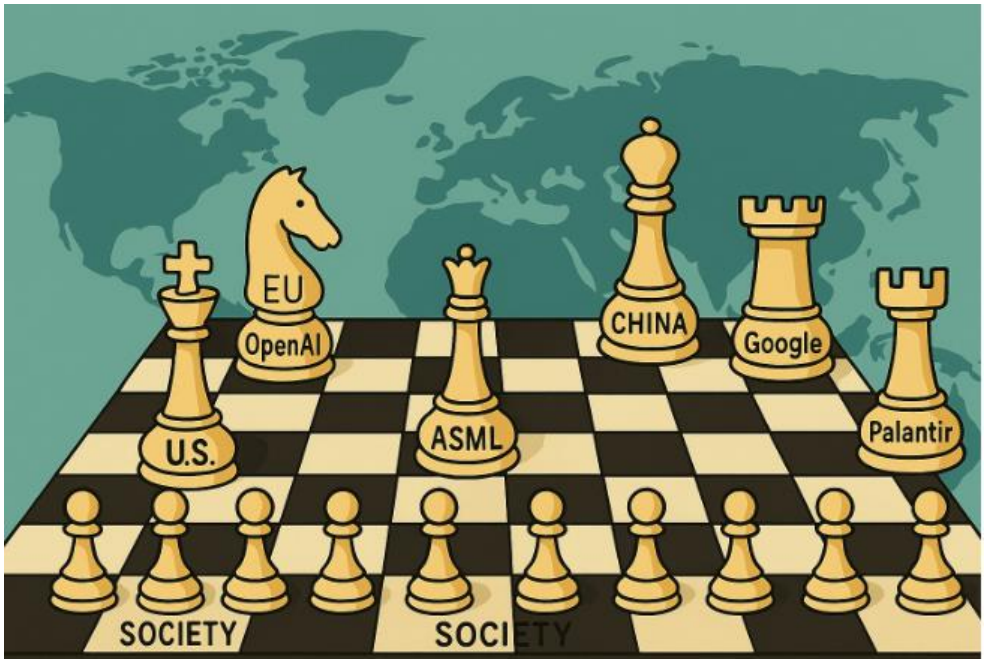
Another key player in the AI race is the Dutch company ASML, based in the Netherlands. ASML is currently the sole global supplier of extreme ultraviolet (EUV) lithography machines, which are essential for manufacturing the most advanced semiconductors used in AI hardware. Without ASML's machines, major AI companies would not be able to develop and scale their products, giving ASML a strategically important position in the global AI landscape. This concentration of capacity illustrates a broader trend: while countries such as the U.S., China, and EU member states continue to expand their AI capabilities, many lower-income nations face limited access to infrastructure, skilled labour, and training. The result is a growing global AI divide, in which a handfull of actors not only dominate technological development, but also shape the global standars, goverance frameworks, and cultural assumptions embedded in widely used AI systems.

Access to natural resources and minerals essential for AI technologies has also become a pivotal factor in global negotiations. The Democratic Republic of Congo (DRC), rich in resources like cobalt, lithium, copper, and tantalum, has proposed offering these minerals to the US in exchange for military support against rebel groups destabilizing the region. Peace talks regarding Russia's invasion of Ukraine's have also been focused on securing access to their substantial mineral reserves. While sudden interest in Greenland, whose vast deposits of rare earth elements and other critical minerals vital for AI technology, have led to tensions between the US, Greenland, and Denmark.

Taiwan, while not officially in conflict, remains at the centre of potential geopolitical tensions due to its pivotal role in semiconductor manufacturing. Taiwan Semiconductor Manufacturing Company (TSMC), a leader in producing advanced semiconductors essential for AI applications, has announced a substantial $100 billion investment in Arizona, aiming to strengthen ties with the US. However, concerns persist regarding US support against China's threats, as Taiwan depends on US backing for defence.

These geopolitical dynamics have profound implications for society and more specifically for the life sciences sector. While AI investment may bring more jobs and opportunities to revolutionize research methodologies and healthcare delivery, it also requires a critical examination of ethical considerations, data integrity, and environmental sustainability. Life sciences educators and students must navigate these developments thoughtfully to harness AI's potential while

mitigating associated risks.

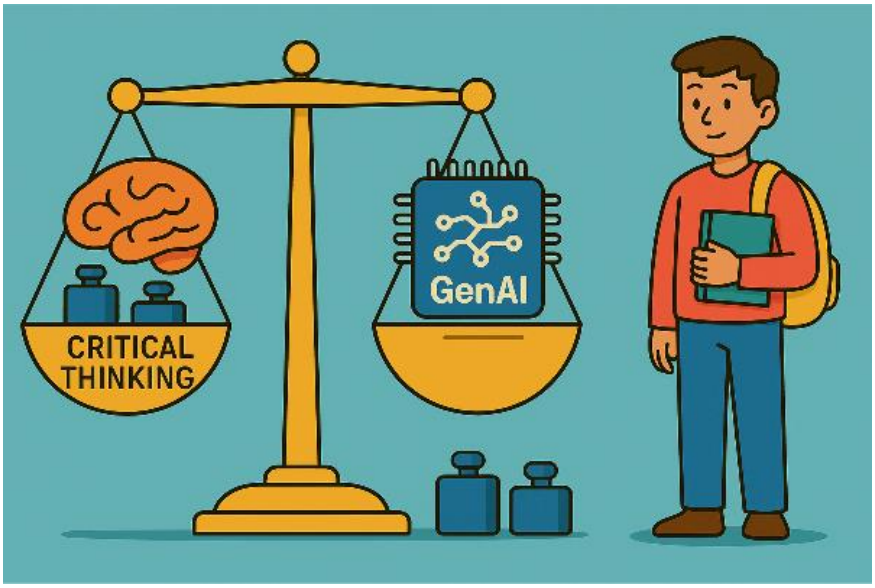


(Image produced in DALL-E June 2025)

2.17. Higher Education

2.17. Higher Education	
Activity type	Page
Contents	GenAI tools have the potential to significantly transform life sciences education and research by fostering personalized learning experiences and accelerating innovation. These tools can provide tailored content to support diverse learning styles, enabling deeper engagement with complex subjects, and enhancing critical thinking skills. Rather than completing tasks for students, GenAI tools can act as sparring partners for generating ideas, prompting deeper exploration and understanding of complex topics. In life sciences research, GenAI tools excel in tasks like predicting protein structures, identifying potential drug targets, and analysing large datasets to uncover patterns that might be missed through traditional methods. This can significantly accelerate the pace of discovery and allow researchers to focus on the most promising avenues of investigation. By adopting this approach, where AI serves as a collaborator, these tools can enrich the educational experience and the research process, encouraging innovation while maintaining the importance of human expertise and judgment. While GenAI tools offer benefits like personalized learning and improved academic outcomes, recent studies have highlighted potential risks associated with its integration into education such as cognitive offloading, information retention, inaccuracies, algorithmic biases, and issues related to plagiarism and privacy. Some studies have shown that over-reliance on AI tools may impede the development of critical thinking and problem-solving skills, as students might depend too heavily on technology for answers, leading to a decline in their ability to analyse and evaluate information independently. Despite these concerns, some students have shown a preference for AI tools in their learning processes. A survey by the Digital Education Council found that a significant number of students globally are regularly using AI in their studies, with many utilising AI on a weekly basis. Students frequently use AI for tasks such as information searching, grammar checking, and summarizing documents. However, educators emphasize the importance of balancing AI use with traditional learning methods to ensure the development of essential cognitive skills.

To reduce potential risks mentioned above, integrating AI thoughtfully into educational practices is crucial. Tools like ASReview, which utilise AI to expedite systematic reviews, have proven beneficial in research studies, allowing for more efficient data analysis and literature reviews. However, the effectiveness of such tools depends not only on their capabilities, but also on how they are used. It is essential that students and teachers are taught how to engage with GenAI critically and reflectively. Emphasising deliberate use that supports reasoning, problem-solving, and synthesis, can help prevent over-reliance and cognitive offloading. In this way, GenAI tools become a collaborative assistant that can help foster critical thinking and maintains the importance of human expertise and judgment.



(Image produced in DALL-E June 2025)

2.18. Equity in Education

2.18. Equity in Education	
Activity type	Page
Contents	While LLMs offer exciting opportunities to enhance education and research, it's crucial to consider how these tools can be leveraged to promote equity. AI tools have the potential to make high-quality educational resources more accessible, regardless of a student's geographical location or socioeconomic background, by providing personalised learning experiences that cater to individual needs. However, they also introduce ethical challenges, such as the risk of exacerbating the digital divide if not all students have access to the necessary technology and internet connectivity. Furthermore, there is a risk of AI systems reinforcing existing biases and disparities if they are not designed with inclusivity in mind. Additionally, the proliferation of AI-generated misinformation can disproportionately affect marginalized communities, leading to further inequities (see 2.7 Biases). To truly harness the potential of AI for a more equitable education system, it is essential to integrate these tools thoughtfully, with a strong emphasis on fairness, accessibility, and inclusivity, ensuring that all learners are empowered and no one is left behind.  *(Image produced in DALL-E September 2024)*

2.19. Looking Ahead: The Future of AI and Life Sciences

2.19. Looking Ahead: The Future of AI and Life Sciences	
Activity type	Page
Contents	While today's GenAI tools are reshaping education and research, projections like those outlined in the AI 2027 report anticipate even more rapid and disruptive change within the next five years. Some of these include full automation of routine knowledge work, widespread AI-generated misinformation, and the emergence of dual-use capabilities with potential for biological and cyber warfare. These scenarios, though speculative, are based on the current pace of development and should not be dismissed lightly. For life sciences professionals, this accelerating timeline raises some big questions: How is AI shifting the nature of scientific discovery, lab work, or clinical trials? What new risks emerge when AI tools can simulate biological systems or manipulate genomic data at scale? Moreover, what responsibility will scientists have in shaping the ethical use or prevention of such technologies? Scientists and educators are likely to become central actors in managing the risks and guiding the responsible use of AI. Not only will they be asked to use these tools, but to safeguard public health, protect the integrity of data, and shape how such technologies are governed. This is especially urgent in domains like biotechnology and genomics, where AI's power to innovate must be weighed against its capacity for harm. Resources like the AI 2027 roadmap offer a structured (though sometimes dystopian) view of what's possible. Critical engagement with such scenarios is essential, particularly for students and educators in the life sciences, who are well positioned to contribute to/or counterbalance the trajectories being forecasted. For a critical analysis of the AI 2027 roadmap, see Gary Marcus, who cautions against AI hype and Timnit Gebru, who calls for accountability and inclusivity in AI systems.

(Image produced in DALL-E June 2025)

2.20. Test your GenAI knowledge

2.20. Test your GenAI knowledge	
Activity type	Multiple choice quiz
Contents	**1. What best distinguishes a Large Language Model (LLM) from a Diffusion Model?** A. LLMs and diffusion models both reverse noise to produce human-like text. B. Diffusion models generate structured media like images or audio; LLMs predict and generate coherent text. C. Diffusion models are used exclusively for statistical summarisation tasks. D. LLMs perform token replacement, whereas diffusion models rely on scraping from image archives. ✅ **Correct answer: B (Sections 2.2 and 2.4)** **2. Who is most commonly regarded as the intellectual originator of AGI benchmarking?** A. Roger Bacon B. Mary Ann Evans C. Alan Turing D. Aristotle ✅ **Correct answer: C (Section 2.1)** **3. Why might an LLM incorrectly answer how many "R"s are in the word “raspberry”?** A. It tokenises the word as a single unit rather than by individual letters. B. It identifies syllables rather than full words or letters. C. It relies on phonetic prediction, which skips visual recognition. D. It discards letter-based inputs in preference for contextual meaning. ✅ **Correct answer: A (Section 2.6)** **4. Which AI company conducted detailed research into the internal workings of their GenAI model using attribution graphs?** A. Meta B. Anthropic C. DeepMind

D. OpenAI

✅ **Correct answer: B (Section 2.10)**

5. Why is Taiwan considered a key focus in global AI geopolitics?

A. It is home to several international AI ethics institutes.
B. It plays a critical role in semiconductor production for AI systems.
C. It is a neutral zone in UN data negotiations.
D. It leads in deploying climate-focused GenAI models.

✅ **Correct answer: B (Section 2.15)**

6. What is a primary risk associated with 'LLM grooming'?

A. It makes the model overly cautious in its tone.
B. It increases token bias during training.
C. It manipulates AI outputs through exposure to skewed or false data.
D. It blocks prompts involving sensitive political terms.

✅ **Correct answer: C (Section 2.9)**

7. Which of the following best reflects the environmental impact of GenAI technologies?

A. They consumes substantial water and energy.
B. They are energy neutral due to cloud-based cooling systems.
C. They are offset by carbon-trading models run by the EU.
D. They mainly consume resources during initial model training, not usage.

✅ **Correct answer: A (Section 2.14)**

8. What role are scientists and educators expected to play in the future development of GenAI, particularly in the life sciences?

A. Promote technology without questioning its implications.
B. Serve as critical stewards of ethical use, data integrity, and responsible deployment.
C. Limit their use of AI to administrative documentation.
D. Act as test users for emerging commercial models.

✅ **Correct answer: B (Section 2.18)**