OVIC
Office of the Victorian Information Commissioner

ADM+S
ARC Centre of Excellence for Automated Decision-Making and Society

GenAI Concepts

Technical, operational and regulatory terms and concepts for generative artificial intelligence (GenAI)

Developed as a collaboration between the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S) and the Office of the Victorian Information Commissioner (OVIC)

Authors
Fan Yang
Jake Goldenfein
Kathy Nickels

Webpage
admscentre.org.au/genai-concepts

Acknowledgement of Country

In the spirit of reconciliation, we acknowledge the Traditional Custodians of country throughout Australia and their connections to land, sea and community. We pay our respect to their elders past and present and extend that respect to all Aboriginal and Torres Strait Islander peoples today.

Suggested citation (document)

Yang, F., Goldenfein, J., & Nickels, K. (2024). GenAI Concepts. Melbourne: ARC Centre of Excellence for Automated Decision-Making and Society RMIT University, and OVIC. DOI: 10.60836/psmc-rv23

Suggested citation (webpage)

Fan Yang, Jake Goldenfein, and Kathy Nickels, 'GenAI concepts', ADM+S and OVIC (Web Page, 2024), https://www.admscentre.org.au/genai-concepts

ARC Acknowledgement

This research was conducted by the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005). The Centre is funded by the Australian Government through the Australian Research Council.

Australian Government
Australian Research Council

Contents

Introduction

Generative AI products and services, such as OpenAI's ChatGPT, Alphabet's Gemini, and Microsoft's Copilot, have sparked substantial interest in the private and public sectors. Organisations are already experimenting with integrated AI services provided by big tech firms, as well as custom procurements from smaller software companies. In the absence of comprehensive AI regulation in Australia, deploying these systems in a responsible, ethical and legally compliant way demands a deep understanding of how they function and the legal and ethical challenges they raise.

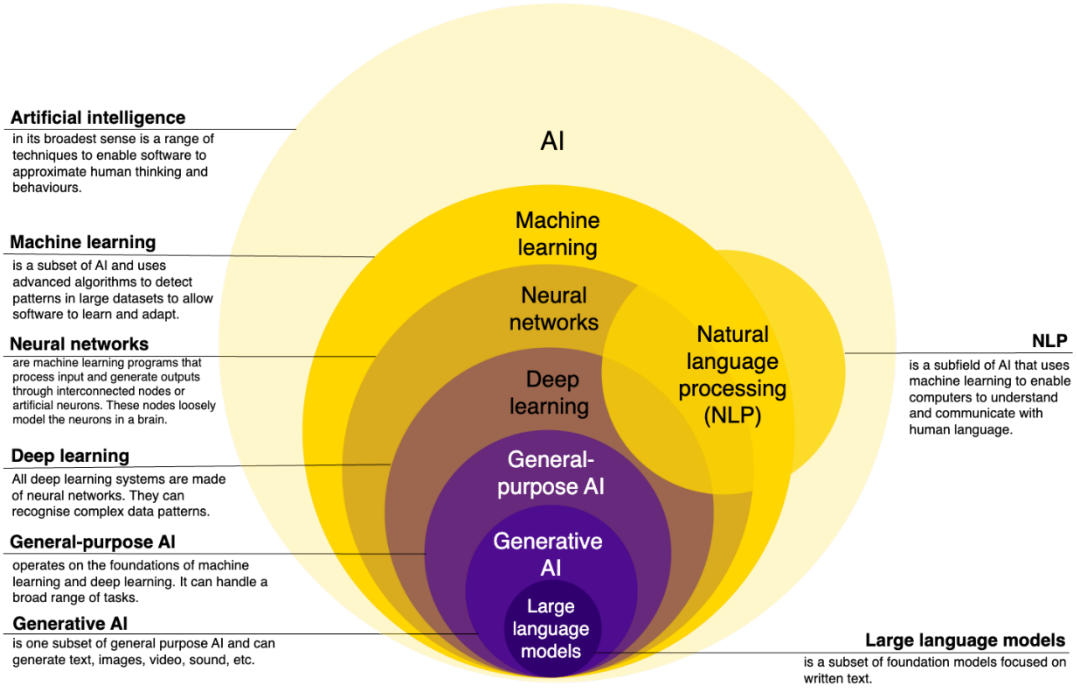
But generative AI (GenAI) is not simple. AI software systems are diverse and rely on complex supply chains and data flows. Coming to grips with the technical, operational and regulatory vocabularies that have emerged around GenAI is a considerable challenge. To help entities interested in GenAI deployments, this publication outlines 42 concepts fundamental to AI software systems. Each concept is illustrated through descriptions, examples and real-world use cases, with accessible language and visual elements to accommodate a diverse range of stakeholders and readerships.

Developed as a collaboration between the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S) and the Office of the Victorian Information Commissioner (OVIC), GenAI Concepts was produced through expert consultation and analysis of academic literature, industry reports and policy guidance.

GenAI Concepts can also be accessed as a regularly updated website at **admscentre.org.au/genai-concepts**

Technical

The following diagram demonstrates the interconnections among fundamental technical terms in the field of artificial intelligence (AI).



Using this diagram as a starting point, this section begins with basic GenAI concepts relevant to daily applications, including prompt and prompt engineering, large language models (LLMs), knowledge cut-off dates and chatbots.

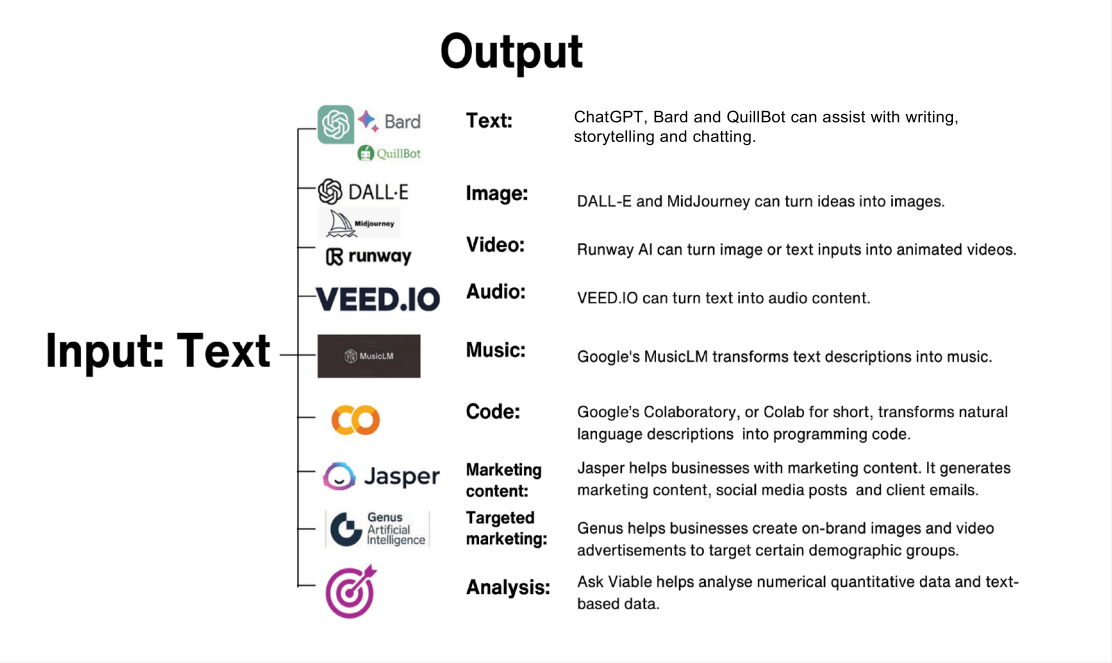
It then moves on to more advanced terms that are often encountered in official documents, such as general-purpose AI, foundation models and frontier AI.

The section further explores specialised technical terms, encompassing open-source and closed-source LLMs, transformer architecture, transfer learning, tokens, reinforcement learning from human feedback (RLHF), diffusion models and inference techniques.

Generative AI or GenAI

Generative AI or GenAI are both short for generative artificial intelligence. These are software systems that create content as text, images, music, audio and videos based on a user's 'prompts'.

Types of GenAI models in the market



Prompt and prompt engineering

A prompt is an instruction, query or command that a user enters into a GenAI interface to request a response from the system.

Because GenAI systems produce text through **statistical predictions** of the most likely next words in a sentence, the responses that GenAI systems produce may not always be the same. This is why GenAI outputs are sometimes described as non-deterministic.

Prompt engineering is the practice of writing inputs for GenAI tools so that the tool can produce optimal outputs.

> **Example**
>
> Basic structure of a prompt:
>
> Acting as a [**role**] perform [**task**] in [**format**] in the [**style**]

Acting as a [role]	Perform a [task]	Show as [format]	In the [style]
Each role provides context and background.	*The task should be clear and specific.*	*How you would like to structure the information generated.*	*This is optional. It shows what tone you'd like to use for the information.*
Marketer Advertiser Copywriter Accountant Lawyer Financial analyst English professor Journalist Project manager Manager Engineer Recruiter …	Headline Presentation Webinar Blog post Book outline Email sequence Social media campaign Product description Cover letter Summary TikTok, YouTube or Instagram Reel video script Sales page / ad copy …	Table List With bullet points Summary HTML Code Spreadsheet CSV file Plain text file Rich text PDF Markdown Word cloud …	Formal Poetic Enthusiastic Shakespearian Accessible Basic English Scientific Objective Neutral Pop culture

As a [life influencer (**the role**)], create a [blog post (**the task**)] about the benefits of daily exercise in a [PDF (**the format**)] in an [accessible and enthusiastic tone (**the style**)].

> **Tip**
>
> If the task is too complicated, break it down into steps to make it easier for the GenAI model to read and interpret.

Machine learning

Machine learning (sometimes seen as ML) is a set of techniques for creating algorithms so that computational systems can learn from data.

A machine learning algorithm is a set of rules or processes that helps an AI system to perform specific tasks, such as finding patterns in data or making predictions based on inputs. In this way, the model's behaviour reflects the data or the learning experience.

Higher quality data helps algorithms improve their accuracy in various tasks, such as recognising faces in photos, predicting the weather or recommending products to buy.

Types of machine learning

- Supervised machine learning:
  An algorithm is provided with labelled data, for example, a collection of pictures of animals with labels for each animal in each picture. The algorithm learns from these examples and tries to predict the correct labels for new, unseen data.
- Unsupervised machine learning:
  An algorithm is provided with unlabelled data and then it tries to find patterns or structures in the data on its own.
- Hybrid machine learning:
  Combines elements of supervised and unsupervised learning approaches and sometimes different types of algorithms to leverage the strengths of different methods based on the developer's needs.

Applications

- **Healthcare**: ProMed (the Program for Monitoring Emerging Diseases) offers an online real-time data analysis and reporting system showing outbreaks of infectious diseases worldwide.
- **Finance**: Fraud detection, automated trading activities and financial advisory services for investors.
- **Marketing**: Product recommendations and news feeds on social media services.

- **Transportation**: Self-driving vehicles, real-time tracking and last-mile delivery optimisation.

Large language models (LLMs)

Large language models (LLMs) are data transformation systems. They are trained with large numbers of parameters, which are numerical values that developers adjust to shape the inputs and outputs of AI models. When a user inputs a prompt, the model generates text content in response.

LLMs are trained on extremely large datasets sourced from websites, blogs, forums and news articles etc. They contain millions or billions of parameters. In the case of OpenAI's generative pre-trained transformer (GPT) models, the first large language model GPT-1 contains 117 millions parameters, GPT-2 contains 1.5 billion parameters and GPT-3 contains 175 billion parameters. Increased parameters often lead to more complex models that can handle more complicated tasks and generate more nuanced text.

Users interact with GPT models through interfaces like ChatGPT. This feedback loop allows the model to continuously learn and improve over time based on user feedback.

Examples

Some notable LLMs are:

- OpenAI's GPT series of models (GPT-1, GPT-2, GPT-3, GPT-3.5 and GPT-4) used in ChatGPT and Microsoft Copilot
- Google's PaLM (Pathways Language Model) and Gemini
- xAI's Grok
- Meta's Llama (Large Language Model Meta AI) family of open source models
- Antropic's Claude models

Related term: **Machine learning**

Knowledge cut-off date

The knowledge cut-off date of a GenAI model is the date when the training data for a specific LLM was last updated. It defines the limitations of the model's understanding and knowledge.

> **Example**
>
> ChatGPT-3.5's current knowledge cut-off date is January 2022. The knowledge cut-off date for ChatGPT-4 is April 2023.

Chatbot

Chatbots are popular applications of GenAI and LLMs. A chatbot is a computer program that interacts with humans through natural language conversations. Some chatbots use LLMs to generate content according to user inputs.

Chatbot functionality is often embedded in customer services (e.g. banking, tax, logistics, ecommerce). Chatbots might be fine-tuned on an organisation's private datasets to answer specific queries.

> **Examples**
>
> Apple's Siri, Amazon Alexa, Google Assistant, ChatGPT, virtual assistants that provide customer services in applications

---

General-purpose AI

General purpose AI is a new paradigm where very large AI models become competent at a wide range of tasks, often without substantial modification or fine-tuning. General-purpose AI systems are often LLMs.

Some general-purpose AI systems are effective at processing a variety of inputs.They can process audio, video, textual and physical data and even complex medical or scientific data with enough training. Part of the commercial appeal of **foundation models** is their capacity for general-purpose applications.

> **Examples**
>
> GPT-3, GPT-4, BLOOM (BigScience large open-science open-access multilingual language model), BERT (Bidirectional encoder representations from transformers), Stable Diffusion and DALL-E. These are designed for general purposes and used for downstream natural language processing (NLP) tasks.

Foundation model

Foundation models, sometimes called **general-purpose AI systems**, provide a basis for future modification and fine-tuning for specific tasks. They are trained on vast amounts of data at scale, including images, text, audio, video and other data types like 3D models, with less emphasis on data quality so that they can be adapted to a wide range of downstream tasks.

> ### Examples
>
> Several companies, including OpenAI, Google, Meta (the parent company of Facebook) and Stable Diffusion, own foundation models. Many open-source foundation models can be downloaded from Hugging Face, GitHub, TensorFlow Hub and PyTorch Hub for further use. Popular foundation models include GPT, BERT and DALL-E 2.

Related terms: **Machine learning**, **Transformer architecture**, **Open-source models**, **General-purpose AI** and **Large language models**

Frontier model

Frontier models are larger than foundation models, with more parameters. They are potentially much more capable than existing models and raise additional **safety** risks particularly when developed for critical applications such as health, social welfare, defense and military.

> ### Challenges
>
> Different definitions of frontier models use different criteria to describe them. These definitions are not necessarily exclusive but rather adopt different assessment approaches, potentially leading to alternative policy discussions and interventions.
>
> - **Based on risk**
>
> The UK government defines frontier AI as consisting of highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities of today's most advanced models (see **AI Safety Summit: Introduction**).
>
> The paper **Frontier AI regulation: Managing emerging risk to public safety** by staff from Google, Open AI and the Centre for Long Term Resilience defines frontier AI models as highly capable foundation models

that could possess dangerous capabilities sufficient to pose severe risks to public safety.

- **Based on computational power**
  Training frontier AI models requires computer power exceeding 1026 integers or floating-point operations (**FLOPS**) (see the **Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence**).
- **Based on risk and computational power**
  The **EU AI Act** categorises a general-purpose AI model as having systemic risk if it has high impact capabilities and the cumulative amount of computation used for its training measured in FLOPS is more than 10^{25} (Article 51, Section 1 Classification rules, CHAPTER V GENERAL-PURPOSE AI MODELS).

Transformer architecture

Transformer architecture makes LLMs possible. Put simply, transformers convert text. This type of architecture was proposed in 2017 by 8 authors from Google in the paper **Attention is all you need**.

The software processes information by tokens and handle sequences step by step. Transformer models process long sequences simultaneously and capture context from the input. They allow developers to train larger networks with more training data at a given scale.

Examples

Google's BERT, OpenAI's GPT, Meta's BART, Bidirectional and Auto-Regressive Transformers.

More information

The central idea of transformer architecture is attention and self-attention.

Attention: When encoding a sentence, the transformer architecture pays attention to each of the words, figuring out which ones are needed to understand the whole sentence and where they appear, giving these higher attention scores.

Self-attention: A mechanism to contextualise words by paying attention to other words that make up its context in a body of text.

Visit **Transformers: The Google scientists who pioneered an AI revolution** for a visualisation of this concept.

Transfer learning

Transfer learning is a model's ability to apply information about one situation to another. In this way, the model builds on its internal knowledge.

In GenAI, products must adapt and work well in different situations. It's like teaching a model skills in one area and then letting the model use those skills in other areas too. In this way, models can learn from one thing and apply it to many different tasks, making them versatile and useful.

> **More information**
>
> A current trained model knows how to distinguish dogs and cats from other animals based on a lot of **data labelling and data annotation**. Now you want your model to identify a new animal, for example rabbits. You don't need to start fresh because cats, dogs and rabbits share similar features like fur, legs and ears but with some differences. You need to fine-tune the model to let it recognise rabbits. The model should be able to apply what it has learned about dogs and cats, the fine-tuned datasets, and apply that knowledge to recognise rabbits. Transfer learning makes the training process faster.

Relevant terms: **Machine learning**, **General-purpose AI** and **Large language models**

Open-source and closed-source LLMs

Open-source LLMs have publicly accessible source code and underlying architecture, allowing developers, deployers, researchers and enterprises to use, modify and distribute them freely or subject to limited restrictions.

Closed-source LLMs have proprietary underlying source code and architecture. They are accessible only under specific terms defined by their developers.

Open-source LLMs and closed-source LLMs can be accessed and deployed via application programming interfaces (APIs), which are sets of rules or protocols that allow two software programs to communicate with each other to exchange functionality. AI developers, deployers or users can integrate data, services and functionalities from other applications through provided APIs, rather than developing them from scratch. The following table presents different features of open-source LLMs and closed-source LLMs.

Open-source LLMs	Closed-source / proprietary LLMs
Description	
Open-source LLMs often mean that a project team has made their model architectures, source code and weight parameters publicly accessible on source code repository hosting platforms such as GitHub, GitLab and BitBucket.	Closed-source or proprietary LLMs do not disclose their source code, how the model is structured, the datasets used to train the system or the training process. These models are often developed commercially and may require licences or subscriptions for their use. Most frequently, we access the model through a provided API and use the model as provided by the owner.
Openness and adaptability	
Other programmers or developers can read the code, audit for bugs or security flaws and potentially contribute improvements back to the project.	Closed-source LLMs typically cannot be examined, audited or modified.
Examples	
Think about open-source projects as public toolkits of different models for developers and users to build their own tools. Access to open-source LLMs can unlock customisation opportunities. OpenAI's GPT models: GPT-2 and GPT-3 can generate human-like text based on input prompts. The code and pre-trained models are available on GitHub, allowing developers / deployers to experiment with and build upon them. Hugging Face's **Transformers** is an open-source library providing easy-to-use interfaces for accessing and **fine-tuning** NLP models. The library includes pre-trained models like BERT, GPT and RoBERTa (robustly optimised BERT pre-training approach), and tools for training and evaluation. BLOOM is an open-source multilingual language model. TensorFlow Models provides a collection of pre-trained models and code examples for GenAI tasks like image and text generation to allow developers to explore and experiment with different approaches to GenAI. PyTorch Hub, like TensorFlow Models, offers a collection of pre-trained models and code examples for GenAI tasks contributed by the community.	GPT-3 and GPT-4 by OpenAI: while GPT-2 is more openly accessible, GPT-3 users must use the provided API, and GPT-4 is a closed-source model. Gemini Pro (updated from Bard) and BERT by Google: • Gemini Pro is open to developers and enterprise customers of Google. It is accessed via the Gemini API. • BERT is openly published but many of its applications and some advanced derivatives need to be accessed via provided APIs. Turing-NLG by Microsoft and its more advanced iterations can only be used internally for enhancing Microsoft's products and services. • ERNIE by Baidu is accessible through a subscription model via Baidu Cloud.

Advantages	
Free to use – many open-source LLMs are free of charge. Customisability – open-source LLMs can be customised and modified for specific tasks. Transparency – some information about the model is made public, providing visibility into the model's inner workings. Community support – open-source LLMs are often supported by communities of machine learning engineers and developers. Software development environments and platforms like Amazon Web Services (AWS), Hugging Face and Azure are emerging as marketplaces where you can develop tools using your selected model.	Legal protection – companies and startups that use closed-source LLMs usually have legal agreements and terms of service that provide legal protection for their businesses. Data security – closed-source LLMs may come with enhanced data security features. Scalability – with the company's resources, closed source LLMs can be scaled more efficiently. Regular update – with resources from proprietary models, LLMs can be constantly maintained and updated. High-level support – companies may dedicate support for model deployment, integration and troubleshooting.
Concerns	
There are different ways of doing open source and the term is contested. However, for GenAI there are open-source models and ecosystems that have their own development, procurement and deployment pathways. Limited resources – open-source LLMs often rely on community volunteers with fewer resources for development, addressing bugs and updating. Security issues and misuses – exploitation of uncensored LLMs for malicious services without appropriate safety checks; **fine-tuned** data containing harmful content might be free for public use and accessible. **OpenAI's Usage policies** include universal policies, **safety best practices** and **moderation endpoint** to ensure their models are used in compliance with the law and without causing harm.	Vendor dependency – companies using closed-source LLMs can be dependent on the AI or machine learning development company. There may be challenges if the development company discontinues support. Limited flexibility and lack of transparency – closed-source LLMs are not suitable for experimentation due to limited access to the model's internal architecture and training data. Licensing and costs.

Token

A token is the smallest unit of data used by GenAI systems.

For text-based models like GPT, a token can be a word, part of a word, punctuation marks, spaces or other elements of the text. For image-generating AI models like DALL-E, a token is a pixel of the image. For audio-based AI models like MusicLM, a token might represent a short sound segment.

Tokens allow AI models to understand, memorise and generate meaningful responses. They play a vital role in memory capacity, determining how much information the AI model can recall.

> **Example**
>
> Different versions of ChatGPT come with different memory capacities. The ChatGPT Plus plan offers up to 8,000 tokens. This means that ChatGPT Plus can only remember information within the last 8,000 tokens during ongoing user interactions.

> **More information**
>
> Counting tokens is especially important when LLMs bill input and output tokens differently. For example, with **Anthropic's Claude 3 Opus** model, the cost for one million input tokens is US$15 and the cost for one million output tokens is US$75.
>
> Both input (the prompt or conversation history) and output (the model's response) tokens contribute to token usage. That is, if the input message uses 10 tokens and the model's response generates an additional 15 tokens, then the user, whether an individual or an organisation, will be billed for a total of 25 tokens.

Reinforcement learning from human feedback (RLHF)

Reinforcement learning from human feedback (RLHF) resembles the human learning process. When we learn a skill our teacher/instructor says things like 'well done' if we do something right or 'let's try it again and differently' to improve our skill.

Application

RLHF works a bit like this: human testers, trainers or programmers observe how AI software conducts tasks and provide feedback through ranks and scores.

The model tries different actions based on the feedback and learns which actions are more preferable based on the scores. Over time, with enough feedback and practice, the model gets better at conducting tasks, making decisions and achieving its goals even without human involvement.

Example

The system provides two possible outputs to a prompt, and the human indicates which one they prefer.

AI products like chatbots are often **fine-tuned** through RLHF. Outputs from an LLM are often conditioned by RLHF according to specific needs, norms, ethical guidelines and standards to make AI products more accountable to the community.

Diffusion models

Diffusion models are used for AI image generation. They work by destroying training data (by adding visual noise) and then learning to recover the data by reversing the noise.

Application

Imagine you have a picture of a dog. A diffusion model would take this picture and add some random changes or destroy a part of it, like making some parts of the picture blurry or changing some colours (i.e. adding noise). Then the model's job is to figure out what the original picture looked like before noise was added (i.e. denoising). Diffusion models learn from lots of examples to figure out how to undo the damage and reveal the original image.

Examples

DALL-E 2, Midjourney, Stable Diffusion

Inference

AI inference is the process of applying trained machine learning models to new, unseen data to derive meaningful predictions or decisions. When users give a GenAI system a prompt, the computational system used to produce the output is called inference. The energy required for inference is much lower than training a model but is still significant and is a large part of the cost of using a GenAI system.

Applications

Almost any real-world AI application relies on AI inference. Some of the most commonly used examples include:

- LLMs – a model trained on sample text can parse and interpret texts it has never seen before.
- Predictive analytics – once a model has been trained on past data and reaches the inference stage, it can make predictions based on incoming data.
- Email security – a machine learning model can be trained to recognise spam emails or business email compromise attacks, then make inferences about incoming email messages, allowing email security filters to block malicious ones.

Operational

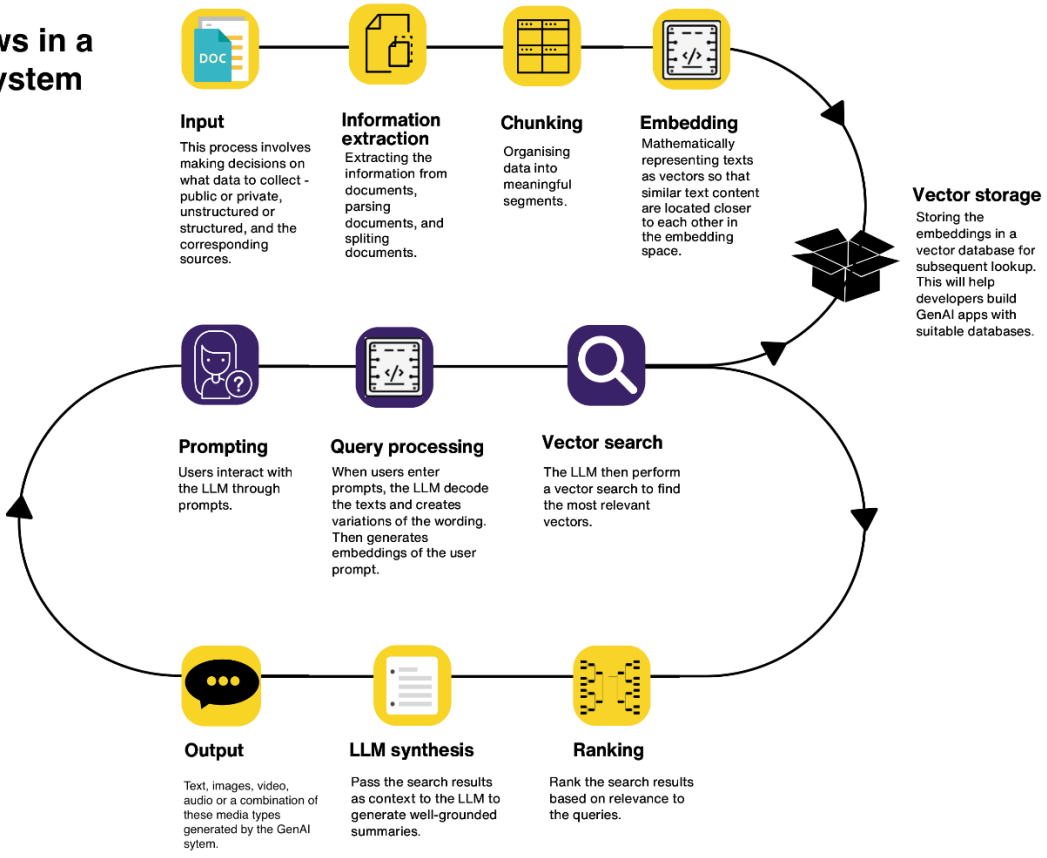
This section begins by illustrating data flows within a GenAI system, followed by an explanation of the general operation of a GenAI system through development, distribution and deployment and use.

The development of GenAI starts with the acquisition of datasets, which often adhere to data licensing agreements. To ensure software recognises datasets effectively, data labelling and annotation are needed to prepare the data for training.

AI libraries and machine learning environments serve as crucial intermediaries in the distribution phase, facilitating the deployment and use of the model.

During deployment and use, models can be fine-tuned by deployers or machine learning / operation teams to ensure they meet user demands.

Data flows in a GenAI system



Development

Datasets

LLMs are trained from different types of datasets. The material in datasets may be protected by intellectual property or information privacy laws. Organisations like **Common Crawl** scrape the public internet periodically and create gigantic datasets that can be used for model training.

Application

Datasets for fine-tuning models are typically found on platform marketplaces like Hugging Face, AWS and Azure. The terms are often commercial, and typically include information about how the dataset was created and its compliance with relevant legal rules.

Most AI companies have trained their data on the open web, but this has prompted several legal challenges under copyright and privacy law.

Data licensing

Entities that control large amounts of text, visual, musical, code or video content may license it to AI firms for a fee. Datasets may also come with licensing stipulations that dictate what others can do with that dataset, for instance whether they can share it, modify it or use it for commercial purposes.

Examples

Organisations with large desirable datasets like **Reddit** or **New York Times** are now seeking financial compensation when AI models train on their data.

On the other end, **AI companies** have arguments against paying for copyrighted content:

- **Adobe** claims that they license all the copyright material used to train their commercial models.
- **Google** claims that AI training is like reading a book. If training could be done without creating copies, there would be no copyright questions.
- **Microsoft** claims that changing copyright law could hurt small AI developers.

Developer

AI developers belong to a broader category of programmers and engineers. They write code and algorithms to enable machines to perform tasks that normally require human intelligence. They build models in-house by training on a mix of public and private data for various applications, from chatbots and virtual assistants to self-driving cars.

In emerging AI regulations, developers are responsible for creating AI software products and complying with regulations.

> **Example**
>
> The **EU AI Act** imposes different obligations on AI 'developers', 'deployers' and 'users'. Many of the regulatory obligations around **safety** and auditing must be satisfied by developers before their product can go to market.
>
> With organisations and individuals assuming overlapping roles within the AI supply chain, third-party organisations have emerged, offering services like Future of Life's **The EU AI Act Compliance Checker** to allow individuals and organisations to identify their roles and obligations under certain legislation.

Data labelling and annotation

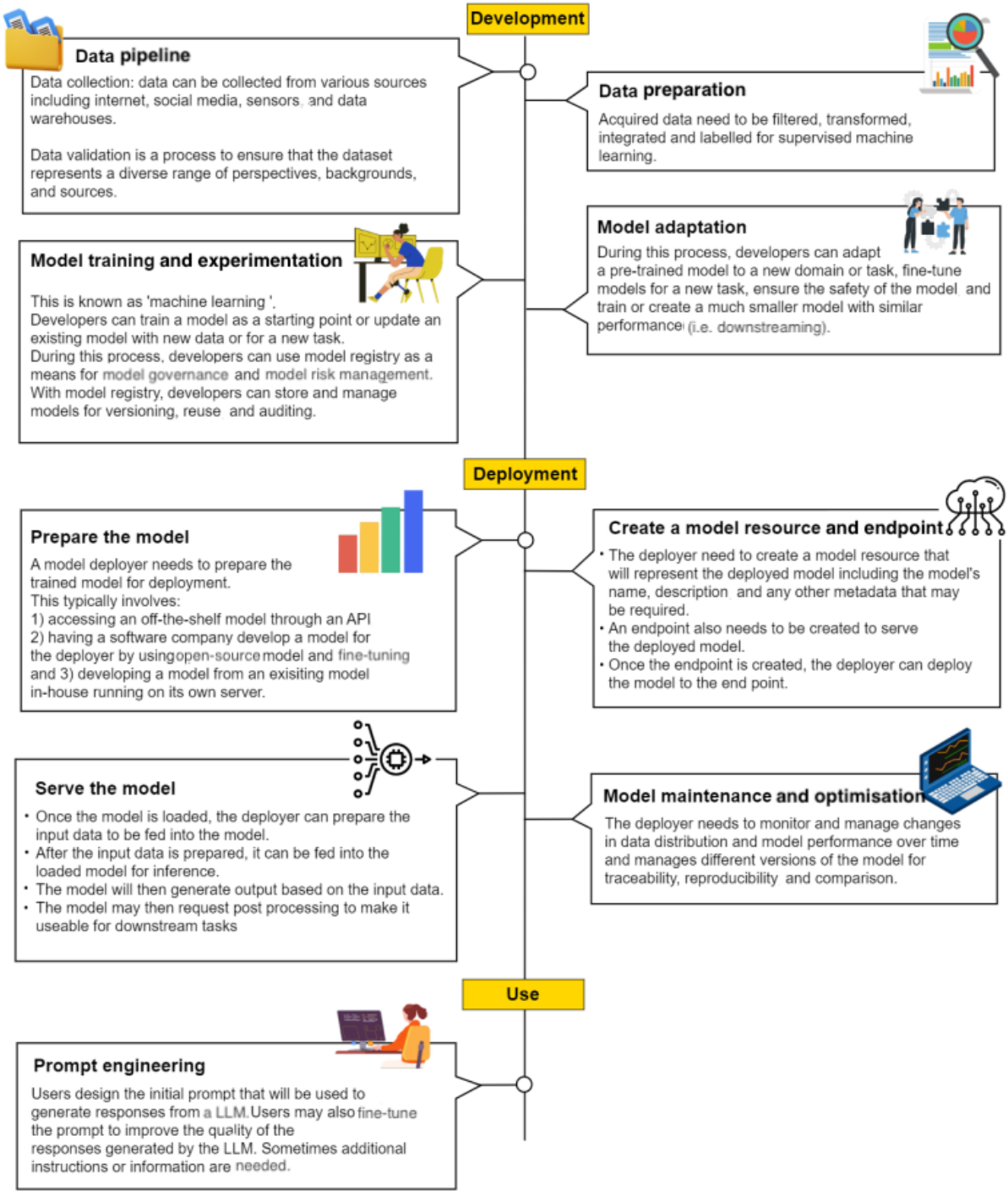
Data labelling and annotation describe the process of tagging or labelling text, images, audio and video data for AI training.

Data annotation	Data labelling
Meaning	
Data annotation provides information about a set of collected data so machine learning algorithms can easily recognise the dataset. Information about data, or known as metadata, includes description, administrative details, statistics and legal information about the dataset. Examples of metadata may include how the data was created, the means of creation, the purpose of the data, the time and date of creation, the location of creation, the standards used, file size, data quality and the source of the data.	Data labelling attaches meaning to different types of data to train a machine learning model. It identifies a single entity from a set of data.
Purpose	
Data annotations can be used for visual-based perception models. It helps models to identify and process objects and people in images and videos, a capability known as ‘computer vision’.	Data labels are used to identify dataset features for NLP algorithms. Data labelling is used for training advanced algorithms to recognise patterns within the datasets in the future.
Applications	
Data annotation is a fundamental element in creating training data for computer vision. Annotated data helps to train machine learning algorithms to see the world humans see.	Data labelling is more complicated than annotation. It identifies key features in the data while minimising human involvement. Real-world use cases include NLP, audio and video processing, computer vision etc.

Distribution

Supply chains

An AI supply chain refers to the process of creating, sourcing and integrating the components needed to develop and deploy AI systems or products.



This diagram builds on Ali Arsanjani's [The generative AI life-cycle].

Concerns

Developers often use pre-made software parts, which makes it challenging to know who's responsible for following the rules and regulations. Understanding AI supply chains requires us to ask:

- where the data used to train the AI comes from
- whether the AI model is open for anyone to use or owned by an entity
- how organisations tweak an AI system by using different datasets and fine-tuning models.

AI libraries

AI libraries include pre-written codes and algorithms for common AI tasks, such as data preprocessing, model training and evaluation.

Examples

PyTorch is one of the most popular machine learning libraries, alongside TensorFlow. It offers free and open-source software released under the modified **BSD licence**.

> **Development:** PyTorch was developed by MetaAI and is now part of the Linux Foundation
>
> **Notable applications:** Tesla Autopilot, Uber's Pyro, Hugging Face's Transformer

TensorFlow is a free and open-source software library for machine learning and AI. It is particularly used for training and **inference** of **deep neural networks**.

> **Development:** TensorFlow was developed by the Google Brain team
>
> **Notable applications:** GE Healthcare, Twitter's ranking system, the image-captioning software DeepDream

The **Hugging Face** ecosystem, including Transformers and the Hugging Face Hub, contains libraries for tasks such as dataset processing, model evaluation and machine learning demos. Its transformers library is built for NLP applications. Hugging Face Hub also allows users to share machine learning models and datasets.

> **Development:** Hugging Face is a French–American company that originally developed a chatbot app targeted at teenagers

> **Notable applications:** partnership with AWS – Hugging Face's products are available to AWS customers and can be used as building blocks for their custom applications

---

Machine learning environments

Machine learning environments are where machine learning models are built, trained and deployed. Provided by the big tech corporations like Amazon and Microsoft, these environments contain fully managed infrastructure and tools to enable reproducible, auditable and portable machine learning workflows across different environments. Deployers can run training scripts or host service deployments, often referred to as computer targets, within these environments.

> **Examples**
>
> Microsoft Azure
>
> - Amazon SageMaker Studio and SageMaker Studio Classic

Deployment and use

Fine-tuning

Fine-tuning happens during model development and deployment. It involves modifying a trained AI model with a smaller, targeted fine-tuning dataset. Fine-tuning maintains the original capabilities of a pre-trained model while adapting it to suit more specialised use cases.

Example

Imagine you have a model that's been trained to recognise birds in pictures, but you want it to be better at recognising Swift Parrots specifically. You deploy a model you found from an AI library, fine-tune it by feeding it lots of pictures of Swift Parrots, adjust its parameters and even score the performance (i.e. reinforcement learning) until it gets better at identifying the target.

More information

Large training datasets are available from development platforms like Github, Azure AWS, Papers with Code and Hugging Face.

The Data Provenance Initiative identified that across Github, Papers with Code and Hugging Face more than 70% of pre-trained datasets had no data licences. For those that did, roughly half were incorrect and were more permissive than the dataset creators had intended.

Read more in **Public AI Training Datasets Are Rife With Licensing Errors** and **The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI**.

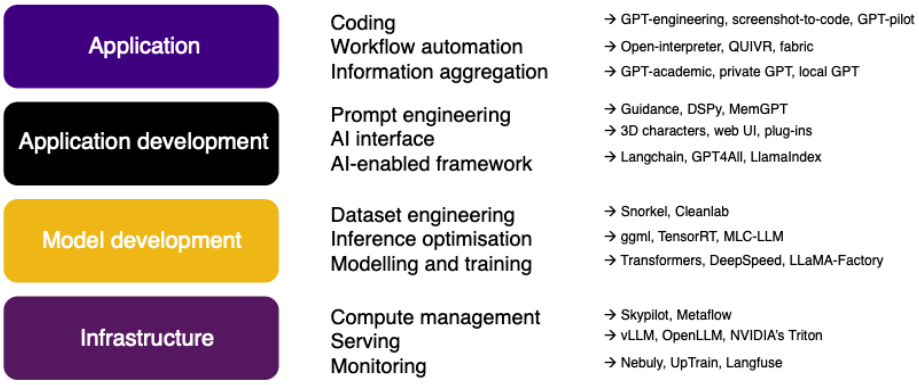
Deployer

While AI developers create the AI system, AI deployers make them available for use in real-word applications. In some cases, the distinction between a developer and a deployer is minimal. However, deployers may be the organisations that make AI tools available to others.

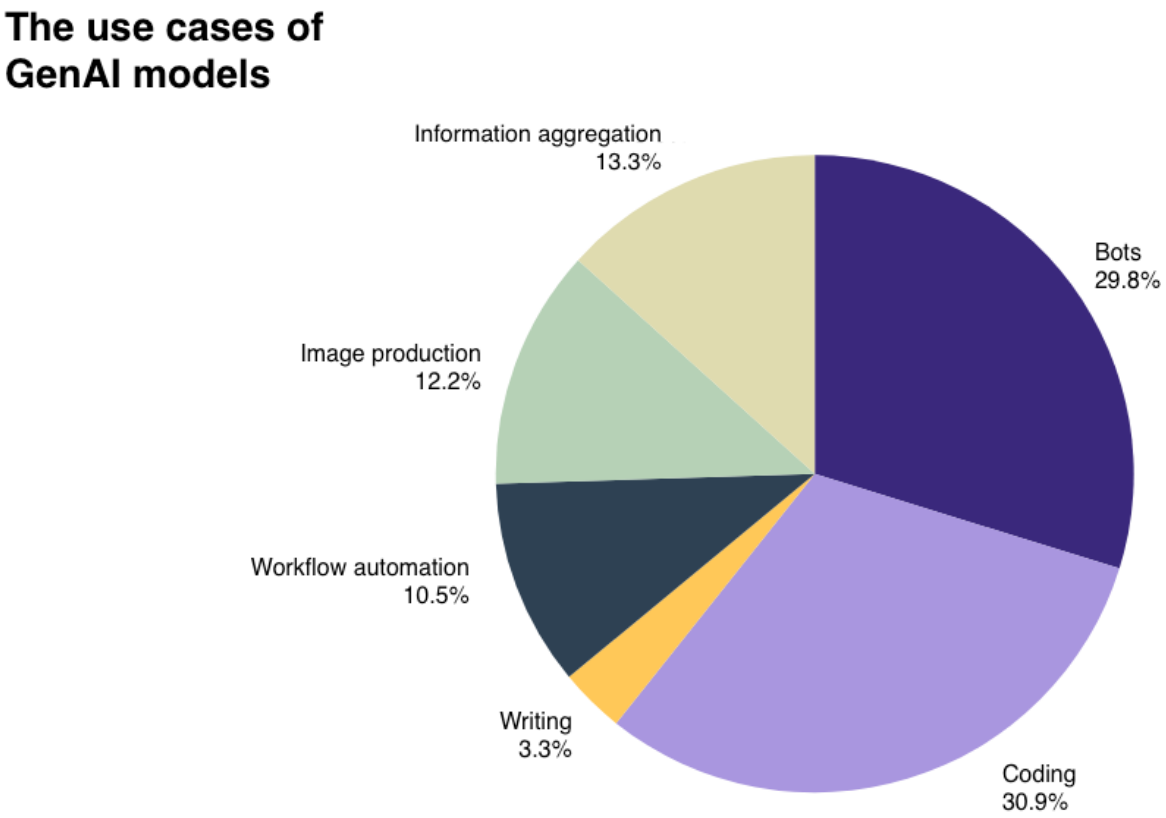
Applications

Let's think about an AI model being composed of 4 layers: infrastructure, model development, application development and applications. Different levels are open to deployment and associated with different levels of technical capacities.

The new AI stacks



AI deployment can be as simple as a dashboard at the application level or as complex as the augmentation of a foundation model through prompt engineering or specific GenAI architecture at the infrastructure level.



As shown in the pie chart, open-source AI tools are deployed for coding, communications (e.g. WhatsApp bots, Slack bots), information aggregation (e.g. asking a model for meeting summaries), image production, workflow automation and writing.

This section references **Chip Huyen's research blog**.

User

Users are the individuals or organisations that interact with AI technology through text, voice, images or other inputs. Users can also be anyone whose rights and activities are influenced by AI system outputs.

> **Example**
>
> We are users when we chat with virtual assistants like Siri or Google Assistant.
>
> Even though individuals do not directly engage with AI models, individuals' rights and activities can be affected by how organisations use AI models.
>
> For example, if AI is used to screen job applications and select candidates for interviews, the AI decision-making system might influence the outcomes.

Regulatory

This section explains the risks and concerns with GenAI systems and how existing regulations, ethical guidelines and practices address these.

GenAI systems raise many concerns, such as privacy breaches, hallucination, safety issues, lack of transparency and ecological impacts.

Existing regulations and legislation, such as copyright laws, data governance and security, demonstrate our capacity to regulate some aspects of GenAI systems. Complementing these are emerging AI-specific guidelines and regulations, including AI standards, guardrails, regulatory sandboxes, human oversight, AI auditing, explainable AI, post-market surveillance and monitoring and FLOPS to respond to increased risks associated with AI.

Risks

AI regulation is developing as risk-based regulation. This approach imposes obligations depending on the risk level of the system. Most regulatory proposals have 3 risk categories: low risk, medium risk and high risk.

The measurement and mitigation of AI-generated risks requires human oversight, AI auditing, transparency and explainability.

> **Challenges**
>
> Measuring risks is not easy! Risk is a spectrum and depends on a number of factors that are difficult to quantify. **Standards** are being created to help developers, deployers and users specify the appropriate level of risk for the AI system.

> **Concerns**
>
> Using AI to play chess is low risk compared to using AI for staff recruitment, which generates medium risks (e.g. gender or racial discrimination). Using AI-enabled robots for medical surgery and in self-driving cars to make real-time decisions generates high risks, as such risks are systematic, perpetual or hard to reverse.

Privacy and data protection

The wide use of AI systems presents challenges to privacy. The datasets used to train **foundation models** and other machine learning systems often include personal information without consent. There have been cases where chatbots reproduce personal information from training data in response to prompts. Privacy and data protection laws impose obligations on entities that collect and process **personal information**.

These obligations relate to collecting, using and disclosing personal information; data quality and security; and deleting, minimising or deidentifying personal information. Protecting individuals' data often requires that individuals consent to the use of their personal information, while systems limit use to only those that an individual would reasonably expect and that relate to the original reason for collection. Privacy and data protection laws also create rights for individuals to have access to information about them held by other entities.

Legislation

The **Privacy Act 1988** is a federal law which does not cover local, state or territory government agencies, except the Norfolk Island administration. Most Australian states and territories have equivalent legislation covering their public sector agencies (see **State and territory privacy legislation**).

For example, in Victoria, individuals' privacy rights are protected by the **Privacy and Data Protection Act** 2014 (Vic) (PDP Act). The PDP Act contains **10 information privacy principles** (IPPs), which are the core of privacy law in Victoria and set out the minimum standard for how Victorian public sector organisations should manage personal information.

Personal information

Under the PDP Act, personal information is information or an opinion about you where your identity is clear or where someone could reasonably work out that it is related to you. It can include your name, email address, postal address, phone number, signature, fingerprint, photographs or surveillance footage of you, comments written about you (whether true or not), your financial details and more (see more at OVIC's **Your privacy rights**).

Challenges

AI models are trained on large datasets and our real-time inputs are sometimes included in the **AI training** process. Organisational data and personal and sensitive information may then be inadvertently disclosed to companies in control of the AI software being used.

When analysed or correlated, a combination of seemingly non-personal information can make individuals identifiable.

Further, AI's predictive capabilities can exceed what an individual discloses, resulting in the collection of new personal information generated by the AI model's output (read more at OVIC's **Artificial Intelligence – Understanding privacy obligations**).

More information

The Office of the Australian Information Commissioner reported a 19% increase in data breaches from July to December 2023, with notable impacts in health services, finance, insurance, retail and the Australian government, where our sensitive

information is stored (read more in OAIC's **Notifiable data breaches report July to December 2023**).

If an organisation uses GenAI services that include accessing a model through an API, then the organisation should ensure that those data flows do not breach relevant privacy and data security legislation, such as the **Victorian Information Privacy Principles** and **Australian Privacy Principles**.

Software providers must provide assurance around how their data is stored, processed and secured.

Hallucination

Hallucination refers to AI models making up facts to fit a prompt's intent. When an LLM processes a prompt, it searches for statistically appropriate words, not necessarily the most accurate answer. An AI system does not 'understand' anything, it only recognises the most statistically likely answer. That means an answer might sound convincing but have no basis in fact. This creates significant risks for organisations that rely on chatbots to give advice about products or services because that advice might not be accurate.

Real case

In 2023, 2 lawyers in New York used **ChatGPT** in their legal research for a personal injury case. Their legal brief, submitted to the court, included 6 fake case citations generated by ChatGPT (i.e. hallucinated detail). The lawyers then conducted a fact check with ChatGPT, which approved the results generated. The lawyers and the law firm received penalties as a result.

Concerns

AI systems cannot 'tell the truth'; they are prediction models. Some people think AI systems only get things wrong occasionally while otherwise telling the truth. That is not true. AI systems make output mistakes based on the input data they use, so it is important to verify the accuracy of the output before relying on the model. This is especially important when people rely on AI systems to make decisions that affect themselves or others.

Tip

One way to identify AI-generated images is by looking at the detail (e.g. hands, teeth, hair, unnaturally smooth skin, accessories, watermarks in the background, stereotypes). Because current AI systems struggle with complex detail they might be trained on simpler stereotypes, leading to odd or unrealistic results.

Safety

In AI regulation, AI safety means that AI should be designed, developed and used in a way that is human-centric, trustworthy and responsible (see **The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023)**.

The main concern of AI safety is to make sure that AI systems are developed, deployed and used in ways that align with the norms and values of the broader public or specific user groups.

Guidelines and legislation

AI safety is central to numerous laws and AI ethics guidelines developed at national and supranational levels, such as **The Bletchley Declaration**, **the EU AI Act**, **The Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence.**

The Australian government is already undertaking work to strengthen existing laws (e.g. the **Online Safety Act 2021**) to address known harms with AI.

Along with ethical guidelines, including **Australia's Artificial Intelligence Ethics Framework** and **Australia's AI ethics principles,** the **Australian government's interim response to safe and responsible AI consultation** describes actions to help ensure that AI is safe and responsible.

Measurement

Safety measures include testing AI systems thoroughly before using them, designing systems to be robust enough to handle unexpected situations without making mistakes and having **guardrails** in place.

Transparency

AI transparency is a fundamental regulatory and ethical requirement; however, transparency has different meanings for different stakeholders (e.g. developers, deployers, users, regulators).

Transparency represents how well users, regulators and other stakeholders understand how the system operates and how it was built. This is important for ensuring trust in AI outputs.

Transparency is related to the **openness** and **explainability** of AI systems.

> **Forms of transparency**
>
> Different stakeholders will be interested in different forms of transparency.
>
> For general public users, transparency can be a sense of what the AI system is doing and why.
>
> For a regulator, transparency can enable auditing how a prediction and decision is made by AI systems in detail or mean transparency in the data used to train a system.
>
> Engineers might be more interested in the internal mechanism and parameters of AI systems.

Related term: **Explainable AI**

Copyright

Copyright law typically applies to original works of authorship, such as literary works, artistic works and computer programs. Copyright material has been included in AI training datasets without permission or rights clearance.

Challenges

The application of copyright to datasets raises complex issues and how rights-holders opt out or demand licensing fees will depend on how the law develops and will vary in different jurisdictions. Some jurisdictions may create exceptions for model training, and others may determine that the training process does not replicate the 'expressive' part of a copyright work and thus is not a breach.

Real cases

There are currently court cases and policy reforms addressing how to manage the inclusion of copyright material in datasets. There is also the risk that GenAI responds to prompts with outputs that violate copyright. This might be an exact reproduction of a text or image or the unauthorised use of a copyrighted character.

Several coders have brought a lawsuit against GitHub, Microsoft and OpenAI because GitHub Copilot was trained and developed on existing open-source code, leading to questions of attribution.

Several visual artists filed a class-action lawsuit against Stable Diffusion, Midjourney and DreamUp, all of which generate images based on text prompts from users. The case alleges that these AI tools violate copyright by scraping images from the internet to train their models.

Getty Images alleges that Stable Diffusion's use of its images to train models infringes on copyrights and that some images even contain the Getty watermark.

Many GenAI systems providers now indemnify users in case an AI output breaches copyright.

Data governance and security

Data governance refers to the system of rules for management and control of data.

Data governance requirements may involve compliance with laws, standards or ethical guidelines to ensure data within an organisation is collected, stored and used appropriately.

Within an organisation, data governance responsibilities include ensuring data quality (i.e. data is accurate) as well as data security (i.e. data is appropriately encrypted and protected).

This may mean additional care when using organisational data to fine-tune a GenAI system or when using organisational or sensitive data as part of a prompt.

Data life cycle management refers to the processes for managing data from creation to disposal, including data archiving and removal.

Challenges

The Australian government will develop a whole-of-government **Data Governance Framework** for public services to ensure data quality, privacy, authority and innovation.

There are also sector-based data governance frameworks, for example, by the **Australian Commission on Safety and Quality in Health Care**, **Australian Institute of Health and Welfare**, **National Archives of Australia** and the **National Health and Medical Research Council**.

In Victoria, there is the **Victorian Protective Data Security Framework and Standards** and the **Privacy and Data Protection Act 2014**.

AI standards

International AI standards serve as a global governance mechanism to help achieve AI policy goals. Standards organisations create standards through stakeholder input and cover various technical and regulatory dimensions of AI systems. Compliance with standards may be required for AI products to be market-worthy; procurers may also require compliance with specific standards. In the EU system, if developers comply with standards, then they likely comply with their obligations under the **EU AI Act**.

Applications

A wide range of national, regional and international organisations are engaged in developing and refining AI standards.

ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system is the world's first AI management system standard that provides valuable guidance for this rapidly changing field. It addresses challenges posed by AI technologies, including ethics, transparency and continuous learning. For organisations, the standard sets out a structured way to manage risks and opportunities associated with AI, balancing innovation with governance.

- **ISO/IEC 23894:2023 – Information technology – Artificial intelligence – Guidance on risk management** provides guidance on how organisations that

develop, produce, deploy or use AI products, systems and services can manage risks related to AI.

- **ISO/IEC 23053:2022 – Framework for Artificial Intelligence (AI) Systems using Machine Learning (ML)** establishes an AI and ML framework for describing a generic AI system using ML technology.
- IEEE is an engineers' professional organisation with a subsidiary **Standards Associations (SA)** whose standards address protocols for digital products, including ethernet and wi-fi. Its AI standardisation processes are part of a larger **IEEE Global Initiative** on Ethics of Autonomous and Intelligent Systems.
- The ITU (International Telecommunication Union) has historically played a role in standards for information and communications technologies, particularly in telecommunications. Following the 2018 AI for Good Global Summit, ITU created a number of **Focus groups** on AI in relation to health, environmental efficiency, autonomous and assisted driving, and data.
- As part of European AI regulation, the European Commission has asked the European Committee for Standardisation (CEN) and the European Electrotechnical Committee for Standardization (CENELEC) to develop **harmonised technical standards** for companies that wish to comply with the EU AI Act.

Specialised bodies may also participate in creating international standards. These bodies can be treaty organisations such as the International Atomic Energy Agency and the International Civil Aviation Organization.

More information

According to the OECD AI Policy Observatory, there are more than **1,000 AI policies** and principles developed by governments, international organisations and private entities.

In Australia, there is the **Australia's Artificial Intelligence Ethics Framework** and **Australia's AI Ethics Principles**. **The Australian government's interim response to safe and responsible AI in Australia consultation** highlights that more needs to be done to ensure the safe and responsible development and deployment of AI.

Guardrails

AI guardrails are predefined standards, limitations and operational protocols to prevent AI systems from making decisions or taking actions that could lead to harmful or unintended consequences.

Guardrails are often sets of algorithms or rules that filter and edit inputs (i.e. prompts) and outputs of GenAI systems to ensure that outputs comply with legal and safety requirements. These guardrails are designed to prevent outputs from breaching copyright, produce political misinformation, create biased or discriminatory information or generate hate speech. This has proved very difficult to achieve in practice but remains an important aspect of AI system design.

> ### Jailbreaking
>
> Jailbreaking refers to techniques for getting around a system's guardrails. This might mean tricking a chatbot into generating outputs that would otherwise be a violation of its policies. Most LLM systems are susceptible to jailbreaking, meaning all guardrails have limited efficacy.

Regulatory sandbox

Regulatory sandboxes are software testing environments where businesses can conduct limited testing of innovations and test their regulatory status.

This allows businesses to receive feedback from experts, including regulators, investors, innovators and other stakeholders, regarding the potential and viability of the innovation.

> ### Application
>
> Regulators and innovators use regulatory sandboxes to offer individualised legal guidance. Sometimes regulatory sandboxes are used to make legal exceptions for certain innovations, enabling innovators and the public to experience the impact of new technologies as if they had already been established as safe and effective.

Legislation

Regulatory sandboxes have been included in the **EU AI Act Article 53** and by the **Australian Securities & Investments Commission**.

Human oversight

Human oversight is the requirement that human actors oversee the output of AI systems to ensure that systems create accurate and accountable results. There are different forms and degrees of human oversight, depending on the context and purpose of the AI system. Human oversight is usually required in high-risk systems.

Application

Humans can intervene in AI systems in different forms and at different degrees:

- **Humans in the loop:** even though AI systems can do many tasks on their own, humans can make decisions, provide approvals or improve the system by correcting or preventing potential AI errors. Humans in the loop may not need say 'yes' or 'no' every step of the way as long as there is still a person watching over the process to make sure everything runs smoothly.
- **Humans in command:** means that humans have ultimate control over the AI system and make all the important decisions.

Sometimes human oversight requires a final human decision-maker, but often this is not the best way to implement oversight or may be impossible.

Legislation

The EU AI Act Article 14 requires that high-risk AI systems have appropriate human oversight from design to operation. Human oversight should be proportional to the potential impact and severity of harm that the AI system can cause.

The following Australian documents refer to and explain the EU rules:

- **Australia's AI Ethics Principles**
- The OAIC **submission** to the Department of Industry, Science and Resources – Safe and responsible AI in Australia discussion paper

AI auditing

AI auditing involves humans – normally researchers, programmers and regulators – looking closely at AI systems to evaluate risk and ensure AI systems act fairly and safely, and comply with relevant laws, regulations and ethical standards.

Different AI auditing methods have different advantages and limits:

- Technology-oriented audits focus on the properties and capabilities of AI systems.
- Process-oriented audits focus on technology providers' governance structures and quality management mechanisms.

Some auditing methods may be simple compliance checkboxes; others may be comprehensive assessments of how AI systems might affect users and other stakeholders.

> **Application**
>
> AI auditing is a rapidly growing field of industry practices. Emerging courses, certifications and associations aim to professionalise the AI auditing industry.

Explainable AI (XAI)

Explainable AI or XAI is a set of tools, techniques and algorithms designed to produce high-quality interpretable, intuitive, human-understandable explanations of AI decisions. Many emerging AI regulations require some degree of explanation for higher risk system outputs.

The goal is not to prioritise complete explainability over performance or vice versa. Organisations should disclose the limits of transparency in the system to find a balanced approach that considers the risks and benefits of each AI application, while also considering human and environmental implication.

> **Application**
>
> The explainability of AI-driven decisions is linked to the **transparency**, fairness and trustworthiness of the decision made, and sometimes the interpretability of AI systems.

Advocacy for explainable AI

Governments

The Australian Government's **The AI in Government Taskforce** outlines transparency and explainability as one of the 4 key principles to ensure the Australian public sector implements AI safely and responsibly.

Research institutes

OpenAI, as an AI research and development company, has been developing a new tool to explain large language models' behaviours.

Civil society

Advocacy entities like the **Electronic Frontier Foundation** (EFF) and the **American Civil Liberties Union** (ACLU) consistently stress the significance of AI transparency, especially when outcomes directly affect individual rights.

Challenges to explainable AI

Complexity vs Simplicity

It is complex to make AI systems explainable. Technical complexity of AI systems has made it challenging for end users to understand how a decision is made by AI systems. Oversimplification can lead to unfaithful explanation of the decision. It is challenging to strike a balance of accuracy, explainability and future accessibility.

Transparency vs Vulnerability

With the decision-making process being made explainable, it can also make the system open to exploitation and attack.

Generalisability vs Particularity

General explainability of AI systems cannot capture specific or intricate patterns in data. Simplification and generalisation can lead to potentially lower performance.

More complex models can identify and use more particular and complicated patterns. Their outputs can be more difficult for people to interpret.

Not all AI applications require the same degree of explainability. For example, AI used in healthcare diagnoses should require higher transparency.

Post-market surveillance and monitoring

Post-market surveillance refers to monitoring the ongoing performance and safety of an AI product or service after it has been released to the market.

Application

Post-market surveillance may involve collecting and analysing data on the use of the product or service in real-life conditions. If issues are detected, information from post-market surveillance can help corrective action, such as updating product design, improving user instructions for use or product recall.

> **Example**
>
> In 2022, Meta **pulled** the public demo of Galactica, an LLM designed for science research, 3 days after its launch due to its inability to distinguish facts from falsehood (i.e. **hallucination**).
>
> In 2024, Google **paused** its Gemini AI model's image generation of people due to inaccuracies in some historical depictions.

Floating-point operations per second (FLOPS)

Floating-point operations per second (FLOPS) in the context of computing and artificial intelligence are often used to measure the processing power or performance of hardware devices like CPUs (central processing units) and GPUs (graphic processing units).

Sometimes **frontier models** are defined by FLOPS, and some regulatory proposals have stricter compliance obligations for systems that surpass FLOPS thresholds.

Higher FLOPS values indicate higher computational power, more complex AI models, faster data processing, better graphic performance and higher energy consumption.

> **Application**
>
> The US **Executive Order** on AI specifies different regulatory requirements for systems that use different amounts of computing power. It states that federal safety and security reporting requirements apply to any model trained using either:
>
> - a quantity of computing power greater than 10^{26} FLOPS
> - primarily biological sequence data and a quantity of computing power greater than 10^{23} FLOPS.
>
> So far, these thresholds are beyond anything on the market.
>
> The largest GPT-3 model required approximately **3.14E+23** = $3.14*10^{23}$ FLOPS for training, with 174,600 parameters and 300 billion training tokens.

There are research institutes, such as **Epoch AI**, tracing and documenting information about AI models, including estimates of their training computation power.

Ecological

Training, deploying and using AI systems contribute to the global CO_2 emissions. Typically, more powerful AI models require more energy. The servers that power AI models also generate considerable heat and are often water-cooled. The amount of water needed to train an AI model is immense. A team of researchers **disclosed** that "training GPT-3 in Microsoft's state-of-the-art U.S. data centers can directly evaporate 700,000 liters of clean freshwater, but such information has been kept a secret." And running GPT-3 inference for 10-50 queries **evaporate** 500 mililitres of water depending on when and where the model is hosted.

LLMs are among the biggest machine learning models, spanning up to hundreds of millions of parameters, requiring millions of GPU (graphic processing units) hours to train and emitting carbon in the process.

OVIC
Office of the Victorian Information Commissioner

ADM + S
ARC Centre of Excellence for Automated Decision-Making and Society

admscentre.org.au/genai-concepts