GENERATIVE AI LANDSCAPE

SHAPING THE TECHNOLOGY OF TOMORROW

KIRTANE & PANDIT

TABLE OF CONTENT

THE ECONOMIST, IN 2017, FAMOUSLY DECLARED THAT DATA, AND NOT OIL, IS NOW THE WORLD'S MOST VALUABLE RESOURCE

1.0 EXECUTIVE SUMMARY

The Economist, in 2017, famously declared that data, and not oil, is now the world's most valuable resource – a statement that remains perfectly valid in today's digital world. Today, a majority of digitally mature companies have an AI strategy in place. Generative AI, a subset of artificial intelligence, is becoming the transformative force, with 65% of organisations reporting regular use of generative AI applications.

Generative AI applications can generate entirely new content based on the prompt or question asked and deliver such content across six main modalities: text, image, audio, video, code, or 3D graphics. According to a 2023 survey, 75% of the use case value was generated across four areas: customer operations, marketing and sales, software engineering, and R&D. Generative AI is already being applied and has shown tremendous utility in sectors such as life sciences, healthcare, banking, software, insurance, and financial services.

The rise of generative AI is reflected by its projected global market growth, from USD 20.9 billion in 2024 to USD 136.7 billion by 2030, registering a compound annual growth rate (CAGR) of 36.7% between 2024 and 2030. Generative AI has the potential to raise global GDP by 7% (USD 7 trillion) and increase productivity by 1.5% by 2033. The AI market in India is also witnessing robust growth, projected at a Compound Annual Growth Rate (CAGR) of 25-35%. India's spending on AI and GenAI, including software, services, and hardware for AI-centric systems, is projected to reach USD 6 billion by 2027.

Globally, the generative AI landscape reflects rapid innovation and funding flows despite significant risks. Organisations can spend between USD 5 to 20 million for tuning, customising, or simply deploying a generative AI model or application, which is why the generative AI market is dominated by a few companies such as OpenAI and NVIDIA. OpenAI's ChatGPT surpassed 1 million users within five days of its launch in 2022, and in 2024, it became the fastest-growing application in history. OpenAI and Microsoft (which is also a substantial shareholder in OpenAI) together accounted for 69% of the GenAI foundation models and platforms market in 2023, and NVIDIA, which was predominantly a gaming company, dominates the data GPUs and AI chips market with an 80% share. In 2023, 90% of the global funding on GenAI foundation models was invested in just three start-ups – OpenAI, Anthropic, and Inflection. For new entrants to the market, the most significant opportunity over the next 3-5 years is in building end-user applications.

Despite the advancements in this new technology, the reality of implementing generative AI is complex. Many generative AI projects fail to move beyond the proofs-of-concept (POCs) stage, and recent studies estimate the failure rate of AI projects between 83% and 92%. The technology also comes with its own share of risks and challenges, the most significant being misinformation and inaccuracy in responses, data security, use of copyrighted or confidential data for generative AI training, and misuse of the technology, for example, deepfakes.

There are also concerns over the ethics of AI, such as bias and discrimination, lack of transparency by AI companies, and a substantial carbon footprint (for instance, OpenAI's GPT-3 released 502 tonnes of carbon emissions during training, which is the equivalent of 1,260 one-way flights from New York to London).

Equally concerning is the fact that generative AI could impact up to 300 million jobs globally, i.e., 23% of jobs, and substitute up to 25% of current work in the US and 24% in Europe while complementing most of the remaining work. On the other hand, generative AI has the potential to create jobs as well. The number of AI and machine learning specialists is predicted to increase by 40% by 2027, thereby creating around 1 million new jobs across all industry verticals, and highlighting the need for at least 80% of the engineering workforce to upskill by 2027.

India has also found itself in the midst of this generative AI boom, with the Indian government committing INR 10,300 crore to the IndiaAI Mission over the next 5 years to build over 10,000 GPUs, and NVIDIA partnering with many Indian companies and start-ups to boost India's generative AI ecosystem. India also has highly skilled AI professionals, and is expected to surpass the US in the number of AI developers on GitHub by 2028.

However, some challenges remain, with some data suggesting that only 15-20% of POCs by Indian companies have been rolled out into production, and nearly 80% of Indian GenAI start-ups reporting earnings of less than USD 100K. Although commercial viability is an issue, the number of Indian generative AI start-ups in 2024 increased by 174, with cumulative funding in these start-ups since 2023 accounting for over USD 750 million (~2% of global funding).

As of 2024, generative AI has passed the peak of inflated expectations on the Gartner Hype Cycle. However, some experts believe LLMs will experience a second wave of innovation. As per Forrester Research, in 2024, 60% of generative AI sceptics will use the technology, whether they realise it or not. The fact remains that we are still decades away from artificial general intelligence (AGI), which is a theoretical AI system whose capabilities will rival the capabilities of humans. The question on everyone's mind is whether the costs of AI models decline enough to see a return on investment (ROI) and whether that one 'killer' application of generative AI is close on the horizon. India, however, seems poised to take advantage of this GenAI boom. Through a collaborative approach to ensure AI's positive impact while mitigating potential risks, India can leverage generative AI to take the top spot as a global tech titan.

GENERATIVE AI IS A BRANCH OF ARTIFICIAL INTELLIGENCE WHICH CAN GENERATE NEW CONTENT, INCLUDING HIGH-QUALITY TEXT, IMAGES, VIDEOS, AND MUSIC.

2.0 OVERVIEW OF GENERATIVE AI

Artificial Intelligence (AI) means intelligence exhibited by machines, particularly computer systems. Broadly, artificial intelligence is a field of research in computer science wherein machines or programs can think and learn like humans to solve problems, make decisions, or carry out tasks.

Below are some examples of high-profile applications of AI:

- Advanced web search engines: e.g., Google Search, Bing, Perplexity.
- Recommendation systems: e.g., Netflix, Amazon, YouTube.
- Virtual assistants: e.g., Siri, Alexa, Google Assistant.
- Autonomous vehicles: e.g., Waymo, Tesla.
- Machine learning and deep learning:
  - IBM's Deep Blue defeated world chess champion Garry Kasparov in 1997.
  - AlphaGo defeated a world champion Go player in 2016.
- Generative AI: e.g., ChatGPT, Meta LlaMa, Gemini (covered in this report).

Generative AI is a branch of artificial intelligence which can generate new content, including high-quality text, images, videos, and music. Generative AI attempts to mimic human intelligence in areas such as natural language processing (NLP), translation, image/ video recognition, etc.

Generative AIs are trained to learn human languages, programming languages, and even complex subjects such as chemistry, biology, and art. While artificial intelligence (AI) might help identify a problem (e.g., alerting when groceries are running low), generative AI can take it further by coming up with creative solutions, such as drafting a recipe with the existing ingredients.

In simple terms, generative AI is a computer that is very good at guessing. At its core, it is a probability engine. Generative AIs are a condensed version of millions of inputs ranging from books, newspapers, internet sources, and many varied databases. When a question is posed to a generative AI model, it breaks down the question into small pieces, and using its 'knowledge database', it returns a coherent output that best answers the question asked.

2.1. AI And Generative AI: A History

Artificial Intelligence may seem like a modern phenomenon due to the sudden popularity of ChatGPT and other generative AI; however, the roots of artificial intelligence date back centuries, whereas general intelligence, such as chatbots, has been around since the 1950s.

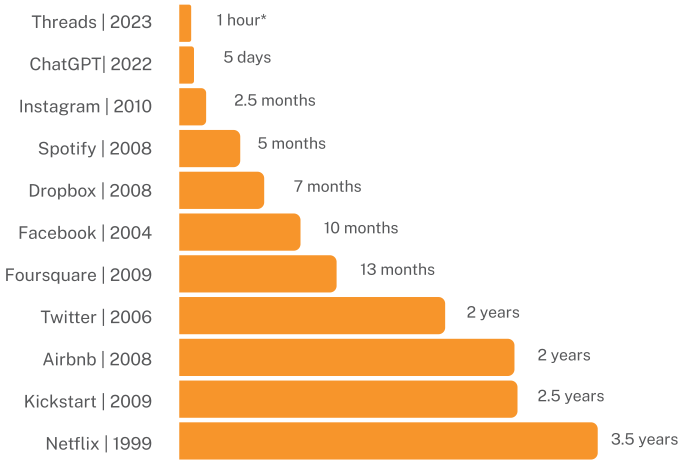
Below is a brief history of AI and generative AI:

- The concept of automated art dates back to the automata of ancient Greek civilisation, which were machines capable of writing text, generating sounds, and playing music.
- Although science fiction had familiarised the world with the concept of artificial intelligence, a paper written by Alan Turing in 1950 on 'Computing Machinery and Intelligence', made a significant contribution to the conceptual groundwork of AI.
- An early example of generative AI is a simple model known as a Markov chain, which was a statistical method introduced in 1906 to model the behaviour of random processes. In machine learning, Markov models were used for next-word prediction tasks, such as the autocomplete function within emails.
- Following the 1956 Dartmouth Summer Research Project on AI, which had an inspiring call for AI research, an early version of generative AI emerged via chatbots in the 1960s.
- ELIZA was an early natural language processing program developed in the 1960s which simulated a therapist by responding to user inputs.
- In the early 1970s, Harold Cohen was creating and exhibiting generative AI paintings created by the computer AARON.
- In the 1970s and 1980s, progress in AI research slowed down due to computational limitations and the lack of scalable algorithms, a period known as the AI Winter.
- The first neural networks were developed in the 1950s and 1960s, however, their capabilities were characterised by small data sets and a lack of computational power. Neural networks only became practical for generating content after the use of big data in the mid-2000s and following major improvements in computer hardware such as graphics processing units (GPUs) used in the gaming industry.
- In 2012, the rise of deep learning, a subset of machine learning using multi-layered neural networks, led to breakthroughs in image recognition, speech processing, and autonomous systems.
- The first practical learning models came in 2014 following advancements around deep learning models such as the Variational Autoencoder (VAE) and Generative Adversarial Networks. In 2017, the Transformer network enabled advancements in generative models, leading to the first generative pre-trained transformer (GPT), known as GPT-1, in 2018. Following GPT-1, GPT-2 was released in 2019, GPT-3 in June 2020, GPT-3.5 in November 2022, and GPT-4 in March 2023.
- The COVID-19 pandemic was a catalyst to the generative AI boom, with more companies moving to remote work models and more professionals adopting AI tools during the pandemic to increase productivity and efficiency. According to IBM's Global AI Adoption Index 2022 report, about 53% of IT professionals stated that they had accelerated their AI adoption in response to the pandemic.

2.2. The Generative AI Buzz

In the past few years, generative AI has been a persistent buzz. ChatGPT, which was built on OpenAI's GPT3.5 implementation, became an instant hit when it was released in November 2022 as a free research preview (it was earlier accessible only through an application programming interface (API)). ChatGPT has a simple, user-friendly interface which can quickly create text, graphics, or video, with a human-like conversational style, and features that allow users to interact, instruct, and fine-tune responses.

ChatGPT shoots past one million user mark in 5 days



According to OpenAI, ChatGPT surpassed 1 million users within five days of its launch – a record broken only by Instagram's Threads app in 2023 (Threads reached a million users within an hour of launch). By April 2024, ChatGPT hit a peak of nearly 2 billion visits in a month, becoming the fastest-growing application in history. According to OpenAI, as of September 2024, ChatGPT had more than 200 million weekly active users, of which over 1 million users were paid business users.

Today, a majority of digitally mature companies have an AI strategy in place. According to the 2024 McKinsey Global Survey on AI, 65% of organisations regularly use generative AI, the majority use case functions being marketing and sales, product development, and IT functions such as chatbots and customer support.

An AI agent is an AI that can plan and execute a task without human intervention, from generating, evaluating, providing feedback and rewriting code to incorporating the feedback and iterating and optimising strategies. The capabilities of current GenAI agents include mastering complex games like Minecraft, online shopping, assisting with research, etc.

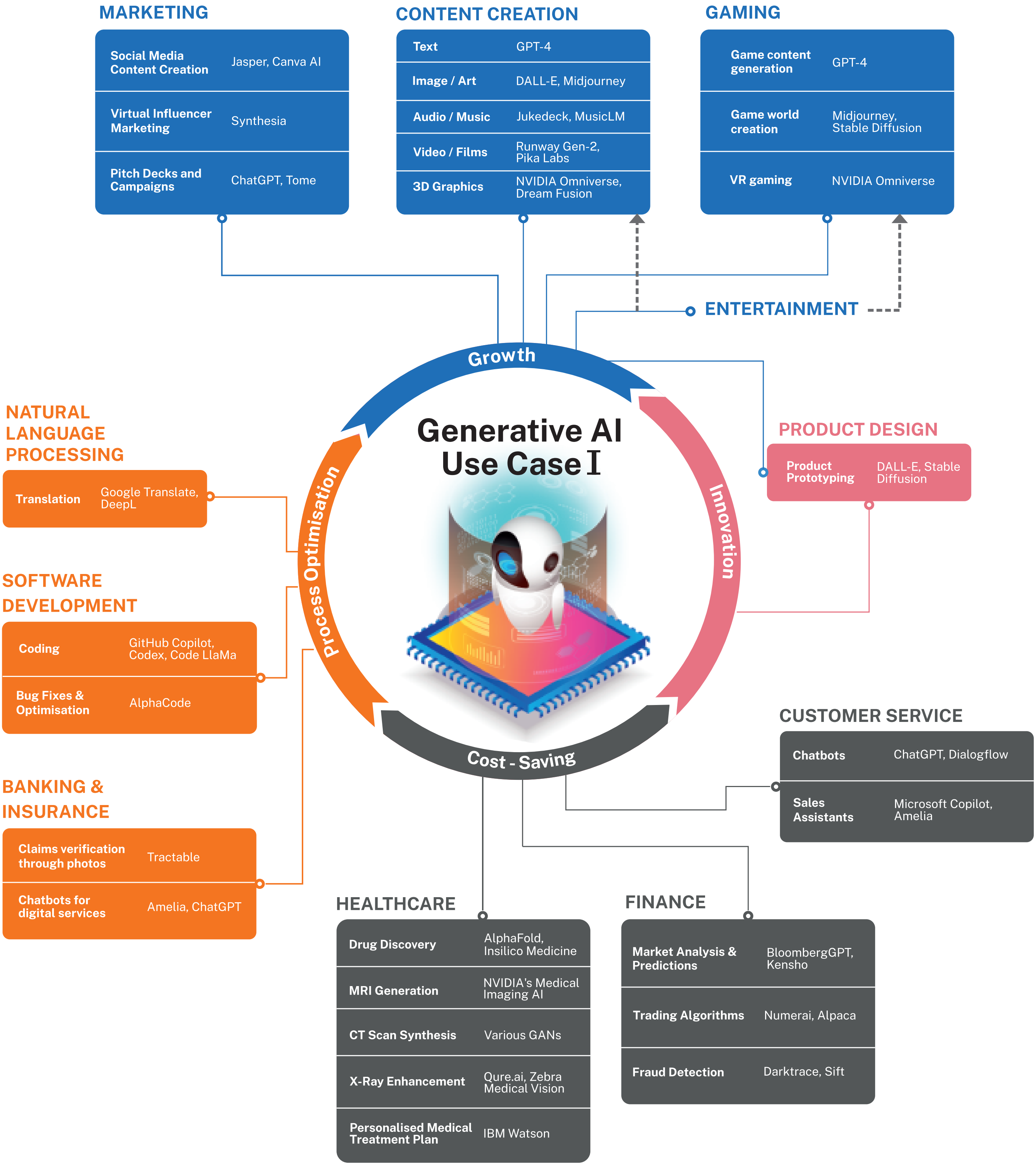
According to a 2024 survey by the Capgemini Research Institute, the pharmaceutical and healthcare sector leads in AI agent adoption (23% of organisations surveyed are already using AI agents for various functions).

It is important to note that generative AI models do not understand language but merely attempt to manipulate it and predict human-like responses. However, generative AI's ability to predict a close-to-accurate output has proved an immediate success, which is why Big Tech companies continue to invest large sums of money in further training large language models.

2.3. What Can Generative AI Do?

Large Language Models (LLMs) such as ChatGPT can generate entirely new content based on the prompt or question asked. The content is delivered across six main modalities: text, image, audio, video, code, or 3D graphics.

2.3.1. Generative AI Use Case I



Generative AI Use Case I

Growth

Innovation

Cost - Saving

Process Optimisation

MARKETING

Social Media Content Creation	Jasper, Canva AI
Virtual Influencer Marketing	Synthesia
Pitch Decks and Campaigns	ChatGPT, Tome

CONTENT CREATION

Text	GPT-4
Image / Art	DALL-E, Midjourney
Audio / Music	Jukedeck, MusicLM
Video / Films	Runway Gen-2, Pika Labs
3D Graphics	NVIDIA Omniverse, Dream Fusion

GAMING

Game content generation	GPT-4
Game world creation	Midjourney, Stable Diffusion
VR gaming	NVIDIA Omniverse

ENTERTAINMENT

PRODUCT DESIGN

Product Prototyping	DALL-E, Stable Diffusion

NATURAL LANGUAGE PROCESSING

Translation	Google Translate, DeepL

SOFTWARE DEVELOPMENT

Coding	GitHub Copilot, Codex, Code LlaMa
Bug Fixes & Optimisation	AlphaCode

BANKING & INSURANCE

Claims verification through photos	Tractable
Chatbots for digital services	Amelia, ChatGPT

CUSTOMER SERVICE

Chatbots	ChatGPT, Dialogflow
Sales Assistants	Microsoft Copilot, Amelia

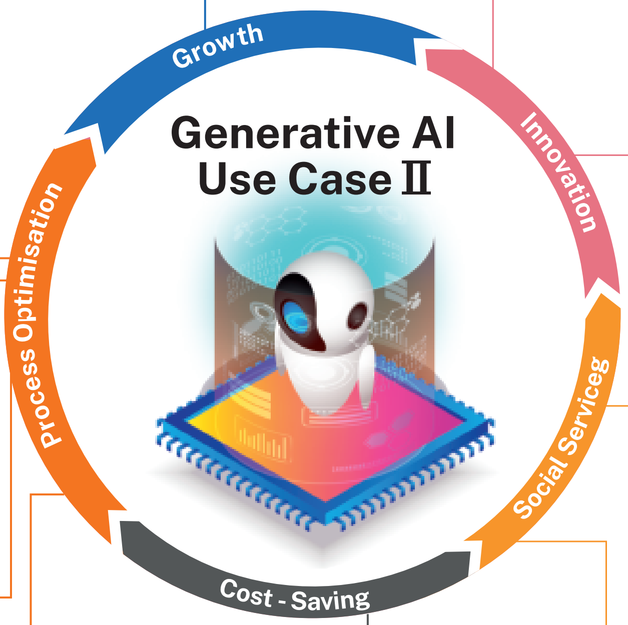
HEALTHCARE

Drug Discovery	AlphaFold, Insilico Medicine
MRI Generation	NVIDIA's Medical Imaging AI
CT Scan Synthesis	Various GANs
X-Ray Enhancement	Qure.ai, Zebra Medical Vision
Personalised Medical Treatment Plan	IBM Watson

FINANCE

Market Analysis & Predictions	BloombergGPT, Kensho
Trading Algorithms	Numerai, Alpaca
Fraud Detection	Darktrace, Sift

2.3.1. Generative AI Use Case II



AUTOMOBILES

Self-Driving Cars	Waymo, Tesla Autopilot
Cruise Control Algorithms	NVIDIA DRIVE, Mobileye
Simulation Training	Meta's AI Research, CARLA Simulator
Traffic Control	IBM Watson IoT, Siemens Mobility AI

AGRICULTURE

Precision Farming	Blue River Technology, John Deere See & Spray
Automated Irrigation Systems	CropX, Hortau
Quality Analysis	Taranis, Descartes Labs
Disease Detection in Plants	Plantix, PEAT
Pest Control	Plantix, PEAT

MINING / OIL & GAS

Seismic Data Analysis & Predictions	NVIDIA Clara, AWS IoT Greengrass
Automation in Energy Optimisation	Google DeepMind, Siemens Mindsphere

MANUFACTURING & SUPPLY CHAIN

Supply Chain Optimisation	IBM Watson Supply Chain, Microsoft Dynamics 365
Quality Control	Landing AI, Neurala
Inventory Management	SAP Leonardo, SymphonyAI

E-COMMERCE

Virtual Fitting Rooms	Zeekit, Perfitly
Personalised recommendations	Amazon Personalize, Salesforce Einstein

DATA ANALYSIS

Data mining simplification	Palantir, Alteryx

CYBERSECURITY

Threat Detection	Darktrace, CrowdStrike Falcon
Encryption Algorithms	IBM Security, Microsoft Azure Security

ENVIRONMENT

Climate Change Analysis and Simulation	Climate TRACE, Google Earth Engine

EDUCATION & LEARNING

Personalised Learning	Squirrel AI, Century Tech
Assessments and Grading	Gradescope, Cognii
Virtual Tutors and Feedback Mechanisms	Knewton, ChatGPT

FASHION

Fashion Design Generation using trend analysis	Heuritech, Stitch Fix

In addition to the use cases outlined above, generative AI can be used across virtually every sector. Below are examples of some industries that have yet to fully adopt generative AI and the potential of generative AI in those fields:

- **Financial and Tax Auditing:**
  - The Institute of Chartered Accountants of India (ICAI) is adopting a small language model called CA GPT by leveraging OpenAI's ChatGPT. CA GPT is being developed to provide tailored support for chartered accountants and students.
  - Generative AI can help CA firms in numerous ways, including, for example:
    - Automating certain tasks such as drafting financial reports and generating audit summaries.
    - Preparing tax strategies based on the company's historical data.
    - Simulating fraud scenarios to test the financial controls of a company.
    - Creating detailed explanations of tax laws for clients.
    - Assisting with regulatory compliances by creating guidelines, scheduling tasks for real-time compliance, etc.

- **Construction:**
  Generative AI can create detailed building layouts or construction plans based on specified parameters. It can also generate 3D models and simulations to help design processes.

- **Energy:**
  Generative AI can help design efficient grids or simulate renewable energy systems.

- **Government and Administration:**
  Generative AI can create simulations for urban planning and crisis management and even craft strategies. It can also help with administrative processes by drafting policy documents or budget proposals.

2.3.2. Generative AI: Success Stories

Some GenAI success stories and accomplishments are outlined below:

Medicine and Healthcare

- SkinVision, an app for early detection of skin cancer, teaches users to examine and assess risks.

- Nvidia, collaborating with Hippocratic AI, has developed a Gen AI-powered healthcare agent that claims to surpass human nurses in patient interaction on video calls, offering empathetic conversations at a fraction of the cost. These AI agents can conduct various healthcare tasks, including appointment scheduling, pre-operative outreach, and post-discharge follow-ups.

Marketing and Branding

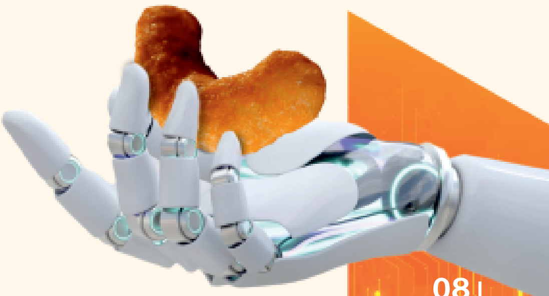
- PepsiCo utilised generative AI to analyse customer feedback, which it then used to refine the shape design and flavour of "Cheetos", thereby boosting market penetration by 15%.

Environment

- BrainBox AI's ARIA, which leverages Amazon's generative AI model, focuses on reducing the carbon footprint of commercial buildings. The AI can predict interior building temperatures and claims to reduce heat, ventilation and air-conditioning (HVAC) energy costs in commercial buildings by up to 25% and greenhouse gas (GHG) emissions by 40%.

Product Design

- Insilico Medicine has identified a new drug candidate to treat idiopathic pulmonary fibrosis using its generative AI platform. The drug candidate is entering phase 2 of clinical trials using one-tenth of the cost and one-third of the time that would have been taken using traditional drug discovery methods.



Coding

- Devin is an AI-driven tool which is a fully autonomous software engineer. The tool acts as a collaborative coding partner where users can control its level of autonomy while it continuously self-learns from GitHub repositories.

Customer Service

- Klarna, a Swedish FinTech firm, has employed generative AI for customer service. The firm claims that the AI assistant can handle a workload equivalent to that of nearly 700 employees, and within a month of deployment, there has been a 25% decrease in repeat calls and a reduction in average handling time from 11 minutes to 2 minutes.

- IndiGo has deployed a customised AI chatbot, 6Eskai, which leverages GPT-4 technology to help users with booking tickets, promotional discounts, seat selection, planning trips, etc.

Business Analytics

Lexi, an AI chatbot created by a Bangalore-based start-up, can provide insights and business reports to entrepreneurs relating to market spend tracking, sales, etc.

Agriculture

- India's KissanGPT, which leverages ChatGPT technology, can guide farmers on issues relating to irrigation, pest control, crop cultivation, and more.

Cybersecurity

- LG's Smart Home AI Agent is an AI robot that can manage smart-home devices without human oversight. The robot patrols the home, monitors pets, carries out household chores and interacts with users.

- Torq's cybersecurity analysis AI agent, Torq Socrates, helps with automating contextual alert triaging, incident investigation, and response, thereby enabling security staff to focus on other priority matters. According to the company, employing this technology will resolve 90% of tier-1 and tier-2 tickets independently.

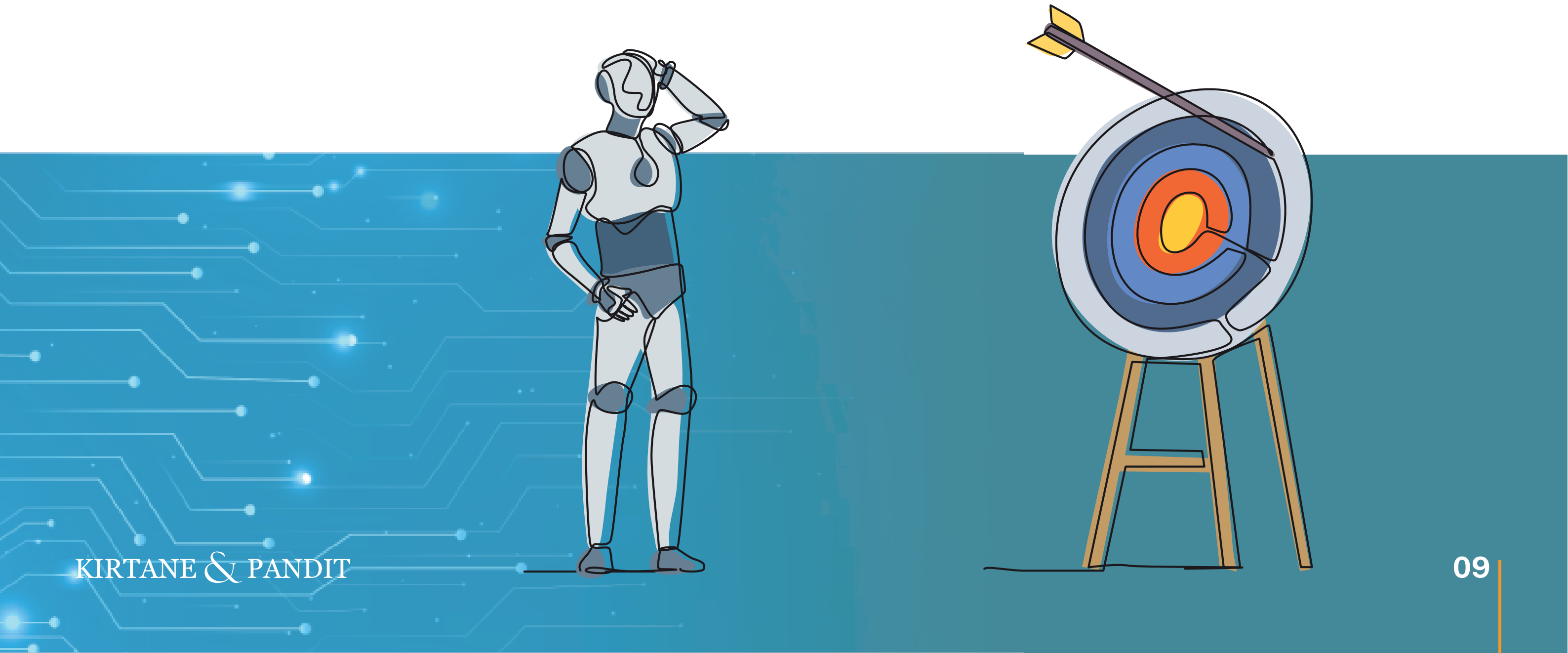
Film

- According to a Gartner article, by 2030, a major blockbuster film will be released, with 90% of the film being AI-generated (from text to video).

2.4. Where Generative AI Misses The Mark

At present, Generative AI cannot fact-check information. A common disclaimer for popular GenAI models such as ChatGPT or Gemini is that the model may hallucinate, that is, it may confidently provide an inaccurate response or pass off made-up information as fact. Further, GenAIs are not inherently creative when it comes to generating content, art, or music.

GenAIs only generate output based on the datasets they are trained with. This also means that there is a cut-off date beyond which a GenAI may not be able to provide information on recent news or articles. For example, OpenAI's latest model, GPT-4o, is a multimodal model with 128 K context and an October 2023 knowledge cut-off.



Some GenAI snags and faux pas are outlined below:

- **Translation:** In 2017, a mistranslation of the Arabic 'good morning' by Facebook's AI-powered translation service led to the arrest of a Palestinian man in Israel.

- **AI vs Human:** In July 2020, a blog written by a college student using GPT-3 went viral and landed the top spot on Hacker News, receiving thousands of subscribers. The fact that a generative AI had written all the blog posts went unnoticed for months. Similarly, in November 2023, a few Sports Illustrated articles were found to have been written by AI generated authors.

- **Voice-Activated Ordering:** In July 2024, McDonalds ended a 3-year test run with IBM where it was leveraging AI to automate orders in over 100 outlets, following several blunders and customer frustration. A viral TikTok video featured two customers attempting to correct their order while the AI kept adding Chicken McNuggets to the order until reaching 260.

- **Legal Arguments:** A legal attorney used ChatGPT to research precedents in a legal suit. The judge found that 6 of the cases submitted in the brief were made up and never existed, and included false names, docket numbers, internal citations, and even quotes.

- **Unethical Advice:** In March 2024, it was reported that a Microsoft-powered chatbot MyCity, which provided information to New Yorkers on starting and operating businesses in the city, was giving incorrect advice which would lead users to break the law.

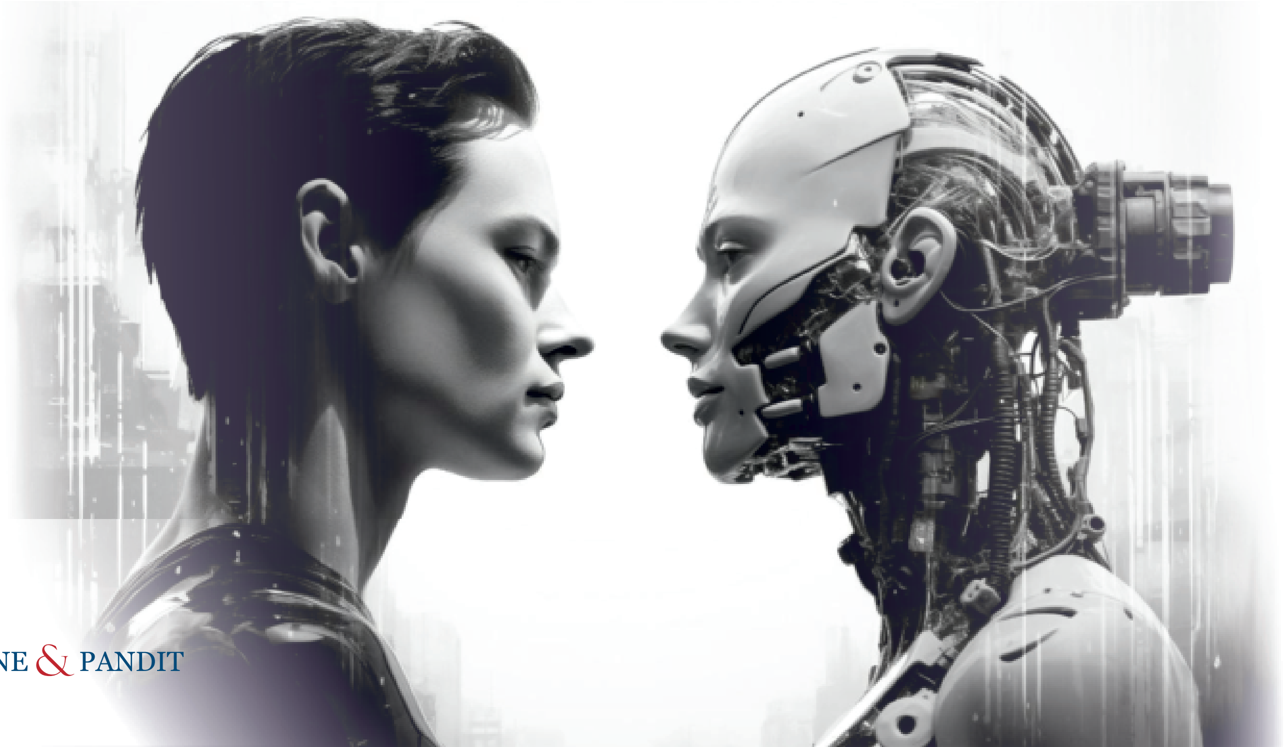
- **Recruitment:** In 2023, an AI-powered recruiting software rejected over 200 applicants applying age filters, namely, female applicants above the age of 55 and male applicants above the age of 60.

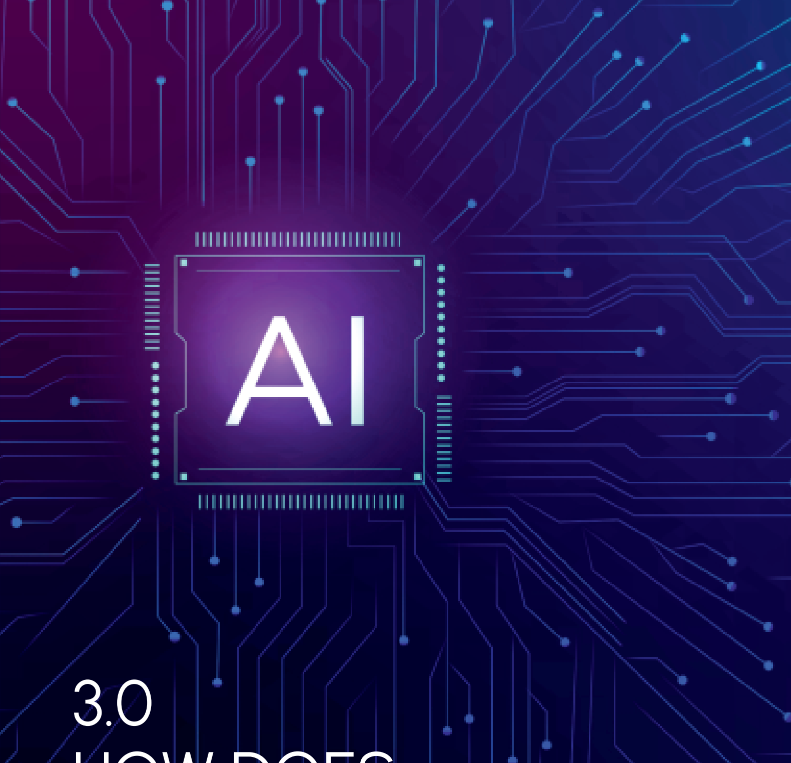
- **Copyrights:** Several authors, including George R.R. Martin and John Grisham, have filed a suit against OpenAI for using their copyrighted material to train the models. The suit alleges that ChatGPT was trained on pirated e-books and, therefore, was able to generate summaries and even unofficial prequels and sequels for the Game of Thrones books. In a similar vein, in October 2024, over 25,000 artists (from fields of music, art, literature, film, and theatre) released a public statement warning that AI training on copyrighted works poses a 'major, unjust threat' to their livelihoods.

- **Deep Fakes:** AIs are increasingly being misused to create 'deepfakes', i.e., images, videos, or audio recordings that are digitally altered to replace the original person with someone else in a way that makes it look authentic. In July 2024, a video went viral that used an AI voice cloning tool to convincingly mimic the voice of US Vice President Kamala Harris.

A free and open-source project called the AI Incident Database has been developed, which tracks and indexes in real-time any harm or potential harm caused by the use of generative AI. The AI Incident Database reported 123 incidents in 2023 (a 32.3% rise from those in 2022). A recommendation to build a similar database for the Indian landscape has been proposed by a panel set up by the Indian government.

It is important to note that the above limitations (covered in detail in the following sections) only indicate that generative AI is still an emerging technology, however, the applications and technology are increasing at an unprecedented speed. For example, by May 2023, the generative AI Claude was able to process 100,000 tokens of text (~75,000 words) in a minute, as compared to 9,000 tokens when it was introduced in March 2023.





SIMILAR TO ARTIFICIAL INTELLIGENCE (AI), GENERATIVE AI WORKS BY USING MACHINE LEARNING MODELS AND DEEP LEARNING

3.0 HOW DOES GENERATIVE AI WORK?

3.1. Understanding the Generative AI Architecture

Similar to artificial intelligence (AI), generative AI works by using machine learning models and deep learning. Machine learning models are very large models that are pre-trained on vast amounts of data. On the other hand, a deep learning network is fed only raw data, from which the software learns independently by analysing unstructured datasets like text documents, identifying priority data attributes, and solving complex problems.

Open AI's generative pre-trained transformer (GPT) models are Large Language Models (LLMs), which are neural network-based language prediction models. LLMs are able to perform multiple tasks because they are trained with a large number of parameters (for e.g. GPT-1 was trained with 117 million parameters as discussed below).

In 2017, the transformer architecture, which was developed by a Google team for translation purposes, served as a breakthrough in the field of natural language processing. Transformers use self-attention mechanisms and positional encoding to achieve parallel computation over the entire text:

- Self-attention mechanisms' allow models to learn dependencies between words, regardless of the distance between the text.
- In a process called 'tokenisation', the input text is split into smaller tokens (tokens are pieces of words where 1,000 tokens equal about 750 words).
- 'Positional encoding' provides information about the order of tokens and word order.
- After self-attention, a 'feed-forward neural network' further processes the information.
- Multiple transformer 'stacks' are used to apply these mechanisms iteratively.

Large Language Models leverage either the encoder, decoder, or both modules of the transformer. For example, GPT-3.5 uses a stack of 13 transformer blocks plus additional components such as layer normalisation, residual connections, and positional embeddings.

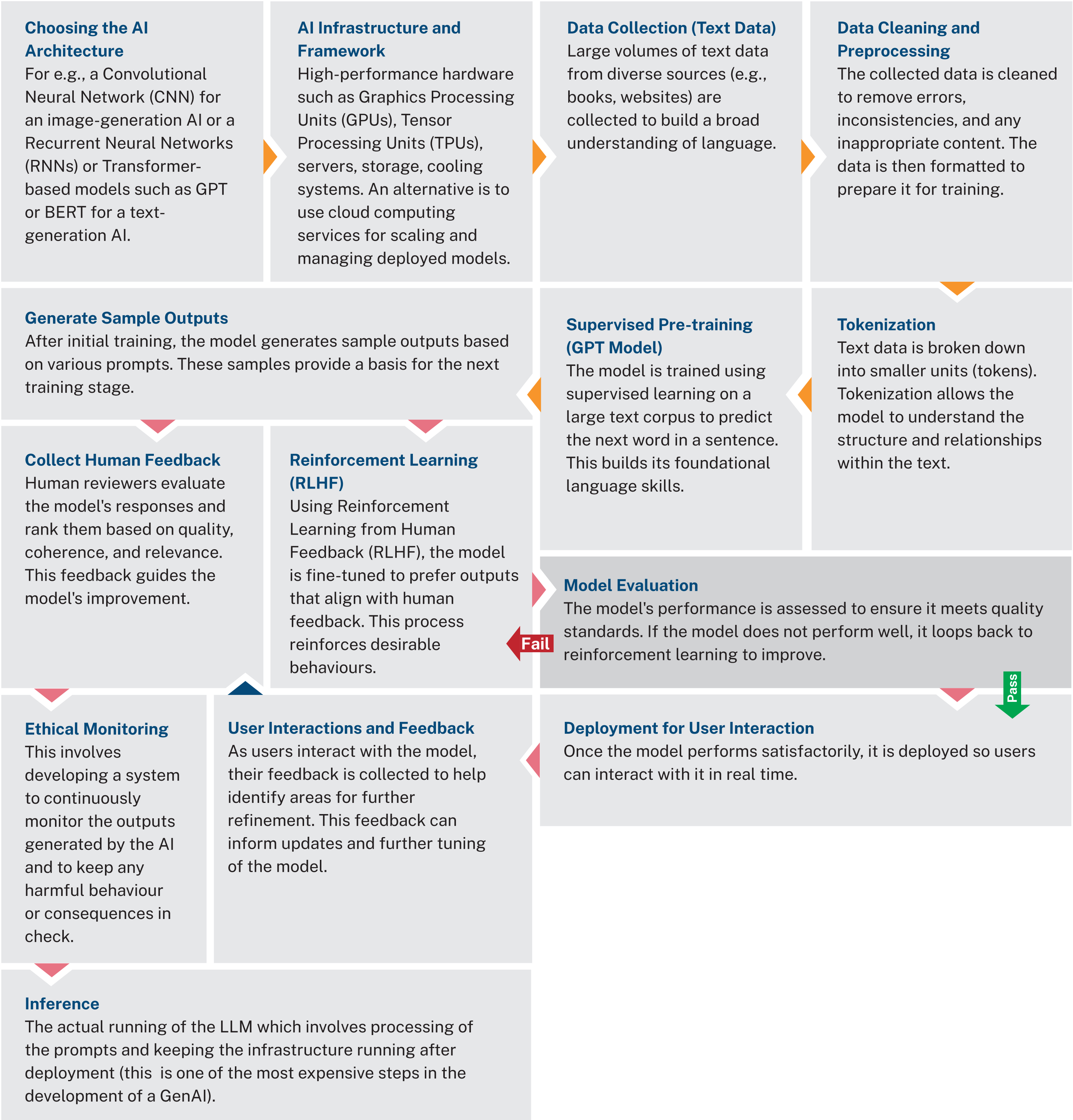
3.2. A Comparison of Popular Generative AI Models

Company / Model	Google – BERT (Bidirectional Encoder Representations from Transformers)	OpenAI – Generative Pre-Trained Transformer (GPT-1 to GPT-4)	Meta - LLaMA 3 (Large Language Model Meta AI)	Google – PaLM (Pathways Language Model)	Anthropic – Claude	Falcon	CodeT5
Transformer Architecture	Encoder only	Decoder stacks	Decoder only	Decoder only	Decoder only	Decoder only	Encoder-Decoder
Features	BERT focuses on understanding the context of each word by looking at the entire sentence from left-right and right-left.	GPT models generate text by predicting the next word in a sequence (a left-to-right model).	Llama is a compact and efficient model which is generally more accessible than other LLMs and, thus, easier to fine-tune.	PaLM was designed as a pathways-based model which is great at multitasking and offers over 100 languages.	Claude is a model designed with a strong focus on responsible AI usage and ethical alignment.	Falcon is an open-source large language model developed by UAE's Technology Innovation Institute. Falcon is scalable and allows users to fine-tune its capabilities for specialised domains.	CodeT5 is an open-source model designed by Salesforce specifically for coding tasks.
Strengths	Tasks requiring a deep understanding of the text, such as answering questions, translating languages, improving search engine results, etc.	Tasks requiring the creation of text, such as writing articles or generating a detailed and coherent response.	Tasks best suited for smaller enterprises, such as research-focused language processing.	Multimodal tasks such as coding, translation, and logical reasoning.	Trained to avoid harmful responses and is sensitive during dialogue and conversations.	Tasks related to customisable natural language processing and research.	Tasks that need code-to-text or text-to-code transformations.
Limitations	Not designed for continuous text generation.	Less suited for tasks that require an encoder-based approach, such as classification of sentences or answering questions.	• May require more fine-tuning or training for specialised domains. • Resource intensive. • Lacks emotional intelligence.	• Resource intensive. • Less advanced than GPT-4	Less optimised for language processing tasks than GPT-4.	Limited multimodality for inputs such as image or video-based prompts.	Less effective for real-time code autocompletion than Codex or Copilot.

3.3. Process of Developing a Gen AI

Developing a model from scratch is expensive and requires huge R&D and in-house developers, machine learning engineers, data scientists, and domain specialists. An alternative is to utilise an existing open-source model or license a pre-trained model.

The steps involved in developing a generative AI model are outlined below:



Choosing the AI Architecture
For e.g., a Convolutional Neural Network (CNN) for an image-generation AI or a Recurrent Neural Networks (RNNs) or Transformer-based models such as GPT or BERT for a text-generation AI.

AI Infrastructure and Framework
High-performance hardware such as Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), servers, storage, cooling systems. An alternative is to use cloud computing services for scaling and managing deployed models.

Data Collection (Text Data)
Large volumes of text data from diverse sources (e.g., books, websites) are collected to build a broad understanding of language.

Data Cleaning and Preprocessing
The collected data is cleaned to remove errors, inconsistencies, and any inappropriate content. The data is then formatted to prepare it for training.

Tokenization
Text data is broken down into smaller units (tokens). Tokenization allows the model to understand the structure and relationships within the text.

Supervised Pre-training (GPT Model)
The model is trained using supervised learning on a large text corpus to predict the next word in a sentence. This builds its foundational language skills.

Generate Sample Outputs
After initial training, the model generates sample outputs based on various prompts. These samples provide a basis for the next training stage.

Collect Human Feedback
Human reviewers evaluate the model's responses and rank them based on quality, coherence, and relevance. This feedback guides the model's improvement.

Reinforcement Learning (RLHF)
Using Reinforcement Learning from Human Feedback (RLHF), the model is fine-tuned to prefer outputs that align with human feedback. This process reinforces desirable behaviours.

Model Evaluation
The model's performance is assessed to ensure it meets quality standards. If the model does not perform well, it loops back to reinforcement learning to improve.

Deployment for User Interaction
Once the model performs satisfactorily, it is deployed so users can interact with it in real time.

User Interactions and Feedback
As users interact with the model, their feedback is collected to help identify areas for further refinement. This feedback can inform updates and further tuning of the model.

Ethical Monitoring
This involves developing a system to continuously monitor the outputs generated by the AI and to keep any harmful behaviour or consequences in check.

Inference
The actual running of the LLM which involves processing of the prompts and keeping the infrastructure running after deployment (this is one of the most expensive steps in the development of a GenAI).

SIMILAR TO ARTIFICIAL INTELLIGENCE (AI), GENERATIVE AI WORKS BY USING MACHINE LEARNING MODELS AND DEEP LEARNING

4.0 THE GENERATIVE **AI MARKET**

According to Markets and Markets, a market research firm, the global market size of generative AI is expected to increase from USD 20.9 billion in 2024 to USD 136.7 billion by 2030, registering a compound annual growth rate (CAGR) of 36.7% between 2024 and 2030.

A 2023 McKinsey report, 'The Economic Potential of Generative AI', gives the following key insights:

- Generative AI could add the equivalent of USD 2.6 trillion to USD 4.4 trillion annually across 63 use cases which were analysed by McKinsey, which include marketing and sales, customer operations, software engineering, product R&D, supply chain, risk, and legal.
- About 75% of the value that generative AI use cases could deliver falls across four areas: customer operations, marketing and sales, software engineering, and R&D.
- The biggest sectors predicted to be impacted by generative AI are banking, high-tech, and life sciences.
    - Across the banking industry, the technology could deliver value equal to an additional USD 200-340 billion annually.
    - In the retail and consumer packaged goods sector, the potential impact is calculated at USD 400-660 billion annually.
- The present capabilities of generative AI technologies have the potential to automate work activities that absorb 60-70% of employees' time.

Generative AI could raise global GDP by 7% (USD 7 trillion) and increase productivity by 1.5% by 2033, according to Goldman Sachs Research.

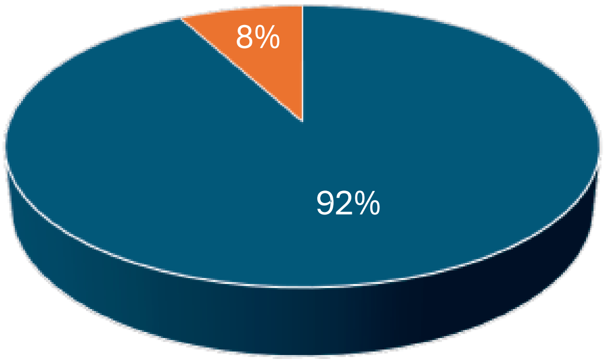
According to Gartner, an American research firm:

- More than 10% of all data will be AI-generated by 2025.
- By 2025, 30% of major marketing messages and over 30% of new drugs and materials to be systematically discovered will be done using generative AI (up from less than 2% and 0% in 2023, respectively).
- By 2026, more than 80% of enterprises will have used generative AI APIs or models in production environments.

4.1. Components of the Generative AI Market

The GenAI market can be broken down into three components: Infrastructure, Foundation Models and Platforms, and Applications.

Data GPUs - Share In GenAI Market in 2023



■ NVIDIA ■ Others (AMD, Intel, Google, Apple)

1. Infrastructure

As of 2024, GenAI infrastructure is the most mature component of the market. There are three types of players in this stack:

1.1 GPUs and Chip Providers
(NVIDIA, AMD, Intel, Google, Apple)

- These companies provide high-performance GPUs that are designed for machine learning, training, and inference.
- GPU companies are also entering into commitments or partnerships with foundation model providers to capitalise early, for e.g. Google and Anthropic.

The NVIDIA Success Story
NVIDIA, a company that primarily designed graphics processing units (GPUs) and chips for gaming, catapulted to success during the generative AI boom. The situation is almost like a gold rush, with all the tech companies racing to hit gold, and in the process, it is the shovel-maker (or NVIDIA in the AI context) that gets rich. The AI revolution revealed NVIDIA's GPUs to be essential for model training and inference. NVIDIA now caters to almost all players in the generative AI market, including OpenAI, Microsoft, and Amazon.

AMD and Intel are among NVIDIA's top competitors. However, many other companies are entering the AI chip market, including:

- Amazon's Graviton4 and Trainium2 reduce the power needed to train AI models.
- Microsoft's Maia 100 AI chip and the Cobalt 100 Arm chip will power its subscription software offerings as part of its Azure cloud computing service.
- Google's new cloud tensor processing unit, TPU v5p, can train a large language model like GPT3-175B 2.8 times faster than the TPU v4, according to Google.
- Newer entrants like Cerebras and SambaNova are developing radical chip architecture that is more efficient for generative AI.

1.2 Cloud Service Providers
(Amazon, Baidu, Google, Microsoft)

- Cloud service companies provide scalable resources for training and deploying generative AI models.
- Amazon Web Services (AWS) currently leads the generative AI market in cloud computing.

1.3 GenAI Service Providers (Amazon, Cohere)

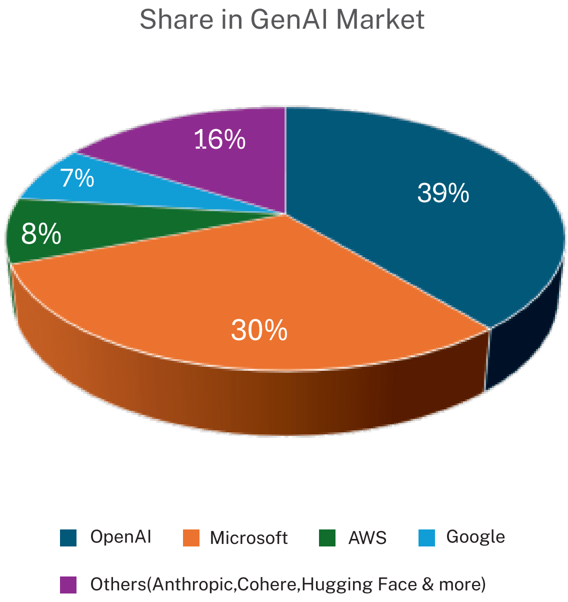
- These companies provide generative AI models for specialised applications such as custom model training and fine-tuning, API integration, developer support, etc.

Despite the emerging competition, by 2023, NVIDIA shares had climbed nearly 300%. In February 2024, NVIDIA's market capitalisation was at USD 3 trillion, and in June 2024, NVIDIA surpassed even Microsoft in valuation.

2. Foundation Models and Platforms

Foundation models are large pre-trained models that are used for tasks such as language processing, image recognition, etc. Generative AI platforms include software that manages generative AI-related activities such as data management, GPU as-a-service, user interface (UI/UX), etc.

- **Foundation Models**
  - Meta, OpenAI's GPT 1 to GPT 4
  - Model Hubs (GitHub, Hugging Face)
- **Platforms** (Google, Hugging Face, Microsoft)
  - Microsoft also promotes the usage of other models via its platform, for e.g. Meta's Llama 2.
  - Amazon Web Services (AWS) provides access to several models, such as Anthropic, AI21 labs, and Cohere, on its platform.



OpenAI and Microsoft (which is also a substantial shareholder in OpenAI) together accounted for 69% of the GenAI foundation models and platforms market in 2023, according to a research report by IoT Analytics. Further, the report projects the generative AI foundation models and platforms market to account for nearly 5% of global software spending by 2030.

3. GenAI Applications

These stand-alone generative AI applications serve as a gateway between foundation models and end users.

- (Canva, Lensa).
- Plugins (Grammarly, AI Art).

4.2. Key Trends in the GenAI Market

- **Classification by Modality**
  - The text segment dominated the GenAI market in 2024 due to its applicability across virtually every industry.
  - The code segment is expected to be the fastest growing region in the next 5 years. According to estimates, only 1% of the global population can code. With generative AI, coding knowledge is no longer an obstacle to creating and developing applications.

- **Classification by Industry / Sector**
  - According to Fortune Business Insights, the healthcare segment held a significant market share in generative AI (23%) in 2023.
  - The BFSI segment also saw increasing adoption of generative AI, particularly for financial services, risk analysis, fraud detection, and customer services.
  - As per a 2024 survey by KPMG, significant use of generative AI includes industrial markets for inventory management (64%), healthcare and life sciences for healthcare document assessment (57%) and media and telecommunications for workflow automation (43%).

- **Classification by Geographical Region**
  - The major activity around generative AI and most of the funding raised by generative AI companies are US-based companies. In the third quarter of 2024, US-based AI start-ups attracted USD 11.4 billion in investment, accounting for over 65% of global AI funding.
  - As per the Stanford AI Index 2024, in 2023, 61 notable AI models originated from US-based institutions as compared to the EU's 21 and China's 15.
  - However, when it comes to filing AI patents, China is leading. In 2022, 61.1% of AI patents originated in China, as compared to 20.9% originating from the US.
  - According to a NASSCOM report on India's Generative AI Start-up Landscape 2024, across the top six global economies leading GenAI adoption, the US ranked first, followed by the UK, EU, Japan, Israel, and India. India registered 2x growth in share of GenAI start-ups, and an overall 4th rank among the top invested economies.

4.3. Funding and Money Flows

In 2024, the GenAI market is already flooded. Major enterprises, including Big Tech companies, are competing heavily to cash in on the AI boom.

In 2022, venture capital firms invested more than USD 2 billion in the generative AI market. Significant investments were also made in GenAI models, for e.g., Microsoft invested USD 10 billion in OpenAI, and Google bought a USD 300 million stake in Anthropic.

The total funding for generative AI further soared in 2023 to USD 25.2 billion (almost 9 times more than 2022 according to the Stanford AI Index 2024). Major players reported fundraising rounds in 2023, including OpenAI, Anthropic, and Hugging Face (Cohere raised USD 270 million in 2023). In 2023, Microsoft announced the integration of GPT-4 into Office 365, and Adobe introduced many generative AI features in Photoshop.

As per the Stanford AI Index 2024, the number of newly funded generative AI companies increased to 99 start-ups in 2023 as compared to 56 in 2022.

According to a NASSCOM report on India's Generative AI Start-up Landscape 2024, 77% of global GenAI funding in 2024 was into companies building foundation models, 90% of which was invested in just three start-ups – OpenAI, Anthropic, and Inflection.

90% of the global funding in 2024 in GenAI foundation models was invested in just three start-ups – OpenAI, Anthropic, and Inflection.

Some GenAI investment highlights of 2024 include:

- OpenAI raised USD 6.6 billion in a funding round in October 2024, which pushed its valuation to USD 157 billion.
- Anthropic raised the second-largest sum among US GenAI start-ups with USD 7.8 billion, followed by Elon Musk's xAI with USD 6 billion.
- According to VC firm Accel, out of the USD 79.2 billion raised by cloud firms based in Europe, the US, and Israel in 2023-24, 40% of all VC funding went to generative AI start-ups.
- Microsoft, Amazon, Google, and Oracle have announced investments in nuclear energy to offset their energy consumption on computational power required by generative AI (discussed below).

According to research by EY Ireland, global VC investment in generative AI is forecasted to reach USD 12 billion by the end of 2024.

GENERATIVE AI, ALTHOUGH OFFERING SUBSTANTIAL BENEFITS, ALSO HAS UNIQUE RISKS AND LIMITATIONS

5.0 RISKS AND **LIMITATIONS**

Generative AI, although offering substantial benefits, also has unique risks and limitations, which, if not checked, can pose serious threats or harm to personal and corporate data as well as society and the environment.

The following section outlines the risks and unique limitations of the new generative AI technology.

Security and Data Confidentiality

Generative AI can pose a threat to confidential corporate or personal data, for example, generating harmful code or drafting convincing phishing emails. Further, such outputs may become difficult to predict or control, leading to unprecedented harm in business or real-world applications.

The unauthorised use of generative AI in the workplace is known as shadow AI and is a growing security concern. Employees may use generative AI to bypass corporate standards without comprehending security risks, which may lead to a leak of sensitive trade secrets or the use of copyrighted material that could impact the company legally.

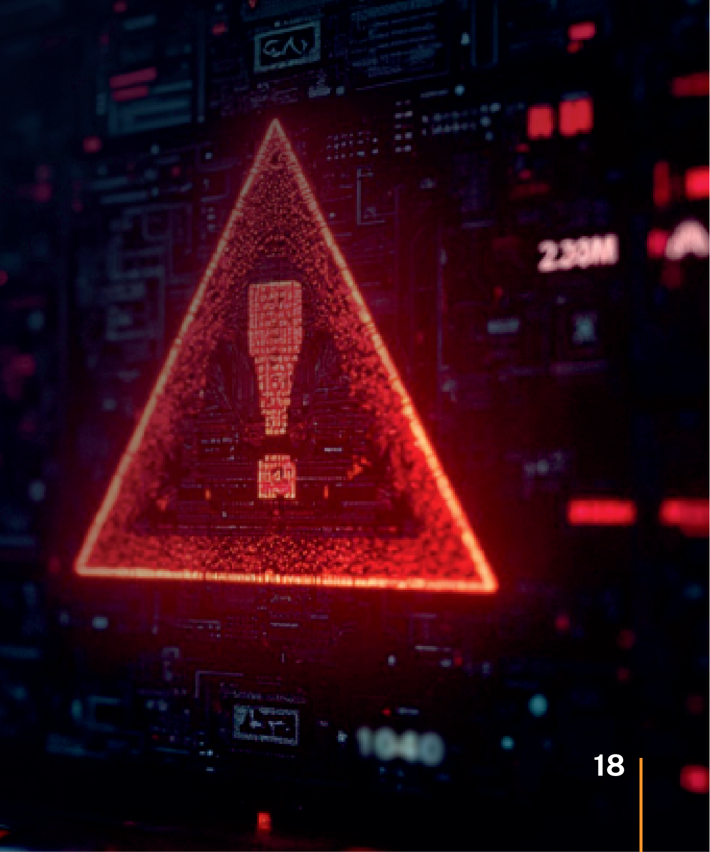
Misinformation and Bias

Generative AI has the potential to produce realistic but false content, including text, images, and videos, i.e., deepfakes. These have been increasingly used to manipulate public perception, spread misinformation, and impact elections or social movements.

Further, since generative AI models learn from large datasets, often derived from internet sources, they can produce and reflect societal, political, or discriminatory biases, which can lead to generative AI producing responses that reflect harmful stereotypes, which becomes difficult to check.

Transparency

Foundation Model Transparency shows that AI developers rank low on transparency with respect to the disclosure of training data and methodologies. This lack of transparency can undermine trust and make it challenging to assess the ethical and environmental impact of these technologies.



Environmental Impact of Generative AI

Training large generative AI models leaves a substantial environmental footprint due to the energy demands of high-performance hardware. Companies such as OpenAI, Google, Anthropic, and Mistral refrain from disclosing carbon emissions when training their models.

However, the carbon footprint and power consumptions estimated for different foundation models as per Stanford AI Index 2024 are as follows:

GPT-3 (OpenAI)

Carbon Emissions during training:
Approximately 502 tonnes of CO_2.

Power Consumption:
Estimated around 1.2 GWh.

Equivalent Activities

- Carbon released by 1,260 one-way flights from New York to London.
- Annual emissions of ~50 average American households.
- Power consumption: Similar to the annual electricity usage of ~110 average American homes.
- Power consumption: Comparable to running 1-2 small data centers for a year.

LLaMA 2 (Meta, 70B model)

Carbon Emissions during training:
Estimated at 291.2 tonnes of Co_2.

Power Consumption:
Estimated at 0.7 GWh.

Equivalent Activities

- Emissions of ~291 one-way flights from New York to San Francisco.
- Equivalent to ~16 average Americans' annual emissions.
- Power consumption: Approximately the annual electricity usage of ~65 average American households.
- Power consumption: Comparable to running a medium-sized office building for nearly a year.

BERT (smaller NLP model for comparison)

Carbon Emissions during training:
Estimated at 72 tonnes of Co_2.

Power Consumption:
Estimated at 200 MWh.

Equivalent Activities

- Carbon released by ~72 one-way flights from New York to Los Angeles.
- Power consumption: Comparable to the annual electricity usage of ~20 American homes.

*The above numbers are estimates and may vary as per location, data center, and energy efficiency.

According to Bain analysts, the rise of generative AI will lead companies to develop gigawatt-scale data centres, consuming 5 to 20 times the power of current facilities and potentially straining both the electricity grid and the labour supply.

However, there is also an upside, since generative AI is also being applied in sustainability efforts, such as optimising thermal energy systems, managing pest control, and improving urban air quality.

Ethical Concerns

Within a few years of its development, generative AI technology is already facing multifold ethical concerns. The use of copyrighted material for AI training (discussed above) has resulted in numerous lawsuits against OpenAI. GenAIs have also been manipulated for data harvesting (such as the Cambridge Analytica scandal). Such misuse highlights the ethical dangers of using personal data without consent, making the monitoring of data handling practices an immense priority in the GenAI Space.

There have also been challenges to user safety owing to the unexpected behaviour of a generative AI:

- ChatGPT once initiated conversations without any user prompts, which OpenAI attributed to a software bug.
- An AI-driven application, Character.AI, which allows users to create chatbot versions of themselves, resulted in harm to the mental health of a teenager and ended in suicide.

The potential for unintended and harmful consequences necessitates continuous oversight and ethical consideration in the development and deployment of generative AI tools.

Public Backlash

The use of copyrighted materials for training GenAIs, as well as the emergence of applications such as AI Art, have sparked public backlash on whether art 'created' by generative AI models (based on true art created by humans) should be displayed publicly and acclaimed. In addition to the lawsuit by authors including George R.R. Martin and John Grisham, 25,000 artists (from fields of music, art, literature, film, and theatre) released a public statement warning that AI training on copyrighted works poses a 'major, unjust threat' to their livelihoods.

In March 2023, over 35,000 prominent technologists, including Steve Wozniak, Elon Musk, and the CEOs of several AI and other companies, added their signatures to an open letter written by the Future of Life Institute, calling on AI labs to pause experiments on AI technologies for at least 6 months. The pause was called for so that AI developers could work on making AI systems accurate, safe, interpretable, transparent, robust, aligned, trustworthy, and loyal, as well as to work with policymakers to accelerate the development of robust AI governance systems.

The open letter followed a 2023 survey of AI experts wherein 36% expressed the fear that AI development may result in a 'nuclear-level catastrophe'.

Loss of Jobs

Generative AI could impact up to 300 million jobs globally.

According to a 2023 World Economic Forum report, 23% of jobs today will be impacted by generative AI.
As per 2023 research by Goldman Sachs, generative AI has the potential to automate around 18% of the total workforce. Further, generative AI could substitute up to 25% of current work in the US and 24% in Europe while complementing most of the remaining work.

The key sectors and occupations that will be impacted are as follows:

- The major industries to be impacted by automation in jobs include banking, insurance, software and platforms, and capital markets.
- The tasks that are most likely to be automated include administrative support and clerical tasks, record-keeping and stock-taking and related tasks.
- The jobs that are least likely to be impacted by generative AI include cleaning, repair, and maintenance, installation, construction, and extraction.

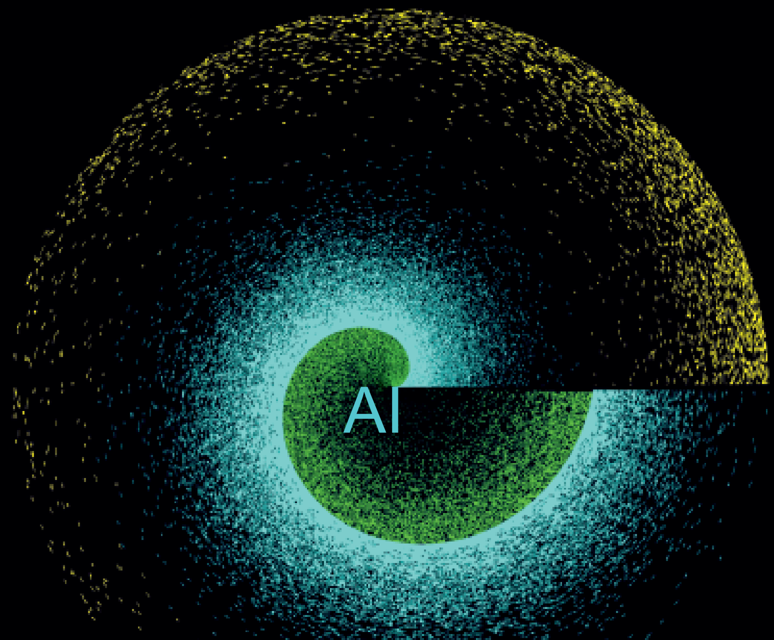
However, there is a silver lining as well, namely, that generative AI has the potential to create jobs. As per the World Economic Forum 2023 report, the number of AI and machine learning specialists will also increase by 40% by 2027, thereby creating around 1 million new jobs across all industry verticals. Further, roles such as data analysts, data scientists, big data specialists and information security analysts will also experience a surge in demand of 30-35% and 31% respectively, generating an additional 2.6 million jobs by 2027, which brings us to the following limitation of generative AI.

Shortage of Skilled Workforce

With generative AI, enterprises are recognising an urgent need to upskill the workforce. The fastest growing jobs, as per the World Economic Forum 2023 report, include AI and machine learning specialists, sustainability specialists, business intelligence analysts, fintech engineers, data analysts and scientists, robotics engineers, electrotechnology engineers, and digital transformation specialists.

Generative AI has the potential to create an additional 2.6 million jobs.

As per Microsoft's 2023 Work Trend Index, 82% of leaders recognise the pressing need to prepare their workforce for the expanding AI landscape. Additionally, organisations like Amazon, Ericsson, and PwC are heavily investing in upskilling and reskilling initiatives to build employees' AI capabilities. In a recent survey, 40% of CEOs cite a lack of AI-related knowledge and skills within their HR team as the biggest obstacle to AI integration.
In India specifically, according to a NASSCOM 2023 report, there is an additional demand for around 1 million professionals in India's AI space (more on this in the following section).



AS OF 2024, AS PER GARTNER, GENERATIVE AI HAS PASSED THE PEAK OF INFLATED EXPECTATIONS

6.0 GenAI HYPE CYCLE AND INNOVATION

6.1 GenAI Hype Cycle

According to August 2024 research by RAND Corporation, 80% of all AI projects end in failure, which is twice the failure rate of projects that do not involve AI. As per Gartner research, nearly one-third of generative AI projects are likely to be abandoned after the proof-of-concept stage, owing to rising costs and unclear business value. Gartner also predicts that by 2028, over 50% of enterprises that have built AI models from scratch will abandon their efforts.

Gartner, an American research firm, creates hype cycles that map the adoption of new technologies.

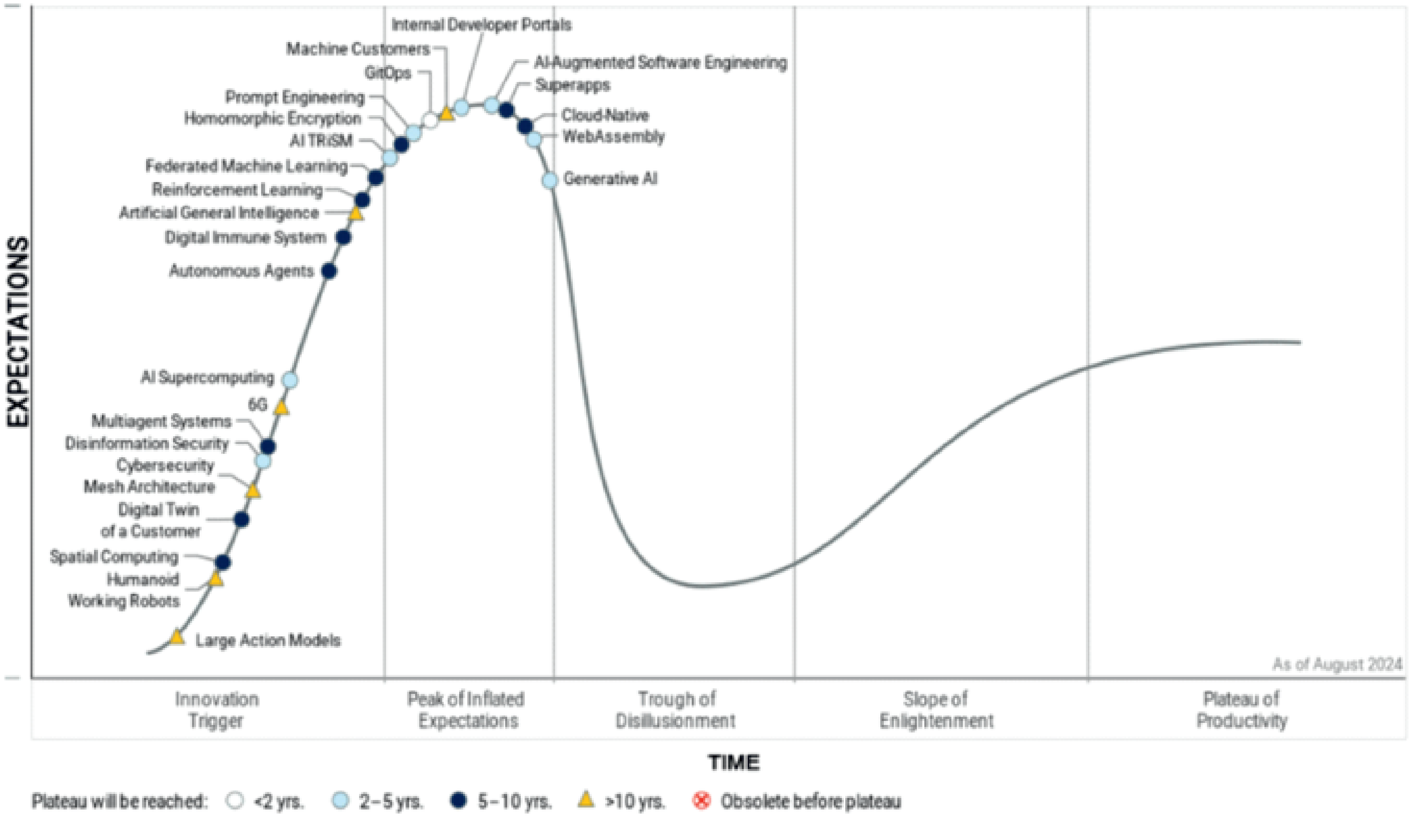
The hype cycle has five phases: the innovation trigger, the peak of inflated expectations, the trough of disillusionment, the slope of enlightenment, and finally, the plateau of productivity. It generally takes between 3-5 years for a new technological innovation to progress through the five stages.

As of 2024, as per Gartner, generative AI has passed the peak of inflated expectations.

The hype around generative AI continued strong from 2022 to 2024, mainly because humans tend to reach optimistic conclusions about a technology that can fluently answer questions and mimic human-like language. Case in point – the hype around 5G technology from 2017 until the 2020s eventually resulted in a slow uptake, considering its adoption challenges and market uncertainties.

With generative AI, however, its unique limitations make it difficult to judge the potential of this new technology.

Studies have shown that generative AI models, while displaying high intelligence on certain benchmarks (e.g. complex medical exams), can fail simple tasks, are unable to solve complex math problems, and can fail to predict results as expected by humans. The generative AI market is also bogged with difficulties such as high investment requirements, a lack of skilled human talent, and a profitability dilemma for generative AI business models (discussed in the following section).



Gartner

On the other hand, some experts believe that generative AI will experience multiple small peaks of expectations and innovations. Other experts suggest that primary technologies such as Large Language Models (LLMs) usually experience a second wave of innovation that involves developing new applications and changing organisational structures to adapt to the technology. As per Forrester, in 2024, 60% of generative AI sceptics will use the technology whether they realise it or not.

6.2. Innovation Around Generative AI

The tremendous pace at which generative AI is developing highlights continuous advancements in AI technologies such as natural language processing (NLP) and generative adversarial networks (GANs).

Some key innovations, applications, and structural developments happening around generative AI include:

Upskilling

A study by economist David Autor found that 60% of today's workforce is employed in occupations that didn't exist in 1940, implying that over 85% of employment growth over the last 80 years is explained by the technology-driven creation of new positions. The generative AI revolution has already caused a surge in demand for skilled data scientists, software engineers, and developers with AI and Machine Learning (ML) skills.

As per Gartner, Generative AI will require 80% of the engineering workforce to upskill by 2027.

Multimodal AI

Multimodal AI models can process more diverse data inputs. For example, generative AI models can use 24/7 camera feeds, analyse them and even learn from that data for better understanding and inference.

Smart Robots

Language processing technology provided by generative AI fused with advancements in robotics has allowed the development of efficient and flexible robotic systems. Smart robots like PaLM-E and RT-2 can ask questions and interact with their environment. PaLM-E and RT-2 variants have an 80% success rate even for unseen objects in their vicinity.

Retrieval-Augmented Generation (RAG)

RAG is a process that allows entities to integrate specific data from internal and external sources, enabling the generative AI to provide relevant and more accurate outputs. For example, using Retrieval-Augmented Generation (RAG) in customer service chatbots produces information that is more reliable and personalised based on company data.

Model Optimisation

Entities have begun focusing on developing smaller, more efficient models (for example, Microsoft's Phi and Apple's MM1) than models that are bigger with respect to parameter count (which are therefore also resource-intensive and costly). Some recent technologies to optimise generative AI models include Low-Rank Adaptation (LoRA), which reduces the number of parameters and speeds up fine-tuning, and Reinforcement Learning from Human Feedback (RLHF), which aligns model outputs to human preferences (see the section on GPT training above).

Fixes for GPU shortages and Cloud Computation

With many GenAI models flooding the market, there is a global shortage of data GPUs and AI chips, and the costs for cloud computing are also rising. Innovation is paramount in the field of high-performance GPUs and alternative computational resources. Companies are designing AI accelerators, such as the NVIDIA Hopper and the AMD MI300, which are designed for next-generation AI workloads. Further, companies such as Google, Amazon, and Microsoft have announced their own AI chips (see the above section for details).

6.3. The Generative AI Value Chain

According to a 2024 McKinsey report, the most significant opportunity for new entrants in the generative AI value chain over the next 3-5 years is building end-user applications.

As per the 2024 McKinsey report, the opportunity size (from highest value to lowest) for new entrants in the generative AI value chain is as follows:

- Services around specialised knowledge on how to leverage generative AI (such as training, feedback, and reinforcement learning).
- End-user applications such as B2B or B2C products that leverage foundation models or fine-tune the models for specific use cases.
- Model hubs that provide tools to curate, host, fine-tune, or manage foundation models.
- Building and deploying core foundation models on which GenAI applications can be built.
- Cloud platforms to provide access to computing hardware.
- Data GPUs and AI accelerator chips optimised for training and inference.

New entrants also need to consider the AI profitability dilemma and its value-cost analysis (discussed in the following sections).

Some researchers believe that large language models are only the first step in the generative AI boom, the destination being artificial general intelligence (AGI), which is a theoretical AI system whose capabilities will rival the capabilities of humans. AGI may be able to replicate human-like cognitive abilities, including reasoning, problem-solving, perception, and learning. For example, AGI will be able to understand and comprehend language and not simply mimic it. According to scholars, we are decades away from AGI – a step where the abilities of an AI are indistinguishable from those of a human (the Turing Test).

74% OF ENTERPRISES (END USERS) USING GENERATIVE AI ARE CURRENTLY WITNESSING A RETURN ON INVESTMENT (ROI)

7.0

ECONOMICS OF GEN AI: A VALUE-COST ANALYSIS

According to a survey by Google Cloud, 74% of enterprises (end users) using generative AI are currently witnessing a return on investment (ROI), with an additional 30-35% anticipating ROI within the next 12 months. Further, 86% of organisations that implemented generative AI saw their revenue increase by more than 6%.

In this section, we analyse the estimated costs for training and running a generative AI model and the profitability dilemma for this new technology. It is important to note that most figures and profitability analyses are based on estimates because exact costs and returns on investment are not disclosed by major players in the market.

Training Costs

Companies like OpenAI and Google do not disclose the precise costs of training AI models. OpenAI CEO Sam Altman estimated that GPT -4 cost over USD 100 million to train. Anthropic CEO Dario Amodei has suggested that by 2025, there may be a generative AI model that costs USD 10 billion to train.

According to the Stanford AI Index 2024, below is a comparison of training costs for different generative AI models:

GenAI Model	Year of Release	Company	Training Cost (USD)
Original Transformer	2017	Google	900
GPT-3	2020	OpenAI	4.6 million*
RoBERTa Large	2021	Meta	0.16 million
PaLM	2022	Google	12.4 million
GPT-4	2023	OpenAI	78 million
Gemini Ultra	2023	Google	191 million
LLaMA	2023	Meta	2.4 million

* Theoretical estimate by cloud provider Lamdba considering a Tesla V100 cloud instance.

Considering these high training costs, most models are not trained continuously and have a knowledge cut-off date. As per the CEO of AI start-up Hugging Face, the process of re-training the second version of the Bloom LLM cost around USD 10,000.

Inference Costs

Inference means the actual running of the Large Language Model (LLM). Training and inference of generative AI models takes place on data GPUs. The primary AI chip made by NVIDIA, which dominates GPUs in the GenAI market, costs USD 10,000.

As per estimates, the inference costs for ChatGPT, which reached 100 million users in January 2023, were around USD 40 million for that month. Estimates suggest that Microsoft's Bing requires around USD 4 billion of infrastructure as running costs.

Other Costs

As per Gartner, organisations can spend between USD 5 to 20 million for tuning, customising, or simply deploying a generative AI model or application.

The cost-effectiveness of developing a generative AI model depends on the specific requirements of the project, the resources available, and the growth plan. The specific factors that drive up costs are as follows:

- **Scale and Complexity:** Large-scale projects are typically costlier, with small-scale projects being relatively cost-effective.
- **Data Quality:** High-quality, labelled data is essential for training effective models but is costly to acquire and prepare.
- **Hardware Requirements:** High-performance hardware, such as data GPUs, TPUs, and servers, are expensive to purchase or rent.
- **Expertise:** A specialised workforce, including data scientists, machine learning engineers, and AI researchers with expertise in generative AI, is expensive to hire.
- **Development Time:** Developing generative AI models often involves experimentation and refinement, extending development time and increasing costs.

Pricing of GenAI Models

Presently, OpenAI's GPT-4o is priced at USD 2.50 per 1 million input tokens and USD 10 per 1 million output tokens, where tokens are pieces of words, and 1,000 tokens equal about 750 words. This pricing is after OpenAI slashed prices as compared to its previous version (GPT-4o is 50% cheaper for input tokens and 33% cheaper for output tokens).

OpenAI's pricing approach ensures that users are charged based on actual usage and depending on the input and output prompts/responses given.

A comparison of pricing by popular Gen AI Models is as follows:

GenAI Model	Company	Pricing	
		Per input token (USD)	Per output token (USD)
GPT-4o	OpenAI	2.5/1M	10/1M0.
Claude	Anthropic	0.008/1K	024/1K1
Gemini 1.5 Pro LLM	Google	3.5/1M	0.5/1M

OpenAI also offers a monthly pricing, which is currently set at USD 20 per month.

GenAI Profitability and Value-Cost Analysis

According to the New York Times, OpenAI expects revenues of USD 3.7 billion in 2024. However, OpenAI has forecasted a USD 5 billion loss (excluding equity compensation) in 2024, thereafter projecting revenues to more than triple in 2025 to USD 11.6 billion and eventually USD 100 billion in 2029. The New York Times further reported that OpenAI plans to raise the monthly subscription price to USD 22 by the end of 2024 and will eventually raise it to USD 44 per month over the next five years. Considering the costs and pricing of models and hardware, generative AI can be cost-effective in the following scenarios:

- Leveraging pre-trained models can reduce development time and costs.
- Using cloud-based platforms can save upfront capital investments; however, these can prove expensive for scaling.
- Enterprises may opt for customised generative AI models for their core use cases to reduce costs and achieve productivity gains.

When it comes to cloud computing, the GenAI market is witnessing a Catch-22, where most of the generative AI cost of compute, such as training and fine-tuning models to storing data and scaling, is done via cloud computing, while at the same time, cloud services are being used for building, training, and running of foundational AI models. This itself is driving up cloud costs (enterprise cloud costs rose an average of 30% in 2023 because of AI, according to a report by Tangoe, a provider of expense and asset management solutions).

Considering the high costs and the soft and hard returns on investment, an enterprise needs to be strategic about investing in generative AI or implementing it in its organisation. Careful analysis needs to be made regarding the skill level of the workforce and whether the productivity gains and time freed up from automation will translate into commercial success for the enterprise.

AS PER THE BCG-NASSCOM REPORT 2024, THE AI MARKET IN INDIA IS WITNESSING ROBUST GROWTH

8.0 INDIA'S GEN AI **LANDSCAPE**

As per the BCG-NASSCOM report 2024, the AI market in India is witnessing robust growth, projected at a Compound Annual Growth Rate (CAGR) of 25-35%.

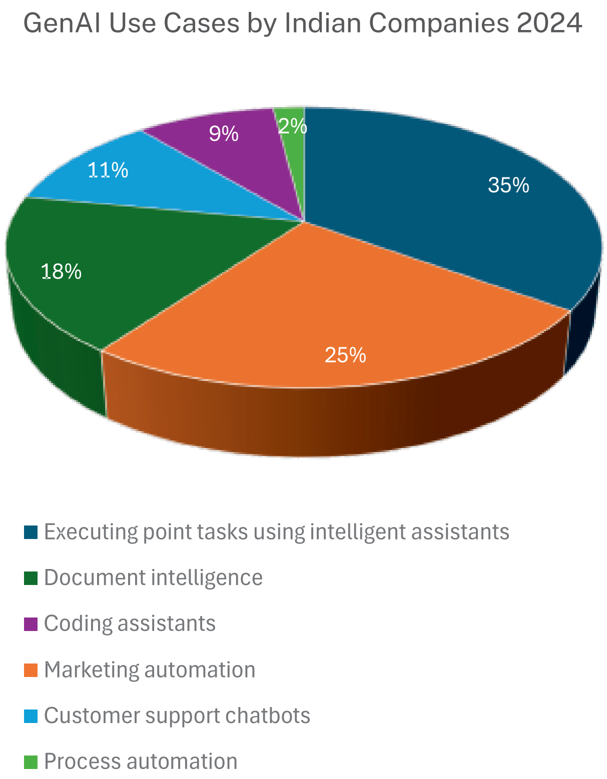
As per the International Data Corporation, a market intelligence firm that provides insights into information technology, telecommunications, and consumer technology markets, India's spending on AI and GenAI, including software, services, and hardware for AI-centric systems, is projected to reach USD 6 billion by 2027, registering a compound annual growth rate (CAGR) of 33.7% for the period 2022-2027.

As per the NASSCOM report on India's Generative AI Start-up Landscape 2024, GenAI funding in India is limited to the early stages of seed and series A rounds, with no growth or late-stage activity. However, in the first half of 2024, cumulative funding in Indian GenAI start-ups grew by 1.5% as compared to the same period in 2023 and accounted for almost 2% of global funding since 2023.

8.1. India's Generative AI Market

India's GenAI space is also witnessing policies and initiatives by three separate groups: large tech companies and business groups, initiatives by the Indian Government, and start-ups developing domestic GPU infrastructure.

8.1.1. GenAI Adoption by Indian Companies



According to a May 2024 report by EY India, only 15-20% of proof-of-concepts (POCs) by Indian companies have been rolled out into production. Among Global Capability Centers (GCCs), this number is between 30% and 40%.

As cost dynamics in the generative AI market evolve rapidly, Indian companies are being cautious in their approach to generative AI. The deployment of LLM applications requires a huge cost with respect to cloud consumption and pay-as-you-go pricing of most models. Even the discounted capacity option for cloud services requires a minimum monthly spend. When used at scale, the costs tend to rise after considering the costs of training and inference (discussed in the above sections), language translation, fine-tuning, etc. In India, there is a lack of domestic availability of a pay-as-you-go commercial computing structure. These are some of the reasons why many POCs have not made it past the boardroom. Further, according to a NASSCOM report on India's Generative AI Start-up Landscape 2024, nearly 80% of start-ups reported earning less than USD 100K.

For now, Indian enterprises may find it more commercially viable to opt for commercial GenAI platforms and an appropriate selection of LLMs based on core use cases.

8.1.2. Initiatives by the Indian Government

The Indian government has introduced the IndiaAI Mission which has a funding of INR 10,300 crore (USD 1.25 billion) for AI projects over a period of five years. A major goal of the IndiaAI Mission is to build a public AI cloud infrastructure through a public-private partnership encompassing over 10,000 GPUs. The INR 10,300 crore funding for the mission will be allocated to diverse objectives, including:

- Development of indigenous LLMs and domain-specific foundational models (the government has released Bhashini, which provides for real-time translation of text and audio in Indian languages).
- Public AI cloud infrastructure through a public-private partnership encompassing a minimum of 10,000 GPUs for access to start-ups.
- Research ecosystem for generative AI within the next five years.
- An application development initiative.
- Start-up financing.
- Safe and responsible AI.

The IndiaAI Datasets Platform, which is part of the IndiaAI Mission, is expected to be live by January 2025. The platform will host datasets and create a collaborative space for developers to train and deploy their own AI models.

The government also plans to establish a National Data Management Office to coordinate with departments and ministries to improve the quality of data and make them available for AI development and deployment. These investments aim to foster the creation of AI applications for the public sector.

8.1.3. Indian Start-ups and Investments in the GenAI Ecosystem

According to a NASSCOM report on India's Generative AI Start-up Landscape 2024, the total number of Indian GenAI start-ups increased to over 240 in the first half of 2024 (meaning 174 new start-ups in one year). 80% of these start-ups provide GenAI assistants. The cumulative funding in these start-ups has been over USD 750 million since 2023. Further, out of 3,600 deep-tech Indian start-ups, 74% focus on AI-driven solutions.

According to EY India's analysis of the top 50 Indian unicorns, approximately 66% are already using AI or GenAI technology. Many GenAI start-ups in India are building functional and domain-specific enterprise applications. For e.g., Ola Krutrim became India's first AI unicorn in January 2024. Krutrim AI, a voice-based AI assistant, offers dynamic ride pricing and can forecast demand and traffic.

AWS recently picked seven Indian start-ups for its Generative AI Accelerator, representing a milestone for India's AI start-ups. These start-ups will gain up to USD 1 million in AWS credits, along with mentorship and exposure. Google has also invested in Indian AI start-ups via their start-up accelerator program.

Most Indian AI companies use international platforms and cloud providers such as AWS, Microsoft Azure, and Google Cloud. For example, Infosys is a leading user of GitHub Copilot and has also partnered with Microsoft to boost the adoption of generative AI and Microsoft's cloud platform, Azure.

However, some Indian start-ups are working to provide domestic alternatives for scalable storage options. For example, Yotta Data Services, one of the bidders for the IndiaAI Mission's GPU tender, has already deployed the first phase with 4,096 GPUs.

With NVIDIA recently investing in India's GenAI space, India's ecosystem is booming. Below are some key partnerships and investments:

- Yotta Data Services and NVIDIA to boost AI adoption: In March 2024, Yotta Data Services received 4000 NVIDIA H100 Tensor core GPUs. Yotta plans to take this number to 32,768 units by the end of 2025.

- NVIDIA partnership with Reliance Industries and Tata Group: This partnership will focus on building GPU infrastructure on top of the GH200 Grace Hopper Superchip.

- NVIDIA's partnership with Yotta Data Services, L&T Technology Services (LTTS) and Zoho Corporation at the NVIDIA summit in India:

  - Yotta launched six AI services on its Shakti Cloud platform, incorporating NVIDIA NIM microservices.

  - LTTS introduced an AI Experience Zone at its Bengaluru hub, focusing on mobility and healthcare. LTTS plans to train over 1,000 engineers on NVIDIA software.

  - Zoho plans to use NVIDIA's platform to enhance its software-as-a-service (SaaS) offerings and invest over USD 10 million.

- Major Indian IT firms like Infosys, TCS and Wipro are also planning to deploy AI solutions using Nvidia's platform.

- Nvidia is also collaborating with various Indian firms to develop AI technologies, including a Hindi language model.

8.2. India's AI Workforce

In 2022, software developers in India contributed 24.2% of GitHub AI projects. The next most represented areas were the European Union and the United Kingdom (17.3%), and then the United States (14.0%).

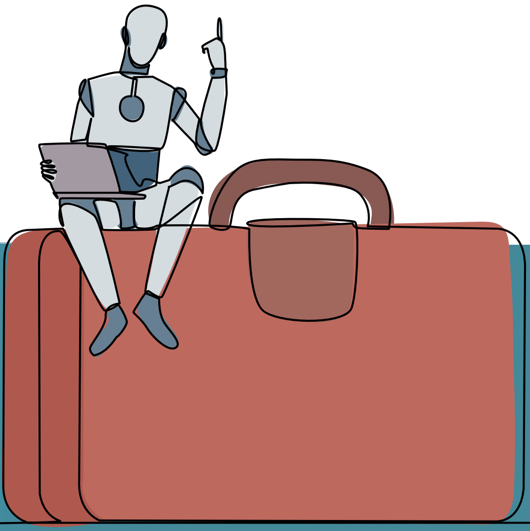
India is expected to surpass the US in the number of AI developers on GitHub by 2028.

The Stanford AI Index 2024 highlights that India tops global AI skill penetration rates and AI talent concentration. India's AI skill penetration is rated at 2.8, surpassing the United States (2.2) and Germany (1.9). Additionally, India's AI talent concentration has seen significant growth, with a 263% increase from 2016 onwards.

According to a NASSCOM 2023 report on India Data Science & AI Skills, AI applications constitute more than 50% of the software segment's market size. The surge in demand for Data Science & AI professionals has showcased India as a principal sourcing destination, with one of the largest qualified talent pool of technical graduates in the world.

However, the NASSCOM report also estimates that by 2026, India will need more than a million data science & AI professionals. In order to stay ahead of the curve, Indian companies need to start upskilling their workforce in AI-related skills. Tata Consultancy Services trained over 300,000 of its employees in Gen AI skills, while Infosys trained 250,000 employees by the end of 2023-24, according to a media report. Wipro trained over 220,000 employees in GenAI skills.

Although India displays a strong ecosystem when it comes to generative AI adoption, it still needs a robust framework and regulation on responsible AI, without which this technology may be very difficult to keep in check. The global regulatory framework on AI is discussed in the following section.



NEED FOR A PROPER REGULATORY FRAMEWORK ACROSS THE WORLD THAT ENSURES THE RESPONSIBLE USE OF GENERATIVE AI TECHNOLOGY.

9.0 POLICY AND REGULATION ON **GEN AI**

Considering the risks and the potential benefits explained in the above sections, there is a dire need for a proper regulatory framework across the world that ensures the responsible use of generative AI technology. Below is an outline of attempts around the world to develop regulations and policies around generative AI.

European Union

The European Union's draft AI Act focuses on preventing harm to individuals and safeguarding human rights. The Act classifies AI tools based on risk level. High-risk applications, such as biometric surveillance, are subject to strict transparency and operational requirements.

The Act also provides for stringent reporting standards for high-risk AI systems and monitoring to ensure that the systems are ethically sound and do not perpetuate bias.

United States

In the US, over 180 bills have been proposed relating to AI and related technologies across federal and state levels. The recent AI Bill of Rights focuses on responsible innovation that prioritises public well-being, human rights, and economic stability. The framework urges that AI products be rigorously vetted for safety before public deployment.

The US has also developed the AI Risk Management Framework, which outlines operational standards for AI developers and a roadmap for the National AI Research Resource to facilitate responsible innovation.

While there is no federal AI law yet, states are implementing their own AI regulations. For example, California's laws require transparency for automated decision-making in certain sectors.

China

China's draft regulations demonstrate a strong focus on government control over the domestic internet and tech space and monitoring of content. China's regulations on deep synthesis technologies prohibit generative AI from producing content that endangers national security, violates public interests, or promotes illegal activities. For example, Baidu's generative AI filters politically sensitive content.

China's AI policy provides for audits of AI-generated content to prevent the dissemination of politically sensitive or false information. Further, China's personal information protection laws must be adhered to when training AI models.

China has also introduced a law regulating private companies' usage of online algorithms for consumer marketing.

United Kingdom

The United Kingdom's regulatory framework, as set out in a White Paper, emphasises five key principles for AI regulation: safety, transparency, fairness, oversight, and contestability. The paper recognises a de-centralised principle-based regulatory framework, including inconsistent enforcement or guidance across regulators.

Canada

Canada has issued the Artificial Intelligence and Data Act (AIDA), which is already part of its digital law. AIDA's framework focuses on AI systems with high societal impact and recommends human oversight, transparency, and accountability. The Act intends to mitigate risks by prohibiting reckless or malicious uses of AI, attempting to balance innovation with ethical considerations.

India

At present, India does not have a specific law addressing generative AI. However, the government has set up a panel to submit recommendations for the introduction of a specialised AI regulatory framework.

The Indian government, in March 2024, issued an advisory instructing large platforms to obtain permission from the Ministry of Electronics and Information Technology (MeitY) before implementing any untested AI platforms deployed in the Indian market. The government has also issued other advisories, including requiring social media platforms to comply with existing rules focused on misinformation powered by AI and deepfakes.

As part of the India AI Mission (discussed above), the government has also selected 8 educational institutes to enhance ethical AI development.

In addition to the acts, draft laws and regulations outlined above, legal experts are also calling for a 'universal copyright standard' in response to the debate on whether copyrighted materials may be used to train AI models. If this becomes a reality, generative AI models will be required to pay a licensing fee to train their models on creative work that is not in the public domain.

AS GENERATIVE AI BECOMES INTEGRATED INTO THE REAL WORLD THERE IS AN URGENT NEED FOR RESPONSIBLE AI PRACTICES

10.0 RESPONSIBLE AI AND THE WAY FORWARD

10.1. Responsible AI

As generative AI becomes integrated into the real world, with implications on personal life, corporates, industry, education, and public services, there is an urgent need for responsible AI practices. Considering the harmful risks discussed above in the report, as well as the huge productivity and potential benefits of the technology, countries around the world are attempting to prioritise transparency, accountability, fairness, and environmental sustainability while harnessing this new technology.

India's Responsible AI Regulatory Framework

As discussed above, governments, including India, are establishing guidelines and regulations to govern generative AI. India's Ministry of Electronics and Information Technology (MeitY) is considering establishing an AI regulatory body and has set up a panel to analyse the impact and opportunities of generative AI. The panel has suggested auditing AI systems as one of the tools put forward to mitigate generative AI risk.

A regulatory framework for India can be developed with inputs from different stakeholders in AI governance:

Policy Development (MeitY)
Setting standards and guidelines.

Implementation by Tech Firms
Alignment with guidelines.

Compliance by Regulatory Bodies
Monitoring compliance

Feedback & Iteration
Continuous feedback and refinement of regulations.

Responsible AI In Decision-Making

Experts in decision modelling and data analytics have defined a four-tier framework for including AI in decision-making processes:

- Human-in-the-loop (HITL): Partially automates some aspects of the decision-making process but has a person actively reviewing the data and making the ultimate decision.
- Human-in-the-loop-for-exceptions (HITLFE): AI is trusted to make many decisions on its own but brings in a human operator for exceptions.
- Human-on-the-loop (HOTL): Allows the AI to make decisions but will refer to people in order to inform how it makes its next decisions.
- Human-out-of-the-loop (HOOTL): AI acts autonomously but with constraints.

When it comes to bias and discrimination while training AI models, a recommended approach is using culturally and linguistically relevant data sets. This approach can especially work in India, whose landscape calls for diverse, regionally adapted AI models.

Sustainable AI

The training of generative AI models is highly resource-intensive and leaves a substantial carbon footprint.

Efforts are underway on a global scale to make generative AI more energy-efficient:

- Google has announced investments in renewable energy and carbon offset programs for its data centres, including those powering AI training.
- Microsoft has announced plans to be carbon-negative by 2030.
- Meta and OpenAI have expressed goals to improve AI's energy efficiency, although there is limited data on specific reductions achieved.

Considering India's commitments under the Paris Agreement, the development of AI in India must align with environmentally responsible practices. India's government is already recommending energy-efficient practices for the development of GPUs as per the IndiaAI Mission. Indian companies, as well as public-private partnerships, can drive research on sustainable AI practices and thus position India as a leader in green and responsible generative AI development.

10.2. Way Forward

Despite the revolutionary boom, generative AI is still seeing a few hurdles, such as:

- Over-flooding of the market with GenAI models and platforms.
- Shortage of data GPUs.
- Power grid shortage.
- GenAI's capabilities fall short of those of humans, for example, in solving complex math problems.
- Commercial viability, especially for newer entrants.

The GenAI market may have passed the peak of expectations, but we are still decades away from artificial general intelligence (AGI). Businesses are still cautious to adopt the technology, primarily due to high cost or lack of skilled personnel. The question on everyone's mind is whether the costs of AI models decline enough to see a return on investment (ROI) and whether that one 'killer' application of generative AI is close on the horizon.

India, however, is poised to take advantage of this GenAI boom, with the government itself funding the IndiaAI Mission and promising to build over 10,000 data GPUs for domestic start-ups. India's generative AI space requires balancing growth and responsibility. Through a collaborative approach to ensure AI's positive impact while mitigating potential risks, India can leverage generative AI to take the top spot as a global tech titan.

Source:

2023 Word Economic Forum report
2023 research by Goldman Sachs
2023 World Economic Forum report
Gartner Research 2024
PIB Release, March 2024
Stanford AI Index 2024
EY India Research 2024
IoT Analytics, 2023
NASSCOM Report on India's Generative AI Start-up Landscape 2024

“Artificial intelligence is not a substitute for human intelligence; it is a tool to amplify human creativity and ingenuity.”

– Kai-Fu Lee

KIRTANE & PANDIT

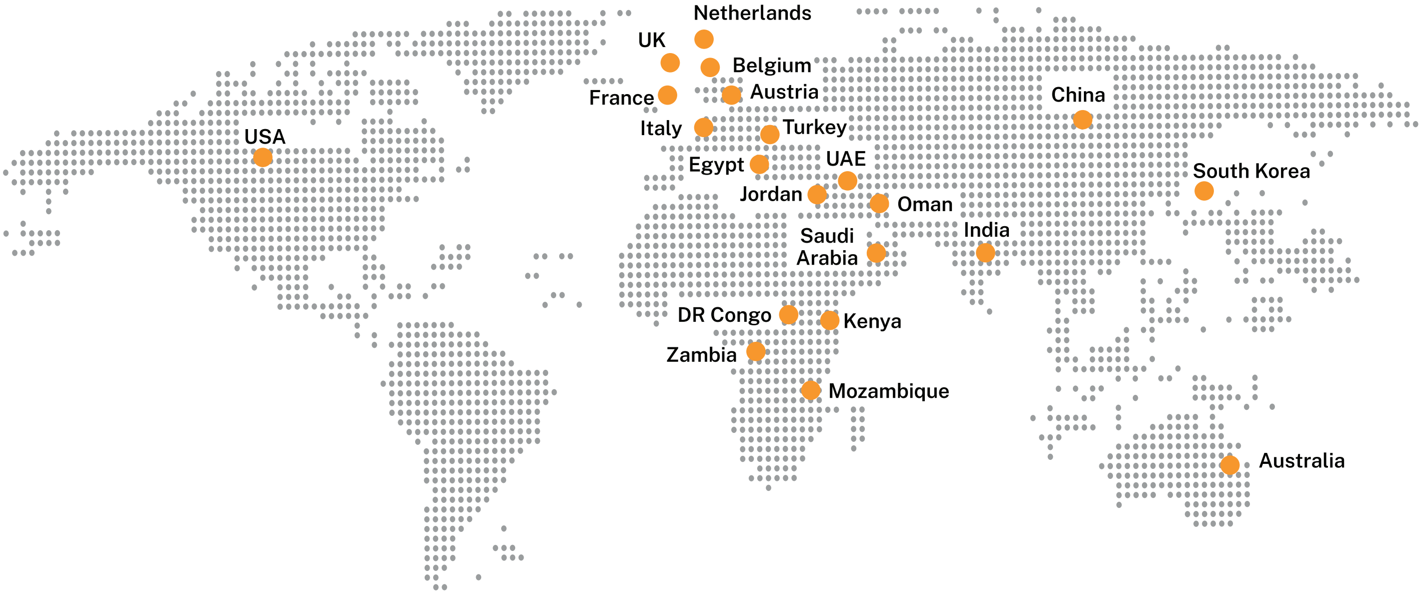
ABOUT US

Kirtane & Pandit LLP, Chartered Accountants is an Accounting, Auditing & Consulting firm with a well established network of financial experts across India.
Our motto 'A Step ahead, Always', reflects our value added approach in delivering sound financial solutions while we partner with you in your journey of growth.

With our extensive experience of 65+ years, we deliver a wide range of professional services in the areas of Assurance, Accounting & Advisory to reputed & listed companies from various industries across the globe.

We are a registered firm with ICAI and also a member of PCAOB, SEC, USA. We feature as category I firm of RBI and C&AG & IRDAI.
We are now one of the few proudly CERT-In Certified Chartered Accountants firms.

Our forward looking approach, technology driven environment, and belief in quality has enabled our employees to think differently & venture into newer emerging areas.



6.5+ Decades of Experience

35+ Partners

15+ Global Reach across countries

1200+ Employee Strength

Operating across India with **7** Offices

Client spread across **30+** Industries

KIRTANE & PANDIT

Pune

5th Floor, Wing A, Gopal House, S.No. 127/1B/11,
Plot A1, Kothrud,
Pune – 411 038, India
Contact no : +91 20 67295100 / 25433104
E -mail : kpca@kirtanepandit.com

Mumbai

601, 6th Floor, Earth Vintage, Senapati Bapat
Marg, Dadar West,
Mumbai- 400 028, India
Contact no : 022 69328846 / 47
E -mail : kpcamumbai@kirtanepandit.com

New Delhi

272, Rajdhani Enclave, Pitampura,
Delhi-110034, India
Contact no : +91-96438 74488
E -mail : kpcadelhi@kirtanepandit.com

Bengaluru

No. 63/1, I Floor, Makam Plaza, III Main Road,
18th Cross, Malleshwaram, Bengaluru – 560
055, India
Contact no : 080 23443548 / 23461455
E -mail : kpcabengaluru@kirtanepandit.com

Nashik

First and Ground Floor, Plot No. 115, Kalpataru
Bunglow, SN- 315/1D, Pathardi Phata,
Prashant Nagar, Nashik - 422010
Contact no : +91 253 2386644
E - mail : kpcanashik@kirtanepandit.com

Hyderabad

401 to 405, 4th Floor, Sanatana Eternal,
3-6-108/1, Liberty Road, Himayatnagar,
Hyderabad - 500 029, India
Contact no : +91 99127 41089 / 94400 55917 /
98480 44743 / 98480 46106
E -mail : kpcahyderabad@kirtanepandit.com

Chennai

No. 128, Old No. 34, Unit No. 1, 6th Floor,
Crown Court, Cathedral Road Gopalapuram
Chennai 600086
Contact no : 044 47990259
E -mail : kpcachennai@kirtanepandit.com

Follow Us On:

kpca@kirtanepandit.com
www.kirtanepandit.com

Authorized By

The Knowledge Management Team

The information presented in this publication is of a general nature, and we strongly advise seeking professional advice before making any decisions based on its contents. Kirtane & Pandit holds no responsibility for any loss or consequences arising from actions taken or refrained from, based on the material provided in this publication.

Kirtane & Pandit is a limited liability partnership registered with relevant authorities. Our registered office is located at 5th Floor, Wing A, Gopal House, S.No. 127/1B/11, Plot A1, Kothrud, Pune-411 038, India. We are an independent entity and not directly affiliated with any other organization mentioned in this publication.