Publication date:
June 2023

Cover created **by Bing Image Creator**

Prompt: "Analyst research on Generative AI for Singapore"

A Special Report Produced by Omdia, for ATxSG:

The Generative AI Revolution: Understanding, Innovating and Capitalising

Sponsored by:

ATxSG

OMDIA

Brought to you by Informa Tech

Contents

Introduction: What to Know, What to Watch, & What to Beware of in Generative AI

Natalia Modjeska, Research Director, AI

"Generative AI (GAI) is a relatively new branch of AI that focuses on the creation of novel artifacts ranging from text and images and video to computer code, music, and more."

You'd be hard pressed to find technology topics of more import and interest so far during 2023 than AI in general and Generative AI (GAI) in particular. Interest and commentary on the topic, moreover, has come from multiple audiences, from the EU's AI Act to Microsoft's $1 billion investment in OpenAI to the recent open letter from numerous technology leaders calling for a pause on AI research.

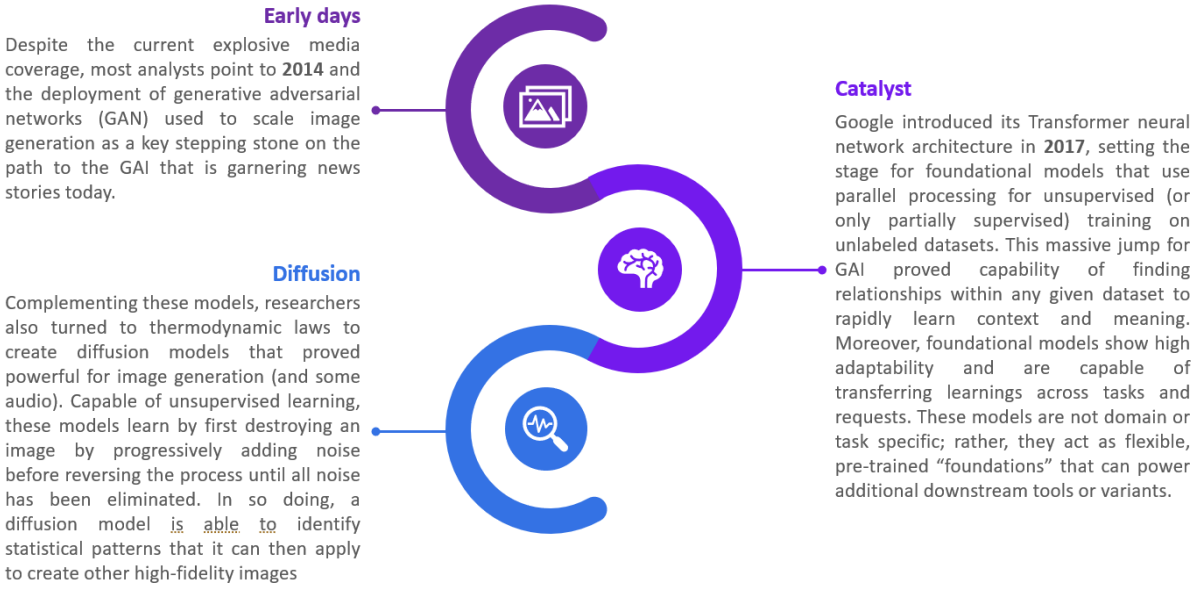
While interest is high, so too is confusion about what lies ahead for the Generative AI ecosystem during this pivotal moment. To cut through the noise and connect the latest developments to a business and industry perspective, Omdia offers this primer looking at the promise, current reality, and potential pitfalls of Generative AI as it stands today. Among other areas covered, it discusses:

- The basics of Generative AI: what is is and where it came from
- The Generative AI ecosystem and its foundational technologies
- Opportunities and risks associated with Generative AI
- Key market drivers and barriers
- Use cases to watch closely
- Expert recommendations on how to best approach this evolving market

Generative AI: What is it and where did it come from?

GAI uses neural networks and deep learning models leveraging large sets of training data to generate novel outputs. In lay terms, GAI creates original information, responses, or content (text, video, audio) based on what it learns from the vast data sets with which it is trained. It's a bit like if you were given a fill-in-the-blank question and used all of the previous information you'd absorbed to predict the right answer, except in this case GAI isn't limited to filling in a single blank but rather can respond to complex requests or tasks.

Figure. The Generative AI Timeline



Source: Omdia

If GAI isn't new, what then is driving this latest wave of interest? First, as we explain in greater detail below, increased computing power, cloud capabilities, and interfaces that make the technology accessible to a wide variety of audiences have acted as key market catalysts. Second, disruptive start-ups like OpenAi (the developer behind ChatGPT) have moved the technology from research project to public tool. Third, large well-established companies like Microsoft, Google, Amazon, and Meta (to name only a few) are putting both their pocketbooks and reputations into additional research and development of use cases.

Did you know? *Explainability remains as much of an issue (and arguably more so) for GAI as for its predecessors: GAI returns answers and creates ostensibly new artefacts, but researchers are unsure of exactly what data points have informed the responses or exactly which patterns GAI uses and how during this process.*

However, we are currently only riding the first wave of this fundamental sea change. In these early days, unknowns abound. We expect 2023 to be a pivotal, defining year that will be filled with experimentation, innovation, but also confusion, missteps and false starts. To cite only a few prominent examples, missteps range from Samsung's ChatGPT-caused security leak to Stability AI's decision to let artists "opt-out" (instead of opt-in) to having their art included in Stability AI's data training sets.

For businesses looking to GAI within this radically changing and radically transformative environment, both great opportunities and risks await. Understanding current market drivers, the emerging ecosystem, along with potential use cases and risks is a necessary first step in successfully approaching this transformative moment.

"The result of this confluence is that we are entering a stage of GAI at scale. Given the enormous social, economic, and ethical implications of GAI, it is no exaggeration to say that we could be witnessing the next transformational technology, as or more disruptive than the origin of the steam engine, car, or internet."

GAI growth factors and market drivers

Today's explosion in GAI would not be possible without a confluence of technical developments. **Increased compute power and cloud capabilities**, for instance, acted as necessary precursors for GAI model training at scale. Looking ahead, demand for accelerated computing is likely only to increase as appetites grow for more and bigger training models. We forecast an expanding need for faster CPUs, more powerful GPUs, and extended cloud computing resources . Moreover, because of the potential economic and sustainability costs of this need for increased compute power, longer term the industry could see a shift toward alternative compute architectures, including model compression and additional reliance on edge computing.

Increased compute power, moreover, simultaneously acts as both a driver for and is dependent on access to vast quantities of training data. **Accessibility to large datasets**, thanks in part to the explosion of social media and online data, functioned as a catalyst for models that power current GAI. Though it sounds remarkable, even greater quantities of training data will be required moving forward. This need has created market opportunity for an emerging trove of **synthetic data** vendors who are using GAI to generate additional data for training other GAI models. While innovative and an area to keep a close eye on, synthetic data does not come without risk, namely the potential of creating a toxic data loop in which errors or biases from original sources are reproduced and thus reinforced in subsequent synthetic data. From this perspective, synthetic data is both a driver and potential barrier to smooth GAI growth moving forward. Because of this dual movement of increasing data needs and the risk of toxic data loops, there is growing interest in training via smaller, smarter, fit-for-purpose datasets, along with improved data curation overall.

Perhaps the most important driver in today's GAI explosion was the emergence of **easy-to-use interfaces** on top of models, personified by ChatGPT and Stable Diffusion. Removing the need for users to have advanced technical or programming skills has democratized GAI and vastly increased

the potential user base. For instance, consider ChatGPT, which reportedly surpassed 100 million users within its first two months of becoming publicly available.[1]

Closely tied to this easy-to-use interface as a growth factor is GAI's ability to show **immediate benefits to users**. Again taking ChatGPT as an example, users receive near real-time answers to queries. Moreover, GAI unites this immediate return with a personalization at scale that makes it even more appealing to users. Outputs can be as unique as the user requests and/or designed to fit user preferences and profiles. In short, unlike many other technological advances that bring promise but little immediate substance, GAI returns tangible and personalized benefits that help prove its utility and have in turn sparked additional interest and use cases.

The **adaptability of foundational models** has likewise served as a catalyst in GAI growth. Because models so effectively transfer learnings from one area to another, there is no need to reinvent the wheel for every new application or use case. Foundational models are just that—foundations upon which additional instances or downstream tools can be built for more specific requirements.

Opportunities and use cases

From a business perspective, these market drivers have led to the development of **a complex GAI stack that touches many audiences and cuts across multiple sectors**. Consumers can directly use an end-to-end solution like ChatGPT. Conversely, Enterprise users could potentially leverage GAI to embed APIs to enhance their own solutions. Independent software vendors (ISVs) could potentially use GAI to create net-new functionality that can then be sold on as a standalone solution.

> *"These growth factors are coming together to fuel a GAI gold rush, with players both big and small staking claim to territory."*

Put another way, although we remain in the early days of this GAI transformation, business opportunities are emerging both for vendors who develop GAI solutions (for consumers or enterprise) and for companies who then incorporate those solutions into their own products. In short, GAI is creating new business opportunities by fueling a multi-faceted ecosystem of users and solutions. Behind all of these, compute hardware companies will need to continue to feed demands for increased CPU, RAM, GPU, and Cloud resources.

One proof point of growing GAI investment, niche cloud providers (like Lamda Labs) that specialize in supporting AI workloads have emerged. As the need for more, bigger, and more frequent training runs grows, these types of companies could be well positioned. Start-ups and companies offering innovating solutions like this are likely to fuel additional growth, but large and established players are always a threat to undercut on price or take over market share. Big-name tech firms that currently dominate cloud compute (think Microsoft, Amazon, Google, NVIDIA, Oracle, IBM, Baidu, Alibaba, etc.) are likely to continue or extend their leadership position as GAI demands grow.

Within this rapidly changing landscape, companies are showing strong interest in the promise and potential of GAI and are experimenting to see how it can help across a number of use cases. For instance, companies are already exploring how to best use GAI to **enhance user experience**. GAI's ability to interpret and return natural, human-like interactions, coupled with non-technical

[1] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

interfaces, promise to make products more approachable and flexible for users. Because of its adaptability, GAI isn't just fueling new products; it also promises to become a vehicle for making existing products better. Microsoft, for instance, has already announced that it will incorporate ChatGPT into Bing, Teams, and many of its most popular Microsoft 365 products.

Customer experience management likewise is an area with strong potential for incorporating GAI. Chatbots, digital engagement tools, and interactive self-help solutions are likely among the first wave of use cases to make use of GAI's natural language interface. GAI's ability to build on a response in ways that mirror human conversation can potentially give automated customer service interactions more flexibility. Given its ability to summarize information quickly, GAI could also be deployed to create additional user or how-to guides in customer service applications.

Marketing teams could likely also reap early benefits from GAI applications. Indeed, as our recent AI in Marketing report shows, numerous companies are already rushing to explore how ChatGPT and GAI functionality can power copywriting, generate unique images and visual assets, develop quick product summaries, create personalized buying experiences, power translation, and more.

Potential use cases within **healthcare and the medical field** are also beginning to emerge. Summarizing patient records, examining scans for abnormalities, and assisting in medical research could be growth areas for incorporating GAI into the healthcare field. Proponents also believe GAI will significantly lessen the time and costs involved in drug research, protein sequencing, and novel chemical compound discovery. Successes are already emerging: algorithms powered by AI recently made front-page news for discovering a new class of powerful antibiotics that could help treat drugs-resistant infections.[2]

"In creating GAI tools, we have set out to build an airplane and instead ended with a UFO that appears capable of interstellar travel. Our challenge going forward will rest not in making our UFO fly, but in doing so responsibly with full understanding, transparency, and trust."

Bradley Shimmin

Chief Analyst for AI Platform, Data & Analytics for Omdia

Software development too is a discipline that could be impacted by GAI's ability to help write better code more quickly. GitHub Copilot, released in June 2021, along with subsequent rivals, have shown the ability to return working source code after receiving natural language prompts. Code optimization, debugging, and documentation are all tasks GAI could help streamline. Looking ahead, GAI-powered coding could radically democratize software development by offering non-technical users a path to bring new products to life. Interestingly, evidence remains inconclusive regarding just how safe GAI-created code is, with some reports showing greater vulnerabilities and some noting the opposite. These types of contrary reports are more evidence of the inchoate state of GAI, with promise and potential currently outweighing definitive empirical evidence.

As this sampling of use cases suggests, GAI promises to bring efficiencies and transform workflows across a diverse number of industry verticals and sectors, for both B2B and B2C audiences. Given the vast potential of GAI, it is undoubtedly tempting for businesses to commit resources early in the hopes of gaining a first-mover advantage. There are, however, a number of risks and potential roadblocks to consider as well.

[2] https://www.bbc.com/news/health-51586010

Omdia review of the generative AI space, forecasting use cases across different industries

Should Generative AI...

In finance

Summarize financial budgets:
Output discrepancies in vital financial data

In marketing

Write marketing copy:
Cost savings, efficiencies, but note risk of copyright infringement and breach of data privacy rules

In communications

Draft emails and chat messages:
Efficient way to create a first draft message

"Current market conditions will spawn significant experimentation and innovation. Enterprises, startups, small to medium businesses, students, criminals, and others will tap into large language models and ask: What can be done with generative AI? Good and bad. In many ways, the next year will be similar to when Apple introduced the App Store—plenty of creative ideas, refinement, and perhaps some solid opportunities to commercialize generative AI."

- **Josh Builta,** Research Director

In customer service

Offer support and self service via chat:
Personalized and always available service

Summarize customer interactions, e.g., in a call center:
Summary with relevant information for contact center

"Currently, generative AI has the potential to improve customer support when there's a human in the loop who can correct/update generated content, if necessary. For example, generative AI can write the first draft of an email or chat message response that can be reviewed by a sender before delivering the message to a recipient."

- **David Myron,** Principal Analyst

In enterprise IT

Generate synthetic data to avoid personal identifiable information (PII):
Enable data sharing and grant data scientists access to what would otherwise be PII

Automate data integration and pre-procesing:
Speed up data preparation for use in downstream analytics workloads

Automate anomaly detection in data:
Early detecton of fraud, outliers or data drift to ensure model robustness and minimize operational risks

Generate charts and dashboard from natural language queries:
Automatic generation of personalized charts and dashboards, increased efficiencies and utilization

Monitor data for compliance:
Maintain compliance with internal and external requirements

Zero ETL - add data sets without coding:
Automatically add new data sources to existing analytical databases

Generate data catalogues; organise and manage enterprise metadata:
Effectively manage metadata that describes a broad spectrum of underlying data sources

In data science

AutoML co-pilot: enable conversational exploration and tuning of machine learning (ML) models:
Help investigation of complex search spaces, data scientists can explore new approaches and iterate faster

Enable conversational exploratory data analysis:
Data practitioners can explore data sets iteratively and faster using natural, conversational language

"There are a lot of ways that generative AI can be useful, but all applications need to be tempered with a healthy dose of practicality. These days businesses are all too eager to jump onto the latest buzzword bandwagon without taking a look at the practical applicability and implications for the enterprise, its customers and its employees. Any application of generative AI should ultimately keep the human at its center."

- **Hansa Iyengar,** Senior Principal Analyst

In software development

Generate, test and debug code:
Raises capability levels, improves output

"To realize the benefits of generative AI, we urgently need self-regulation, active governance, ethical leadership, and intentional integrity. In the exuberant pursuit of growth, transformation, and profit, those attributes will help us build more than a second generation of "weapons of math destruction" and more than a "new and improved" assembly line for disinformation, surveillance, propaganda, and (deep)fakes of all kinds. The goal is to use this technology to help humanity address pressing concerns facing us all today and build a sustainable future for generations to come."

- **Natalia Modjeska,** Research Director

In travel & hospitality

Generate trip itinerary, compare prices, handle booking:
Plan an end-to-end travel experience

In entertainment

Write a movie script:
Risk of copyright infringement

Compose music and sound effects:
Risk of copyright infringement

Create personalized TV, movies and video:
Higher engagement and satisfaction

In media

Write books, both fiction and non-fiction:
Risk of copyright infringement, hallucinations (non-fiction)

Transfer images between domains, e.g., photo to a sketch or painting:
Enjoy artistic hobbies very easily

"Although generative AI can boost creativity and productivity, it can also undermine job functions. Generative AI may be able to produce convincing outputs, but these can contain errors that could cause harm, particularly in sensitive use cases. Generative AI risks creating toxic data loops and ramping up the ways that synthetic media can be used to inflict harm. Taken together this can amplify discrimination, bias, misinformation, manipulation and abuse. All these issues must be addressed if generative AI is to be a force for good and produce the benefits and the opportunities of which it is so capable."

- **Eden Zoller**, Chief Analyst

In telecommunications

Generate synthetic data to improve intelligence in networks:
Improve fault and security management

Improve network design and engineering:
Useful planning tool

Train field technicians:
Improve efficiency of field technicians

In retail

Enable "Try before you buy", e.g., in fashion:
Personalized shopping experiences

In manufacturing

Generate novel designs:
Significant cost reduction for prototyping

In smart home

Be a conversational interface to smart home devices:
Enable devices to understand a wider array of commands

In medicine

Develop new drug molecules:
Speed up drug discovery and time to market

Generate synthetic data to train medical algorithms:
Alleviate data sharing challenges in healthcare

"Despite the fact that generative AI is still in its early stages of maturity, some forms are already in use in the healthcare sector. As this technology matures, it holds the potential to address some of the burdens and challenges that the industry faces. The use cases where generative AI are being and can be applied in the future are broad and far-reaching across the sector, from improving drug research and development programs and enhancing patient experience, empowerment, and privacy to reducing the administrative burden inherent in many of the sector's activities."

- **Andrew Brosnan**, Principal Analyst

In healthcare

Book and triage patient appointments, e.g., via chat:
Breach of data privacy rules and potential misdiagnosis

Summarize patient records and doctor's notes:
Output discrepancies in vital patient data

"The big question is whether one can replace a doctor with generative AI and in what contexts that might occur. It is already feasible, arguably. For example, because ChatGPT is not claiming to be specifically for medical use, as things stand I believe it cannot be regulated as a medical device."

- **Felix Beacher**, Practice Lead

"In creating generative AI tools, we have set out to build an airplane and instead ended up with a UFO that appears capable of interstellar travel. Our challenge going forward will rest not in making our UFO fly, but in doing so responsibly with full understanding, transparency, and trust."

- **Brad Shimmin**, Chief Analyst

Risks and potential market barriers

As with any gold rush, big winners will generate the most noise. But racing too quickly forward is likely to create issues for others. To start with, it's important to remember that all of the problems inherent to AI in general are carried into GAI as well. Foundational models are only as good as the training data itself. Any **biases or inaccuracies** in the underlying data can be replicated within the information that GAI returns. This potential to unknowingly reinforce erroneous or potentially harmful data, moreover, is complicated by the lack of explainability within GAI systems: it is currently impossible to trace which sources a GAI system has used to inform its answers or content, and researchers are not entirely sure even how a GAI system comes to the conclusions it returns.

In addition to this risk of unknowingly perpetuating biases or incorrect information, there is also the risk of **bad actors intentionally spreading misinformation** through the use of GAI. Many commentators have pointed to GAI as a tool that could be used for fraud and abuse at scale. For instance, consider the possibility of feeding GAI information so that it could mimic not just the ideas but the tone and style of a particular writer, thinker, politician, or even ordinary citizen. There could also be sabotage attempts by hackers, rogue investors, activists, or competitors generating large quantity of fake but plausible high-quality social media posts, reviews, and videos that disparage a company, its products, or leadership, all of which could lead to reputational and material losses. Yesterday's deepfakes or phishing schemes could be supercharged by GAI, with potentially disastrous consequences, eroding the trust which is the foundation of business.

Likewise, GAI presents difficult challenges on the **IP, copyright, and privacy** front. Remember that GAI creates something "new" based on something old. We are only at the tip of the iceberg for the likely numerous legal challenges filed by writers, artists, or even coders who believe their work has been unfairly co-opted and used by GAI to advance other projects. Potential legal issues could result as well if personably identifiable information (PII) is included, inadvertently or intentionally, in any of the training data or used by GAI.

Some of the very factors that make GAI a potential positive can also create negative repercussions that could disrupt markets. For instance, GAI's ability to help with routine task automation, content creation, and customer service assistance have also sparked fears of **worker displacement and unemployment.** Unlike with robotic automation, moreover, job displacement is likely to cut across both blue- and white-collar work.

Similarly, while we have pointed to enhanced compute power has acted as a catalyst for GAI growth, it also could become a market barrier moving forward. **Power consumption** has increased dramatically as CPUs and GPUs require more and more resources, prompting **environmental concerns** that will likely need to be addressed if current GAI growth curves continue.

The greatest risk, however, could be to a company's **reputation** if a serious data inaccuracy, copyright infringement, bias, or privacy issue develops due to the GAI solution implemented. This risk is particularly acute given the lack of explainability or ability to track which source material has contributed to GAI content. In short, if something goes wrong with a company's GAI, users and customers are likely to place blame on the company, not the technology itself.

Recommendations and looking ahead

We are at a watershed moment for GAI, with risk and reward both potentially waiting for those businesses considering investing in this powerful ecosystem. Without question, GAI is a disruptive technology. But it is disruptive to existing cultural and ethical norms as well. In that sense, we are in

the midst of a moment of cultural lag, where the technology, as usual, has raced ahead faster than the legal, political, societal frameworks have been able to change.

Omdia therefore recommends a deliberate, cautious approach toward GAI. Yes, there could be first-mover advantages for those ready to move forward now, but pitfalls (some known and some likely to still emerge) are also ahead. For every story of success in this rapidly changing and still developing market, there are likely to be an equal number of cautionary tales.

A well-thought out approach to GAI that includes the following key parameters will keep businesses positioned to strike the most effective risk-reward balance.

- **First do no harm.** Rather than immediately investing in a pre-packaged or self-service solution, begin by building internal expertise and understanding of the technologies underlying GAI and the ecosystem itself. For companies that lack the internal resources for this type of knowledge building or that will need to rely on outside partners, this is not a time to accept solutions blindly. Asking tough questions about what's underneath the hood of any GAI-enabled product and pressing for clear documentation and answers is key during this moment of growth and uncertainty. A good vendor will act as your partner and share not just the potential benefits of GAI but also what to watch out for.

- **Establish clear internal guidelines and guardrails.** Given the potential risks that accompany the space, companies should at minimum ensure that GAI polices explicitly map back to corporate privacy, security, and ethical guidelines. Establishing a clear set of rules governing use and development of GAI, along with enforcement mechanisms, can help create company-wide alignment in this transformative moment.

- **Monitor a changing landscape.** Staying abreast of developments is key during this time of rapid changes. This means not just watching the industry itself but also keeping a close eye on regulatory discussions and enforcement. Governments and regulators are moving quickly and ignorance will not be an excuse for companies that fall on the wrong side of new rules and limitations to GAI.

- **Start with low-risk, small-scale explorations.** Because the potential return is so great and GAI tools so easy to use, it can be tempting to jump in fully. Remember that we are still in a very early phase, where experimentation, confusion, and instability are likely to dominate. By starting small, either through internal development or with lower-risk vendor implementations, corporations can develop first-hand expertise and experience in this critical space while limiting potential fallout or monetary investment should anything go wrong or the market shift suddenly.

As with the emergence of other societally transformative technologies, GAI is currently balancing between two poles. On the one hand, it promises to usher in tremendous innovation and opportunities. On the other hand, it also is introducing incredible disruption and risk. Given its transformative potential and this evolving balancing act, organizations should create a careful and well-researched plan of action today to be best positioned for reaping full benefits tomorrow.

For additional details and commentary regarding benefits, risks, and the GAI market outlook, please see the following related Omdia reports: The Rise of Generative AI: A Primer and Generative AI: Market Landscape 2023.

An invitation to readers

Generative AI is fast-developing, high-profile technology with a vast array of potential applications from across every industry – with the ability to touch every stakeholder from government, to enterprise, to consumer. The challenge for all is in keeping up to date with the rapidly evolving products, their potential use cases, and the readiness of enterprises to adopt them.

At Omdia, we are tracking the evolution of Generative AI through our team of expert analysts to cover every angle – from across the AI, Cybersecurity, Vertical Markets, Enterprise IT, Service Provider and Media & Entertainment teams.

This special report from Omdia pulls a snapshot view from those analysts, by compiling a range of their insights on Generative AI as it stands right now. Read on to hear the views of our analysts across three broad areas of coverage:

- **Ecosystem -** Who is leading the conversation? From vendors, to enterprise, and with governments/regulators.
- **Technology -** What's the next level consideration beyond the initial hype? From hardware to software, including cybersecurity considerations.
- **Value-driven Use Cases -** Considering the specific impact for vertical industries and applications.

Ecosystem

Who is leading the conversation? From vendors, to enterprise, and with governments/regulators.

Regional/ National strategies addressing the growth of Generative AI

Bradley Shimmin, Chief Analyst, AI & Data Analytics

"Such efforts may sound straight forward, but AI-related legislation requires the creation of a living agreement capable of taking into account the needs of numerous, often competing entities"

The act of legislating artificial intelligence (AI) did not begin with the introduction of generative AI phenomenon, OpenAI ChatGPT, in November 2022. And yet, this groundbreaking chatbot service together with the rapidly exploding ecosystem of competing and divergent large language models (LLMs) have forever changed the way countries and regions view the opportunity and threat posed by generative AI.

For many, LLMs like ChatGPT have taken humans off the edge of the map to a place where there "be monsters," figuratively speaking. An open letter drafted by the Future of Life Institute in March 2023 and signed by more than a thousand AI innovators and practitioners called for a global six month ban on the training of AI systems larger than GPT-4 -- the most recent LLM behind ChatGPT.

What would prompt the very people responsible for creating generative AI to call for a pause? Three highly disruptive factors are at play here, each intertwining with the others to force many governing agencies into alternating states of either fight or flight, depending on the news cycle.

- **#1 -** The surprising scope of emergent abilities displayed by consumer-facing chatbots like chatGPT (version 4 in particular).
- **#2 -** An extreme and accelerating pace of corporate and academic innovation arising in association with transformer-based LLMs and text-to-image models like Stable Diffusion.

- **#3 -** High profile litigation surrounding ownership of and rights to training data, giving rise to further questions surrounding liability and accountability.

Nonetheless, within this report, Omdia will analyze a number of select, potentially bellwether regions to better ascertain how governing bodies are likely to approach the problem of simultaneously controlling and exploiting this fascinating new era of generative AI.

Regulating Generative AI

To date, AI-related legislation has sought to help the initiating institutions harness the transformational potential of AI for its own economic and societal benefit. In general, AI regulations across all regions incorporate several common elements, including:

- - Data privacy protection and ownership
- - AI ethics and responsible development practices
- - Intellectual property and copyright ownership
- - Liability and accountability
- - Safety and security
- - Research investment and promotion
- - Talent development and workforce support
- - Data sharing and access
- - Standardization development and support
- - Inter-agency and cross-border collaboration

Such efforts may sound straight forward, but AI-related legislation requires the creation of a living agreement capable of taking into account the needs of numerous, often competing entities. Those include public and private sector companies as well as research institutions, data providers, standards bodies, other governing agencies, investment agencies (plus the startups they fund), and international partners.

The resulting tug of war between interested parties has made it very difficult for large-scale legislation to make its way into law within a short period of time (see Figure X). Consider the General Data Protection Regulation (GDPR). Written by the European Commission for the European Union (EU), this data privacy legislation took more than four years to emerge and another two years to enforce compliance. The much more recent and AI-specific legislative document from the EU, the Artificial Intelligence Act (AI Act) has taken only three years to move from a white paper to committee vote (scheduled for April 26th, 2023).

Figure. Global legislation summary

Region	Country	AI Legislation	Focus of Legislation	Date Initiated or enacted	Current Status
North America	United States	The Algorithmic Accountability Act	Protections for individuals from automated systems	2020	Proposed
		The National AI Initiative Act	development and use of trustworthy AI in the public and private sectors; workforce preparation	2022	Active
		Global Catastrophic Risk Management Act	Protections against catastrophic use of AI and other technologies	2022	Proposed
		California Consumer Privacy Act	Data privacy and protection	2023	Active
	Canada	Artificial Intelligence Data Act (AIDA)	AI transparency and explainability; research innovation and collaboration	2022	Proposed
South America	Brazil	Brazilian Data Protection Law (LGPD)	Data privacy and protection	2020	Active
Western Europe	European Union	GDPR (General Data Protection Regulation)	Data privacy and protection	2018	Active
		EU AI Act	AI regulation transparency and ethical guidelines	2021	Proposed
	United Kingdom	UK National AI Strategy	AI research innovation and ethical guidelines	2021	Proposed
Central Europe	47 Member States	Council of Europe (CoE) and European Convention on Human Rights	Updated to identify and develop standards capacity-building solutions for intersection of humans and AI	1950	Active
Asia Pacific	Singapore	Personal Data Protection Commission (PDPC)	Guidelines covering data privacy and security	2020	Active
		The Info-Communications Media Development Authority (IMDA)	Organization leveraging PDPC to protect citizen rights	2016	Active
	China	A Next Generation Artificial Intelligence Development Plan	Mandating standards of use (ethics, development, access) of data and AI	2017	Active
		Administrative Measures for Generative Artificial Intelligence Services	Guidelines ensuring that AI abides by shared human values	2021	Proposed
	India	National Strategy for Artificial Intelligence	AI research innovation workforce development and ethics	2018	Active

Source: Omdia

And yet, even with its experience in governing data and AI concerns, the EU AI Act seems rushed comparatively, especially given last minute arguments over semantics (foundational model vs general purpose model) and the act's general approach (regulating according to capacity to do harm). Given the extreme degree of difficulty and time involved in drafting legislation of this size and scope,it is plain to see that governing bodies are simply not capable of keeping pace with the rate of technological innovation and varying nature of public reaction to generative AI.

Thinking globally, acting locally

While most governing bodies still favor this scale of legislation, Omdia sees a much more diverse array of approaches emerging from multiple players that emphasize a more tactical, arguably reactive response to generative AI. Within the short span of Q1, 2023, Omdia has noted several such "reactions'' taking shape across Though highly divergent in nature, these efforts all seek to put the figurative genie back in the bottle, to hopefully create enough time for society to more fully accurately assess the actual privacy, security, accountability, and ethical risks posed by generative AI.

Italy: In early April 2023 the Italian data protection authority, Garante, temporarily blocked OpenAI's ChatGPT chatbot, citing a supposed data breach and a lack of compliance with GDPR privacy regulations. At issue is OpenAI's inability to provide information regarding the company's legal right to process potentially personal data in training ChatGPT. Another concern involved a lack of tools capable of helping users manage the use of their data and recognize inaccurate information.

Shortly after this announcement, OpenAI's CEO Sam Altman promised to address these concerns directly but did not go into any details.

It is important to note that Germany was rumored to follow suit, but as of the writing of this report, Italy stands alone in taking such action. Italy's lone action, which took place within the EU, therefore, illustrates the difficulty in enforcing AI regulation across borders, even where legislation calls for a unified front.

India: Taking a decidedly contrarian approach compared with the global call for regulation, lawmakers in India have stated that they will not attempt to control the growth of generative AI within the South Asian market. The Indian government, which plans to create several public projects featuring generative AI, feels that controls such as those proposed by the EU AI Act will ultimately stifle innovation. India feels that it can better govern generative AI through innovation and application.

Singapore: Singapore has long been a proponent of AI, having already rolled out several public-facing projects. Rather than seek to place comprehensive controls on generative AI, the country is instead looking to already established methodologies that call for companies to perform a self-assessment before releasing an AI product of any kind. Built by the Info-Communications Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC), the AI Governance Testing Framework and Toolkit guides innovators toward the creation of what the government calls a minimally viable product (MVP) that's capable of meeting at least eight discrete governance principles.

United States: Historically, the US government has to date taken a hands-off approach to AI governance, choosing to publish guidelines and leaving most legislation to play out on a state by state basis. Over the last year, however, the country has reversed course and is now pushing for country-wide regulation, calling for an AI Bill of Rights in late 2022 as one example.

The rapid emergence of generative AI, however, has not only accelerated existing but has also sparked several reactive efforts spanning both the public and private sectors.

For example, many in the government are beginning to look to older pieces of legislation such as Section 230, which protects against exposure to content posted by others, to tackle generative AI. Similarly, the US Copyright Office in March 2023 suggested that AI-generated work be copyrighted, opening up several new avenues of regulation and control regarding prompt engineering. As a collection of self-governing stats, the US is also evolving several concurrent but independent policies. For instance, in January 2023, New York passed the New York City Bias Audit Law (Local Law 144), which may be used to govern LLM training data. And in January 2020, the California Consumer Privacy Act (CCPA) became effective and will likely play a role similar to GDPR in potentially slowing generative AI.

China: Standing in stark contrast to the United States, China has been quite active and timely in directly addressing emerging AI technology. In 2022, for example, the country introduced new Deep Synthesis Provisions, which targets deepfake technologies spanning audio, video, text, and images. Falling into alignment with with these provisions, in early April 2023, the Cyberspace Administration of China (CAC) issued a set of draft measures (titled Management of Generative Artificial Intelligence Services) that directly controls the use of AI to generate content (code, images, text, video, et al.). If adopted, these measures will require companies providing such AI services to meet a set of very stringent requirements. The most striking requirement concerns the gathering of training data, where providers must be able to prove the accuracy, objectivity, veracity, and diversity of their training data.

> *"Current efforts point to a chaotic patchwork of regulations limited to the state and/or country level, each focusing not so much on the models themselves but rather on control of the data used to train those models"*

European Union: The EU was very early to draft laws regulating both data- and AI-related concerns. Early laws such as GDPR, which went into effect in May 2018, have created a sizable impact globally, influencing subsequent legislation such as CCPA in the United States. In what overtly seems to be a case of perfect timing, the European Parliament is currently readying the Artificial Intelligence Act. First introduced in 2021, this act introduces several potentially disruptive mandates covering issues specific to generative AI, including data quality, accountability, transparency, and human in the loop (HiTL) oversight.

Interestingly, the AI Act will not pay much attention to AI systems classified as posing minimal risk (e.g. a spam filter). High risk systems (ostensibly ChatGPT), however, will have to undergo extensive scrutiny. The goal, as stated by EU representatives, is to preserve the safety and fundamental rights of all EU citizens. However, continuing infighting over the inclusion of language specific to generative AI as well as semantic differences regarding what a system means to be high or low risk, could either slow the enforcement of the AI Act or worse stifle EU innovation.

In summary

With so many divergent voices currently endeavoring to control generative AI, it is unlikely that we will see any sort of coordinated, consistent, and comprehensive plan of action emerge on a regional level. Rather, current efforts point to a chaotic patchwork of regulations limited to the state and/or country level, each focusing not so much on the models themselves but rather on control of the data used to train those models.

Such controls are of course a continuation of much older battles over privacy, ownership, and the right to monetize information -- both information belonging to individuals or companies and information belonging to AI itself (i.e., whatever is being generated by generative AI).

Authority over issues such as bias, inclusiveness, accuracy, et al. are certainly on the docket. However, given that current researchers are still grasping to understand how LLMs do what they do, it is unlikely that any legislation can be written that would effectively mandate its use.

One mitigating factor (a wild-card, perhaps) that has recently emerged, which may serve to simplify this complex picture is the recent introduction of several LLMs such as Databricks' Dolly, Stability AI's StableLM, and Nomic-AI's GPT4All-J. These models, most available for commercial use under the lenient Apache 2.0 license, stand as alternatives to super-large LLMs like OpenAI ChatGPT and Google Bard in two important ways. First, they can be run on local hardware and do not cost very much to train. And second, they espouse openness, spanning model code, weights, and training data.

Certainly this level of transparency opens up LLMs to "bad actors." However, it plays a much more crucial role in that it enables more researchers to work on understanding LLMs, and it makes any kind of regulatory control easier to understand and enforce. In this way, these diminutive LLMs can help law makers and users of generative AI to more fully apprehend the current and future impact of generative AI upon established social, cultural, economic, political, and technical infrastructures.

Written 23rd April.

The standard way to deal with non-standard technology

Lian Jye Su, Chief Analyst, Applied Intelligence

"Not surprisingly, there are several limitations to standardization in AI governance. The most glaring one is the lack of generative AI-specific standards"

The most appropriate non-regulatory approach to achieve robust governance of generative AI is via standardization. Since the early day of AI democratization, various Standard Development Organizations (SDOs), such as ISO, IEEE, ETSI, ANSI, and DIN, have introduced AI standards that help to facilitate industry-wide alignment and collaboration, leading to rapid AI adoption. In addition, efforts from industry-specific SDOs, like ITU, also provide guidelines for vertical-specific use cases not covered under global AI standards.

Here is a list of standards issued or in the process of development by key SDOs related to AI risks and governance:

Figure. List of standards issued/in process of development

Code	Title	Published by	Published on
DIN SPEC 92001-1	Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Meta Model	DIN	April 2019
DIN SPEC 92001-2	Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness	DIN	December 2020
ETSI GR SAI 004 V 1.1.1	Securing Artificial Intelligence (SAI) – Problem Statement	ETSI	December 2020

ISO/IEC AWI TS 29119-11	Information technology — Artificial intelligence — Testing for AI systems — Part 11	ISO/IEC	December 2020
ISO/IEC TR 24029-1	Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview	ISO/IEC	March 2021
ETSI GR SAI 005 V 1.1.1	Securing Artificial Intelligence (SAI) – Mitigation Strategy Report	ETSI	March 2021
ETSI GR SAI 002 V 1.1.1	Securing Artificial Intelligence (SAI) – Data Supply Chain Security	ETSI	August 2021
ISO/IEC TR 24027	Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making	ISO/IEC	November 2021
ANSI/CTA 2096	Guidelines for Developing Trustworthy Artificial Intelligence Systems	ANSI	November 2021
IEEE 2941	IEEE Standard for Artificial Intelligence (AI) Model Representation, Compression, Distribution, and Management	IEEE	March 2022
ISO/IEC 38507:2022	Information technology. Governance of IT. Governance implications of the use of artificial intelligence by organizations	ISO/IEC	May 2022
ETSI GR SAI 001 V 1.1.1	Securing Artificial Intelligence (SAI) – AI Threat Ontology	ETSI	June 2022
ISO/IEC DTR 24368	Information technology — Artificial intelligence — Overview of ethical and societal concerns	ISO/IEC	September 2022
P2941.1	Standard for Operator Interfaces of Artificial Intelligence	IEEE	November 2022
ISO/IEC 23053	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)	ISO/IEC	November 2022
ISO/IEC 23894	Information technology — Artificial intelligence — Guidance on risk management	ISO/IEC	February 2023
ISO/IEC DIS 42001	Information technology — Artificial intelligence — Management system	ISO/IEC	Under development
ISO/IEC DIS 42005	Information technology — Artificial intelligence — AI system impact assessment	ISO/IEC	Under development
ISO/IEC DIS 42006	Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems	ISO/IEC	Under development

Source: Omdia

Not surprisingly, there are several limitations to standardization in AI governance. The most glaring one is the lack of generative AI-specific standards. Standard formulation tends to be time-consuming. The process is slow because it is based on consensus between all stakeholders and involves contributors from around the globe who are for the most part volunteering their time and expertise. A standard is first drafted by international experts, before being reviewed by national bodies who will review, make additions, and ultimately vote on the final draft. In the case of generative AI, the lengthy process means most, if not all, standards are developed before the emergence of ChatGPT. Also, standards are non-legal binding frameworks and cannot be enforced unless backed by strong legislative authority and mandate.

Nonetheless, as generative AI is becoming an AI technology that will see global adoption, having regional and national agencies working with SDOs is a necessity. Compared to legislations that are often regional or country-specific, SDOs often operate at a global, if not regional, level. Since standard development efforts are mostly industry-led, generative AI-related standards from these SDOs will likely be accepted worldwide, facilitating industry-wide adoption and alignment. At the same time, governments can lean on the expertise of private sector and AI developers. The formulated standards will be based on a solid foundation of technical mastery and market understanding, ensuring their relevancy, applicability, and acceptance.

Of course such process will take time. With the mounting pressure to legislate and monitor generative AI applications, Omdia believes the industry will actively work together to address this gap. When developing an universally-accepted generative AI governance standard, Omdia believes it must meet the following criteria:

- Shared values: The governance standard must be based on core societal values that prioritize the wellbeing of all human beings, free from any bias, discrimination, and abuse.
- Wide participation: Considering the widespread adoption of generative AI, the standard formulation process must involve as many keystakeholders as possible, including generative AI developers, end users, leading generative AI technology suppliers from hardware, software, and services, regulators, public agencies, and auditors.
- Clarity and transparency: The standard must be clear in its objective, implementations, and outcomes. All parties involved should be able to understand and master the methodologies and approaches proposed in the standard to govern generative AI. More importantly, end users of generative AI must be able to observe and appreciate the benefits of the standard implementation.
- Roles and responsibility: The standard must also propose clear roles and responsibilities for all stakeholders, especially the roles of generative AI developers and implementors in ensuring the performance, transparency, and compliance of generative AI solutions.

Infrastructure Investments

The promises and potential of generative AI has led to a flurry of investment into training and developing various foundation models. According to OpenAI, the AI training workloads double every 3.43 months. Given the amount of compute resources to support AI training, it is not surprisingly that organizations are targeting High Performance Computing (HPC) and supercomputers to further develop their AI models.

At the moment, China and the US continue to expand their arsenal of supercomputers. China tops the Top 500 Supercomputers ranking with 162 entrants, while the US is a close second at 127 entrants. Both countries have been actively accumulating large computing resources on top of the AI design and development capabilities of top AI vendors and startups in these two markets. Both countries have also actively funding universities and research institutions to promote fundamental research in AI.

Meanwhile, the UK government is the first to announce its intention to build an exascale supercomputer dedicated for generative AI. The supercomputer is designed to analyze large datasets more quickly, train generative AI models in a more efficient manner, and optimize the cost required to develop generative AI. The entire investment is estimated to cost GBP 900 million or US$ 1.1 billion.

Since generative AI is very much an advancement in AI software, the UK government has taken a step further by looking into creating a “sovereign” foundation model for generative AI. The country believes that it will prevent UK-based tech startups from losing out to their counterparts in the US and China, as most, if not all, of the foundation models are developed by startups from these two top markets.

Recommendations

Considering all the aforementioned approaches, Omdia would recommend the following to all jurisdiction and public agencies around the world who wish to leverage a non-regulatory approach:

- Be proactive instead of reactive: Governments and public agencies must be proactive at identifying emerging trends in generative AI and non-regulatory approaches to expand infrastructure support, provide the right incentive, and mitigate risks and concerns.

- Active participation and contribution to SDOs: Governments and public agencies must actively participate in relevant global SDOs for AI. Public-private collaboration is critical to ensure the standards developed are relevant, applicable, and enforceable.

- Close consultation with generative AI developers and users: Aside from participating in SDOs, governments need to engage with local generative AI developers and users to understand their AI development needs in terms of hardware and software, future plans and direction for generative AI, concerns around legislation and standardization before determining the right approach. Such understanding will lead to a trusting relationship that enables a formulation of a non-regulatory framework to leverage the growth in generative AI.

Written 9th May.

Generative AI: Who's running the race?

Andrew Brosnan, Principal Analyst, AI in Life Sciences

The emergence of GAI over the past six months is sparking an inflection point for AI market adoption. It is unprecedented. This inflection point was largely triggered by an elegantly designed natural language interface, ChatGPT, which essentially unleashed the potential of LLMs. LLMs have been around for a few years, but few people outside of data scientists bothered to experiment with them. ChatGPT brought the power of an LLM to almost anyone.

Current market conditions will spawn significant experimentation and innovation. Enterprises, startups, SMBs, students, criminals, and others will tap into LLMs and ask: What can be done with GAI? Good and bad outcomes will result. In many ways, the next year will be similar to when Apple introduced the App Store—plenty of creative ideas, refinement, and perhaps some solid opportunities to commercialize GAI.

"The trajectory of GAI follows the classic lifecycle of techno-logical innovation adoption. We are in the very early phases where there are no precedents, but there are lots of experimentation, lots of startups, and lots of confusion"

During this experimentation period, the challenges for GAI remain significant—chiefly, how to address misuse, bias, cost, and copyright/IP issues. GAI does not solve underlying challenges that face NLU. It will be messy. Can and will guardrails be put in place for using LLMs?

Despite the messiness and risk, there is a significant opportunity. Companies should experiment with GAI, as long as they do so with common sense discipline. Namely, they should

- Use business disciplines—what problem are you trying to solve?
- Practice AI risk management—responsible AI

Nobody can really know where this will end up, but I believe the trajectory of GAI follows the classic lifecycle of technological innovation adoption. We are in the very early phases where there are no precedents, but there are lots of experimentation, lots of startups, and lots of confusion.

Some of the more promising developments to watch for are going to be how LLMs integrate APIs into other software—different/competing applications and LLMs, different interfaces, and private domains. In particular, when will LLMs be able to narrow their focus and tap specific domains only?

I believe we will soon be entering a time of significant competition between LLMs. In theory, there is money to be made from entities building GAI applications—and the only way they will be able to do that is by accessing LLMs.

For now, the players that are going to "win" in GAI are most likely the tech firms that provide cloud compute and or will create LLMs—Microsoft, AWS, Google, NVIDIA, and IBM. At the same time, the players that are best positioned to innovate in and commercialize GAI are likely the following:

- A small group of software as a service (SaaS) companies that are AI innovators.
- Companies that lead the world in embedding AI into their products and services already—Adobe, Salesforce, IBM, SAP, Oracle, and AWS.
- Sophisticated conversational AI companies like Amelia, Verint, LivePerson, Interactions, and several other customer service virtual assistant specialists.

The big tech consulting firms and systems integrators—Accenture, Deloitte, PWC, Bain, BCG, WiPro, and a host of others—will have to work hard to come up to speed since so many enterprises rely on them for broader-based technological and digital transformation.

Introducing the key players for AI

This section features a selection of most, but not all, key industry players. It is not intended to be a comprehensive list of all the key players in the GAI field.

Open AI

- OpenAI created ChatGPT and DALL-E, the pioneering GAI applications, and the GPT LLM. Through its partnership with Microsoft, OpenAI's ChatGPT and other GAI capabilities will be embedded into a wide range of commercialized Microsoft products, which could be a significant accelerator for GAI use cases—or, conversely, reveal the challenges to current approaches.

Stability AI

- Stability AI develops a software tool that uses its Stable Diffusion (DL model) for text-to-image generation. Stability AI's model offers notable differences from the GPT-3 Transformer model from OpenAI, especially in that the model is much smaller, making it capable of running on PC hardware. Stable Diffusion and Open AI's DALL-E could change the way we create and consume visual content and augment the prototyping and design lifecycle for a wide variety of industries.

Microsoft

- The software and cloud vendor that is present in every enterprise is almost unavoidably going to be the channel that brings AI into many companies. Microsoft has focused its AI-related research and development on making the technology manageable, secure, and compliant. It has also focused on

competing for its share of the historically enormous compute spend as a major and price-competitive provider of virtual machines as well as one of the fuller ranges of managed AI services.

- Microsoft will provide OpenAI with up to $10bn of credit for Azure cloud services in exchange for part of OpenAI's future profits until the $10bn is earned out. Microsoft also receives a license for OpenAI's full suite of AI models. It is likely that Microsoft will try to fold GAI into as many of its applications and lines of business as possible, both to hasten the earn-out of OpenAI shares and to reinforce its competitive position.

Google

- Google has an important role in GAI throughout the technology stack. Google Cloud already hosts two LLM players, Cohere and Anthropic, with the potential to add more; in addition, Google has created and operates four LLMs: BERT, PALM, MUM, and LaMDA. The new GAI interface Bard is a gateway to LaMDA, while PALM is the LLM engine behind Minerva, a GAI model that addresses quantitative reasoning (to solve mathematical and scientific questions). The stakes are very high for Google/Alphabet. Omdia believes they will get GAI search right over the next two years. They may lose some search market share regardless.

NVIDIA

- NVIDIA is the market leader in AI silicon. Its data center GPUs accounted for 70% of the market in 2022 by revenue. Since 2018, NVIDIA has been responsible for a succession of landmark server GPUs both for model training and for inference scale-out. Most GAI projects will be trained on their GPUs. Omdia expects NVIDIA to remain the market leader out to 2027. Although alternatives will eat into its market share, overall growth in the industry will be enough for it to continue to grow, and GAI is already showing signs of being a boost to its business.

- As the software moat will continue to protect NVIDIA, much of the adoption of alternative silicon will happen through the hyperscale channel, where developers do not necessarily have to engage with the hardware layer.

AWS

- As a global leader in cloud services hosting, AWS is deeply involved in driving forward and supporting the GAI marketplace. AWS provides a rich platform (Bedrock) for the development and hosting of LLMs centered around its SageMaker brand. This includes not only tooling, but also training and inferencing hardware tuned specifically to support AI at scale (as with LLMs).

- AWS has invested heavily in the creation of LLMs and has partnered with leading LLM vendors, including OpenAI, to support both training and inferencing. Omdia expects AWS to grow rapidly in the GAI space over the next few years, building on its global dominance as a cloud provider, coupled with its already mature AI platform, Amazon SageMaker, and growing portfolio of GAI solutions, including Polly, Lex, Transcribe, and Translate.

Hugging Face

- Created in 2016 with the goal of making Transformers and NLP, in particular, more readily available to a broad range of users. The company has greatly influenced the market through its focus on

collaboration, openness (particularly in supporting the BLOOM LLM), and innovation (via its very popular Transformers library).

BLOOM (BigScience)

- The BLOOM LLM was created as a part of a year-long workshop launched by Hugging Face and led by many international researchers from 70 countries and 250 institutions, including those involved in the creation of OpenAI's GPT model.
- Capable of a wide range of tasks, this pre-trained model features 176 billion parameters and can support 46 languages and 13 programming languages.
- As an open-source project that was created entirely in the open, BLOOM stands as a unique and potentially disruptive force within the GAI market, promising to make LLMs available to all researchers without any kind of corporate cost or control.

Oracle

- Oracle has been employing generative capabilities to optimize SQL queries and query plans. Like other hosting providers, the company seeks to leverage Oracle Cloud Infrastructure (OCI) resources to form an optimal LLM training and inferencing platform. A key capability is the company's ability to access high performance computing (HPC) resources (e.g., large-scale GPU clustering) that are tuned to Stable Diffusion, BLOOM, DreamBooth, et al.
- Oracle's prowess in handling large datasets at scale, especially across hybrid cloud-premises deployment scenarios, will enable the company to gain a lasting foothold in the GAI marketplace, particularly as enterprises begin contextualizing (e.g., aligning) LLMs using their own training data.

Cohere

- NLP platform that offers a variety of services in three major categories: classification, generation, and embedded functions.
- Cohere's platform can support common programming languages such as Python, Go, Node.js, and shell script, allowing API calls to be easily integrated into disparate application endpoints.
- Cohere as a software tool is important in this space, as it allows integration into end-user applications with a focus on users with limited or zero code development experience.

Anthropic

- Supporting GPT-based solutions and building on its own implementation of ChatGPT (Claude), Anthropic is currently focusing on several areas of natural language, human feedback, reinforced learning, code generation, review, and interoperability and predictability.
- The company is looking to evaluate and develop tools around automating model behavior discovery and AI self-improvement feedback for scaling bias and reducing AI harm.

Written 3rd March.

Profiling a key player: The Microsoft view on Generative AI

Alexander Harrowell, Principal Analyst, Advanced Computing for AI

The following section consists of two articles published to Omdia.tech.informa.com in Q1 2023.

"Even a startup as prestigious as OpenAI is willing to sell equity remarkably cheaply in exchange for committed funding and infrastructure"

Microsoft is determined to keep OpenAI on-platform, but only on the cheap

Microsoft's $10bn investment in OpenAI reflects the company's determination to keep OpenAI on the Azure platform more than anything else. A deal structure represents the parties' expectations for the future. This deal's structure, which is complex in the extreme, tells us that the future potential of artificial intelligence (AI) is enormously valuable but still very uncertain, and consequently subject to a substantial discount. What is absolutely certain, however, is that getting there will need a huge amount of GPU compute capacity, and selling it will be good business. In the short term, the financial "vibe shift" of 2022 means that even a startup as prestigious as OpenAI is willing to sell equity remarkably cheaply in exchange for committed funding and infrastructure.

Microsoft is in for $10bn, although terms and conditions apply

On January 9, the blog Semafor revealed that Microsoft was preparing to make an investment of as much as $10bn into OpenAI. It's worth noting that there are currently two distinct news stories circulating regarding Microsoft and OpenAI – separately, the Financial Times reported that the company and some other venture capitalists are considering a tender offer for existing shares of OpenAI held by employees and others that would value OpenAI at $29bn. This, however, doesn't represent new investment into OpenAI, but rather an opportunity for Microsoft to increase its stake and for existing shareholders to exit.

The second Microsoft investment is more interesting and comes in the form of a complex deal under which Microsoft contributes $10bn to OpenAI. OpenAI undertakes to repay the $10bn to Microsoft in the form of a first charge of over 75% of its profits and, once the $10bn is repaid, to issue enough

new shares to grant Microsoft 49% of its equity, with the OpenAI Foundation getting 2% and the founders the rest. Some or all of the $10bn may be in kind rather than cash, in the form of Microsoft Azure service credits. However, the distinction is less important than it usually is, as OpenAI's biggest use of cash is almost certainly its Azure infrastructure-as-a-service (IaaS) bill. During its period as a non-profit organization, it sometimes reported spending 66% of its net income on cloud computing.

This deal structure contains elements of several different financial instruments with different implications, discussed below.

A complex deal structure

For a start, there is the sale of Azure infrastructure services to OpenAI. Depending on whether the $10bn is cash or service credit, this is either implicit (Microsoft provides the cash, knowing that it will come back in Azure sales) or explicit (Microsoft provides OpenAI with Azure credits it cannot use elsewhere). This is simple enough. This sale is vendor-financed, with OpenAI not needing to come up with any cash up-front, instead repaying through a charge over its profits.

Once the $10bn is paid off, Microsoft gets just under half the ordinary shares in OpenAI. The founders' 49%, plus the OpenAI Foundation, make up a bare majority and will presumably control the company – unless Microsoft has already acquired a substantial holding from the tender offer. This element in the deal has probably aroused more comment than any other; if OpenAI was to really succeed, Microsoft has acquired a very large stake in what could be a very valuable company cheaply. Also, Microsoft is not taking much risk on the OpenAI shares, as it only gets stock after OpenAI has made at least $12.5bn in profits.

As for the vendor financing, this in itself is routine in the IT industry. Unlike a normal vendor-financing deal, though, OpenAI is only required to make repayments once it is profitable, and the repayments are pro-rated to its profitability. As such, the vendor financing does not have a fixed term and the effective rate of return Microsoft receives will vary depending on how long it takes for OpenAI to pay it off. This financing has a curious mix of equity-like and debt-like properties – it is debt-like in that it has a fixed face value to repay, but it is equity-like in that it does not have an interest coupon or a fixed term; it is junior to ordinary debt, and the holder takes on the risk that there will be no profits.

What the structure tells us

The payout structure we just described has three possible outcomes depending on OpenAI's future performance. In all cases, OpenAI gets $10bn of IaaS capacity up-front.

- **Things go really well** – OpenAI both triumphs as a research organization and finds a way to commercialize its discoveries. The company is valued in the trillions. Microsoft has bought half of it for $10bn that might not even be cash, has its IaaS business, and has the opportunity both to fold OpenAI technology into products such as Office, Dynamics CRM, and Visual Studio, and to resell it via Azure. OpenAI has sold half the equity, but who cares?

- **Things go moderately well** – OpenAI makes it to profitability and begins paying Microsoft off gradually. Microsoft has a reliable IaaS customer and the prospect of an equity kicker. OpenAI is grateful for the escape hatch of not paying Microsoft during its bad quarters.

- **Things go badly** – OpenAI goes bankrupt. Microsoft loses out on the fraction of the $10bn of Azure service that has been drawn down and used, less any repayments OpenAI has made. Asset

recovery on unused cloud service credits, handily, is 100% as Microsoft can just sell the unused capacity – it does not need to search OpenAI's data center. Microsoft can easily live with this, while OpenAI is no more bankrupt than it would otherwise be.

It's clear from this list that Microsoft has gained a large amount of exposure to the potential upside in exchange for an investment that, although substantial, amounts to relatively little value-at-risk at any one time, with the further twist that it has also made sure of a large and growing Azure account that preferentially buys its highest-margin products such as top-line GPU instances. There is also the possibility of more IaaS business from developers in the OpenAI ecosystem or enterprises running their own instances of OpenAI infrastructure, as well as some marketing value in the glamor of being associated.

OpenAI, for its part, has locked in financing for probably its biggest use of cash on very easy terms – for as long as it takes to reach profitability, it will be paying nothing on $10bn of its infrastructure. To do so, though, it has had to give Microsoft a big share of future equity. A sweetener in this is the possibility that Microsoft might become a major customer and value-added reseller as well as a supplier. In general, it helps that OpenAI's biggest supplier has a major stake in its success.

It has been suggested that Microsoft is interested in using OpenAI technology in Office and potentially other products such as Bing, and on January 16, Microsoft announced general availability of OpenAI Service for Azure, a managed services product that includes GPT-3.5, DALL-E 2, and Codex and, in future, ChatGPT. This is a client version of the deployment Microsoft is using for Github Copilot, as well as AI-powered features in PowerBI and the Designer image generator app. However, this is only general availability with caveats – the signup form requires around 29 answers and specifically excludes customers that don't have a Microsoft key account manager.

Written 19th January.

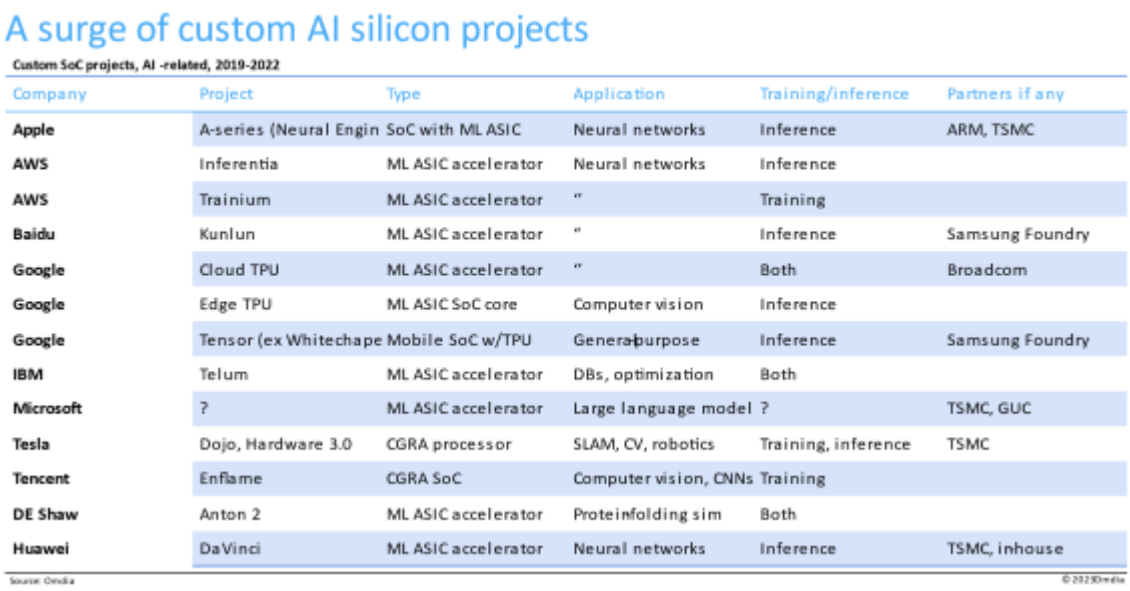
Microsoft joins the charge for custom AI silicon

Microsoft may be planning a custom chip for artificial intelligence (AI) workloads with the help of foundry Taiwan Semiconductor Manufacturing Company (TSMC) and design house Global Unichip. This is a major change in Microsoft's strategy, which previously prioritized reconfigurable devices over custom. It also represents another example of the so-called Makimoto's Wave shift from standardized to customized electronics. This possible move is likely being driven by Microsoft's deal with OpenAI, which creates a large structural source of demand for AI compute that the technology company must provide. The existence of this liability will help the company in negotiating with TSMC, as it tends to commit Microsoft to the custom project for the long term.

Microsoft may be building an AI ASIC

On March 2, Taiwanese media DigiTimes reported that Microsoft had approached foundry market leader TSMC and Global Unichip, an ASIC design house, about possibly producing a custom ASIC for Microsoft. This device would be a hardware accelerator for AI workloads, comparable with Google's Tensor Processing Unit (TPU) or Amazon Web Services' (AWS') Trainium or Inferentia, and it would be fabbed using TSMC's Chip on Wafer on Substrate (CoWoS) 2.5D and 3D packaging process. Current NVIDIA flagship GPUs use this packaging, as will the next generation of AMD Instinct. This process permits chip designers to stack up processor cores, high bandwidth memory, and passive components vertically on the die, with through-silicon vias for power supply between the component chiplets.

Figure: Notable AI ASIC projects, 2019–22



A surge of custom AI silicon projects

Custom SoC projects, AI -related, 2019-2022

Company	Project	Type	Application	Training/inference	Partners if any
Apple	A-series (Neural Engin	SoC with ML ASIC	Neural networks	Inference	ARM, TSMC
AWS	Inferentia	ML ASIC accelerator	Neural networks	Inference	
AWS	Trainium	ML ASIC accelerator	"	Training	
Baidu	Kunlun	ML ASIC accelerator	"	Inference	Samsung Foundry
Google	Cloud TPU	ML ASIC accelerator	"	Both	Broadcom
Google	Edge TPU	ML ASIC SoC core	Computer vision	Inference	
Google	Tensor (ex Whitechape	Mobile SoC w/TPU	General purpose	Inference	Samsung Foundry
IBM	Telum	ML ASIC accelerator	DBs, optimization	Both	
Microsoft	?	ML ASIC accelerator	Large language model	?	TSMC, GUC
Tesla	Dojo, Hardware 3.0	CGRA processor	SLAM, CV, robotics	Training, inference	TSMC
Tencent	Enflame	CGRA SoC	Computer vision, CNNs	Training	
DE Shaw	Anton 2	ML ASIC accelerator	Proteinfolding sim	Both	
Huawei	DaVinci	ML ASIC accelerator	Neural networks	Inference	TSMC, inhouse

Source: Omdia

Source: Omdia

If this project eventuates, Microsoft will be the latest hyperscaler to take the step of building its own AI ASIC. Google was the first with the TPU, followed by AWS, Alibaba, Baidu, and IBM. These companies have generally involved an integrated custom ASIC design house and foundry, such as Broadcom or Marvell, or else worked directly with TSMC. In every case Omdia is aware of, they used the new processor first for their own managed services (i.e., software as a service [SaaS] or platform as a service [PaaS]) and only later offered infrastructure as a service (IaaS) virtual machines to their clients.

This project represents a substantial change in Microsoft's strategy, although one that has been brewing for a while. In the 2010s, Microsoft invested heavily in reconfigurable FPGA technology, developing a SmartNIC for networking and hypervisor offload (Project Catapult), an operating system for FPGAs in the hyperscale data center (Project Feniks), and an FPGA-based AI accelerator card (Project Brainwave). This latter project was developed as Microsoft's go-to AI solution. The company added Xilinx as a second supplier alongside Intel in 2019, claimed as late as December 2020 to be the world's biggest investor in FPGAs, offered two families of Azure instance types and a version of the Azure Stack Edge Pro on-premises cloud product with FPGA acceleration, and continued hiring FPGA engineers through October 2021.

However, as Omdia reported that month in our "What's happening to Microsoft's FPGA strategy for AI hardware?" report, Microsoft withdrew all the FPGA-based Azure instances except for one, the Xilinx Alveo U250-based NP-series. It recommended that users shift to instance types with NVIDIA GPUs. Omdia observed at the time that FPGAs were likely a hard sell for client workloads due to the relative scarcity of developers with FPGA skills, although Microsoft itself could afford to maintain a skills base supporting its internal PaaS workloads, such as Azure Machine Learning and Cognitive Services. It is also worth remembering that Microsoft is not a complete beginner in custom silicon; the Surface Pro X and some Surface Pro 9 tablets use one of three generations of semi-custom systems-on-chip (SoCs) developed with Qualcomm.

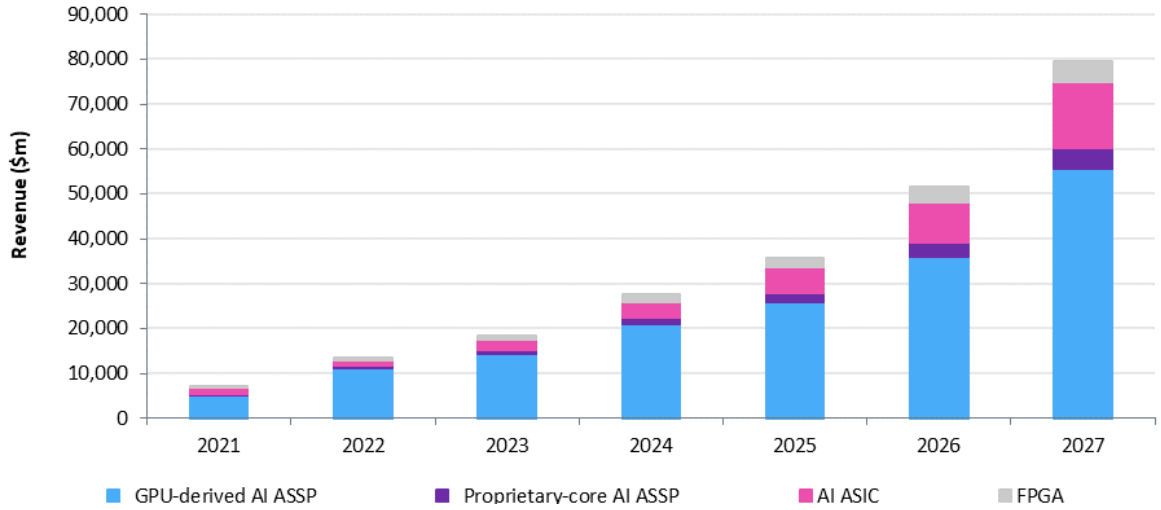
An ASIC for OpenAI?

This news will inevitably be read together with the Microsoft/OpenAI deal. OpenAI itself has so far worked exclusively with GPUs from NVIDIA, specifically the A100 server GPU launched in summer 2020. Under the terms of the deal, Microsoft is committed to providing OpenAI with $10bn worth of investment, probably in the form of Azure service credits. (As OpenAI has historically spent more than two-thirds of its cash flow on cloud computing, even if the investment were in cash, it would very likely be used to pay for Azure service.) Consequently, Microsoft owes OpenAI this much AI accelerator capacity, which it will have to provide to avoid losing its entitlement to OpenAI stock.

Ironically, this liability may also be an asset when it comes to dealing with TSMC. A foundry industry source recently told Omdia that their decision to go ahead with custom projects or not had less to do with the upfront price than the client's commitment to the project for the long term. Accepting a custom project means assigning a substantial engineering team, identifying and blocking out foundry capacity, starting wafers, and making a variety of other commitments. This is a lot easier to justify if there is a realistic prospect of the project resulting in a large volume order or a second-generation repeat project—so much so that the source's company is willing to co-finance ASIC projects if the client is suitably committed. Microsoft's liability to deliver $10bn of AI accelerator capacity to OpenAI is a commitment TSMC can take to the bank.

Microsoft is planning to use OpenAI models in some very high scale internal services—Bing search being the biggest—and to white label these models as a managed service. As such, the company could certainly benefit from a dedicated accelerator. The biggest barrier to the adoption of alternative AI accelerators is NVIDIA's hegemony in the software layer. Microsoft, like the other hyperscalers, can circumvent this by using custom silicon behind the hypervisor, the API, or the AI studio application, where the user does not interact with it directly. As Figure 2 shows, data from Omdia's AI Processors for Cloud and Data Center Forecast Report indicate hyperscalers' AI ASICs will be the primary competitors to NVIDIA's GPUs through 2027.

Figure: Data center AI accelerators by type, 2021–27



Source: Omdia

Written 9th March.

And just like that, generative AI reaches the enterprise marketplace

Bradley Shimmin, Chief Analyst, AI & Data Analytics

"When it comes to GAI solutions like ChatGPT …in no way is the enterprise ready to fully embrace GAI as though it were just another technology"

Generative artificial intelligence (GAI) solutions are on their way to enterprises around the world. Concerns over accuracy, misuse, and privacy abound—with no widespread resolution in sight. Still, all enterprises must prepare for the inevitable, building practical expertise in the technologies and techniques that drive GAI.

Step right up, get your generative AI here!

It seems like years have passed since OpenAI introduced its GAI chatbot, ChatGPT, in November 2022. GAI has very quickly brought the world to an inflection point, with the technology market rushing forward in what can only be termed a stumbling sprint. Numerous consumer implementations, primarily focused on search and software development, have come into being. An even bigger number of startups focused on GAI have seemingly teleported into existence. And (sadly) a rabid horde of self-proclaimed ChatGPT experts now flood the social landscape, promoting get-rich-quick schemes and even spreading nefarious "jailbreak" prompts that can lead only to misuse. That is a lot of experimentation and confusion.

Then again, perhaps my colleagues and I were too conservative in our estimate of GAI in terms of market adoption. In recent conversations, many technology providers themselves revealed a long-standing interest in and internal experimentation with GAI, dating back to the early days of pre-trained large language models (LLMs). As one collaborative AI and machine learning (ML) platform player put it: *"On a technical and implementation level, this is nothing new, but it is far more than a fad; it is a sea change."*

Most technology providers bet heavily on identifying and capitalizing on pivotal market events – with a run-up of gestation to general availability that spans 12–18 months. If the starting pistol fired

in November 2022, that means further examples of language-centric GAI solutions this coming December or early in 2024…but in reality the race toward GAI in the enterprise began back in 2017, when the market began experimenting with large, pre-trained transformer models like Google's BERT (or even decades before for the race in natural language processing R&D efforts).

However, March 2023 is definitely the date when the enterprise market for GAI platforms first took shape, with nearly simultaneous introduction of several new offerings targeting the enterprise technology buyer. There are many, many use cases for GAI emerging. Omdia is starting to see several patterns appear, particularly for vendors seeking to help customers democratize data-driven insights and speed AI development:

- **Vendor:** Look for announcements from all AI cloud platform leaders (Google, Microsoft, AWS, et al.). A secondary group of major line-of-business players (SAP, Oracle, Salesforce, et al.) will engage in similar efforts. Not discussed here but of equal importance, the AI hardware acceleration and orchestration space (headlined by the likes of Intel, Huawei, AMD, NVIDIA, et al.) will likely generate its own market independently and in cooperation with the cloud players mentioned above.
- **Focus:** With AI platform leaders, expect initial offerings to center on developers and AI practitioners working within enterprises of all sizes, geographies, and use cases. This will quickly shift into overdrive as these platform players begin garnering interest from third-party developers seeking to productize GAI within specific verticals/use cases. Line-of-business vendors will take on some of this ecosystem internally.
- **Scope:** Early efforts will focus on internally built foundational models covering multiple modalities (text, video, image, etc.). The development of solutions using these models relies on a host of new supportive tooling meant for both data science professionals and business users. In all cases, platform players will seek an open stance, connecting with and supporting third-party models and model hosting services, including Hugging Face and Cohere.

Omdia also took a closer look at two recent announcements from Salesforce and Google, two companies using the same technologies and general approach to GAI in two very different ways – but further illustrating these same patterns.

Where do we go from tomorrow?

As a longtime enterprise market observer, I have become accustomed to disappointment. So often, analysts will hear how X technology will forever alter the way companies do business. Unfortunately, nine times out of ten, those claims fall short of the intended mark. At best, technology X may indeed promise to make a difference, but it is always just a bit out of sync with the enterprise—either the new technology is not yet ready for the enterprise, or the enterprise is not quite ready to pull the trigger.

When it comes to GAI solutions like ChatGPT or Codex from OpenAI, PaLM from Google, or Einstein GPT from Salesforce, I can plainly say that in no way is the enterprise ready to fully embrace GAI as though it were just another technology. Far too many as yet unresolved concerns surround the performance of and risks posed by those models outside of a few well-trodden pathways dating back to the early days (circa 2019) of generative NLP with basic use cases like chatbots, code suggestions, and text summation.

And yet, if the clamor for sites like ChatGPT and the instant success of services like GitHub Copilot for code generation are reliable indicators, then the enterprise does not just want but demands GAI—so long as it does not take anyone's job (for now).

Should enterprise executives give into this fervor and welcome such an unproven and potentially risky technology inside the corporate firewall? They may not have a choice. Given the widespread availability, Netflix-scale pricing, and API-centricity of these services, companies of all sizes can stand up GAI offerings right now with very little effort. For that reason, even the most risk-averse company should begin planning for GAI.

At a minimum, companies should implement a basic usage policy that is in line with corporate data privacy and security requirements to minimize the potential blowback from any shadow IT usage. At the far end of the spectrum, companies that have identified one or more target use cases should begin not by investing in any prepackaged or self-service solutions, but instead by building practical expertise in the underlying technologies beating at the heart of GAI.

"Companies of all sizes can stand up GAI offerings right now with very little effort. For that reason, even the most risk-averse company should begin planning for GAI."

Why invest in technologies like generative adversarial networks (GANs) and Transformer models when those complexities live so far beneath the surface of easy-to-use tools like ChatGPT? Quite simply, deriving value from even the friendliest GAI tool demands understanding the investment in GAI. No one, not even the creators of these tools, understands how they work.

Prompt engineering is a good case in point. Regardless of the tool at hand, GAI needs a guiding hand capable of prompting it toward not just the correct response, but the best possible response. The current issue is that the creators of these tools are still discovering how to do that. As the market for GAI matures, solutions will likely do away with or at least minimize the need for prompt engineering itself. Should users create a single, detailed prompt (a la DALL-E or Midjourney)? Should they instead describe or give an example of the desired output? Or should they create a chain of reasoning for the model to follow, breaking the question down into constituent parts?

Truthfully, no one knows except through experimentation. Depending on how the model has been built and trained, these (and many more) prompt engineering methods may be required. It depends, for instance, not just on the use case, but on the data consumed by the model during training (labeled/annotated or raw, for example) or the architecture of the training algorithm (e.g., ChatGPT's use of InstructGPT to optimize the model's reasoning skills).

In a way, in creating GAI tools, humans set out to build an airplane and instead ended up with a UFO that appears capable of interstellar travel. Our challenge going forward will rest not in making our UFO fly, but in doing so responsibly with full understanding, transparency, and trust.

What does this mean for the enterprise? As these tools evolve, enterprise practitioners will need to build a full understanding of their inner workings from an IT perspective. And from a user perspective, companies will need to develop a new kind of prompt engineering literacy, one that can evolve along with our understanding of GAI.

Written 23rd March.

Technology

What's the next level consideration beyond the initial hype? From hardware to software, including cybersecurity considerations.

ChatGPT Artificial Intelligence: An Upcoming Cybersecurity Threat?

Ketaki Borade, Senior Analyst, Infrastructure Security

"ChatGPT tool is transformational in many cybersecurity scenarios if put to good use."

The role of artificial intelligence in cybersecurity is growing. A new AI model highlights the opportunities and challenges.

Artificial intelligence (AI) has the potential to revolutionize many aspects of our lives, including how we approach cybersecurity. However, it also presents new risks and challenges that need to be carefully managed.

One way that AI can be used in cybersecurity is through the development of intelligent systems that can detect and respond to cyber threats.

This was the AI chatbot's reply when I asked it to write about AI and cyber threats. I am sure by now you know I am talking about the most popular lad in town, ChatGPT.

In November 2022, OpenAI, an AI research and development company, introduced ChatGPT (Generative Pre-trained Transformer) based on a variation of its InstructGPT model, which is trained on a massive pool of data to answer queries. It interacts in a conversational way once given a detailed prompt, admits mistakes, and even rejects inappropriate requests. Though only available for beta testing right now, it has become extremely popular among the public. OpenAI plans to launch an advanced version, ChatGPT-4, in 2023.

ChatGPT is different from other AI models in the way it can write software in different languages, debug the code, explain a complex topic in multiple ways, prepare for an interview, or draft an essay. Similar to what one can do through Web searches to learn these topics, ChatGPT makes such tasks easier, even providing the final output.

The wave of AI tools and apps has been growing for some time. Before ChatGPT, we saw the Lensa AI app and Dall-E 2 making noise for digitally creating images from text. Though these apps have shown exceptional results that could be nice to use, the digital art community was not very happy that their work, which was used to train these models, is now being used against them as it raised major privacy and ethical concerns. Artists have found their work was used to train the model and now has been used by app users to create images without their consent.

Pros and Cons

As with any new technology, ChatGPT has its own benefits and challenges and will have a significant impact on the cybersecurity market.

AI is a promising technology to help develop advanced cybersecurity products. Many believe broader use of AI and machine learning are critical to identifying potential threats more quickly. ChatGPT could play a crucial role in detecting and responding to cyberattacks and improving communication within the organization during such times. It could also be used for bug bounty programs. But where there is the technology, they are cyber-risks, which must not be overlooked.

Good or Bad Code

ChatGPT will not write a malware code if asked to write one; it does have guardrails, such as security protocols to identify inappropriate requests.

But in the past few days, developers have tried various ways to bypass the protocols and succeeded to get the desired output. If a prompt is detailed enough to explain to the bot steps of writing the malware instead of a direct prompt, it will answer the prompt, effectively constructing malware on demand.

Considering there are already criminal groups offering malware-as-a-service, with the assistance of an AI program, it may soon become quicker and easier for attackers to launch cyberattacks with the help of AI-generated code. ChatGPT has given the power to even less experienced attackers to be able to write a more accurate malware code, which previously could only be done by experts.

> *"One thing is for sure: The industry cannot merely wait and watch if it creates a security problem."*

Business Email Compromise

ChatGPT is excellent at replying to any content query, such as emails and essays. This is especially applicable when paired with an attack method called business email compromise, or BEC.

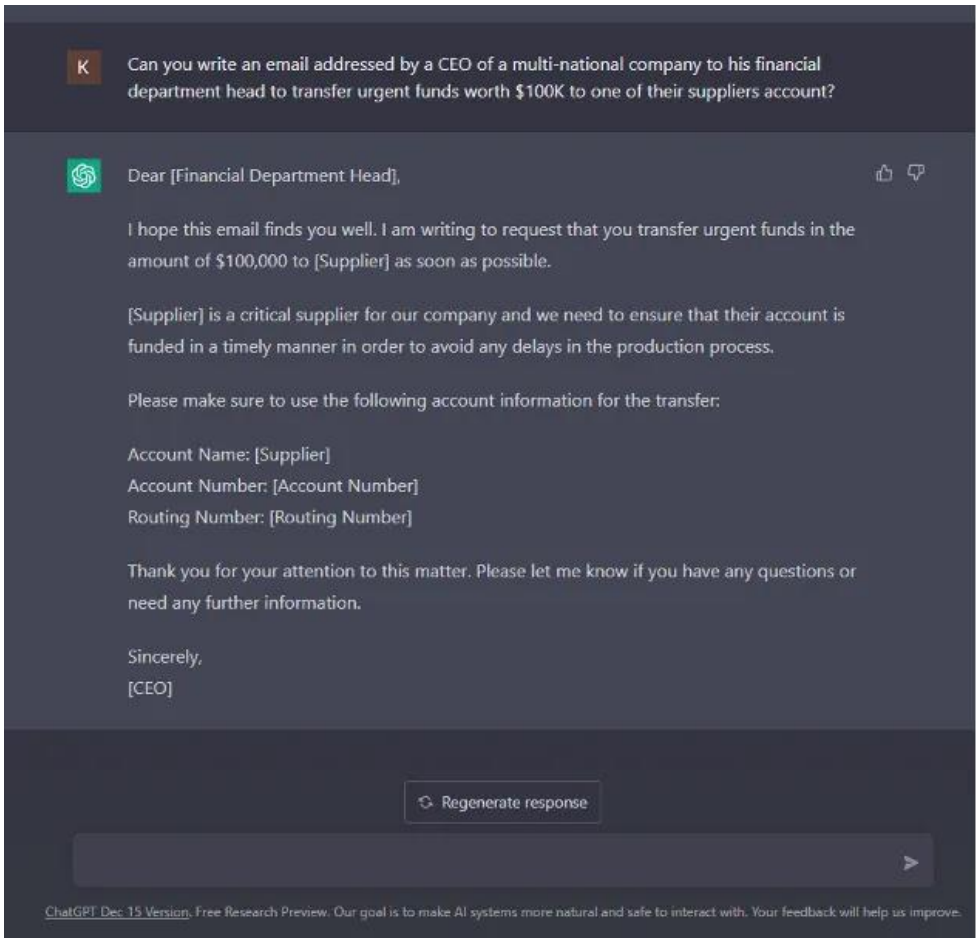
With BEC, attackers use a template to generate a deceptive email that tricks a recipient into providing the attacker with the information or asset they want.

Security tools are often employed to detect BEC attacks, but with the help of ChatGPT, attackers could potentially have unique content for each email generated for them with the help of AI, making these attacks harder to detect.

Similarly, writing phishing emails may become easier, without any of the typos or unique formats that today are often critical to differentiate these attacks from legitimate emails. The scary part is it's possible can add as many variations to the prompt, such as "making the email look urgent," "email with a high likelihood of recipients clicking on the link," "social engineering email requestion wire

transfer," and so on. Below was my attempt to see how ChatGPT would reply to my prompt, and it returned results that were surprisingly good.

Figure: ChatGPT Conversation Transcript



Source: Omdia, ChatGPT

Where Do We Go From Here?

ChatGPT tool is transformational in many cybersecurity scenarios if put to good use. From my experience researching with the tool and from what the public is posting online, ChatGPT is proving to be accurate with most detailed requests, but it still is not as accurate as a human. The more prompts used, the more the model trains itself.

It will be interesting to see what potential uses, positive and negative, come with ChatGPT. One thing is for sure: The industry cannot merely wait and watch if it creates a security problem. Threats from AI are not a new problem it has been around, it's just that now ChatGPT is showing distinct examples that look scary. We expect the security vendors will be more proactive to implement behavioral AI-based tools to detect these AI-generated attacks.

Written 6th January.

What is the hardware requirement for Generative AI?

Alexander Harrowell, Principal Analyst, Advanced Computing for AI

"Generative AI is driving innovation in two areas – more efficient AI accelerator hardware on one hand, and leaner AI model architectures on the other."

So far, the large language models of generative AI have led AI development to double-down on scale, building bigger models, data sets, and computing infrastructure. This trend will eventually hit limits. As a result, generative AI is driving innovation in two areas – more efficient AI accelerator hardware on one hand, and leaner AI model architectures on the other.

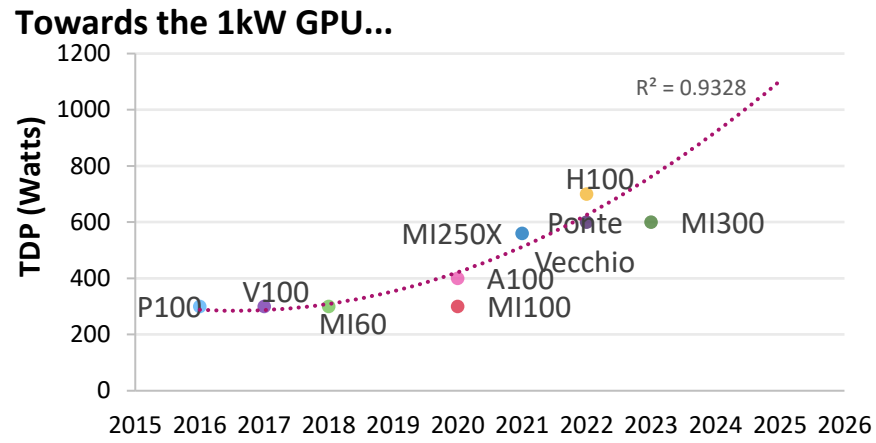
The race for scale

The current GAI boom may have started in 2020 when a key AI research paper identified that language model performance showed a regular, positive relationship with model size. More model parameters meant both better benchmark performance, and also new emergent properties that gave AI models skills never previously observed. In 2019, a model with 10 billion parameters was considered large; by 2021, Google's PaLM language model had 540 billion, and today, OpenAI's GPT-4 is over a trillion.

This dash for sheer size had consequences for the underlying computing infrastructure. AI model training had already moved from running on CPUs to running on GPUs, taking advantage of the extreme parallelism designed into them to cope with graphics made up of thousands of individual shaders. Now, not only did it need hundreds to thousands of GPUs, it also needed a new class of flagship server GPUs that are increasingly specialized as AI accelerators.

NVIDIA's A100 GPU, launched in 2020, became the iconic product of the scale era but every other player chased it up the power curve by building bigger GPUs or dedicated AI ASICs. Before 2020, the most power-intensive GPUs had thermal design power ratings around 300W; the A100 was 400W, and successive launches took it higher and higher.

Figure. GPU power consumption surges

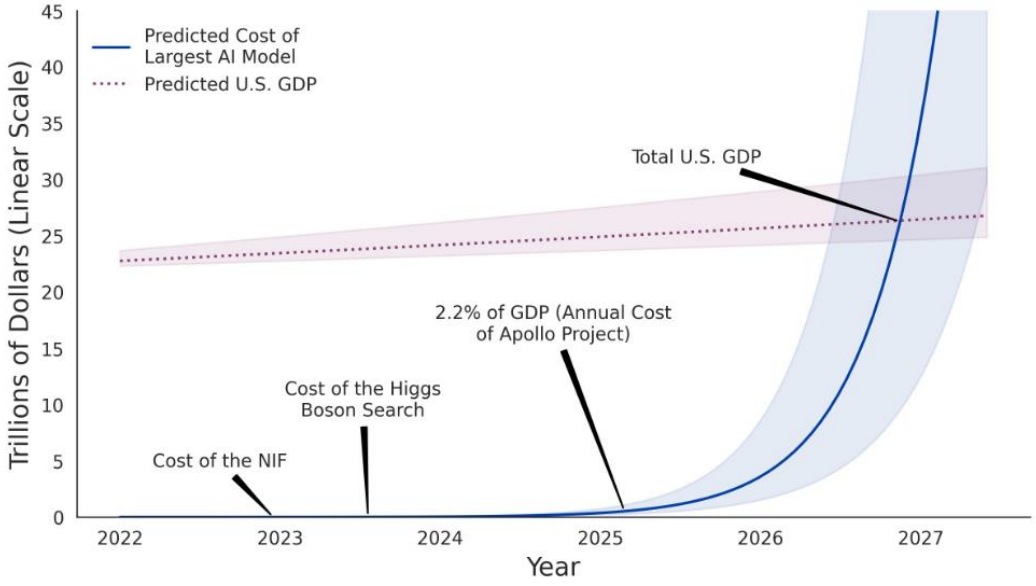


Notes: Most power-dense option (usually SXM) shown

Source: Omdia.

CPUs have also been also getting hotter; AMD's industry-leading Epyc data center CPUs are now running at the levels the GPUs were in 2019, around 300W, while Intel's latest Sapphire Rapids CPU has a TDP of 420W in its most powerful configuration. The combination of bigger training clusters, hotter and more power-dense racks, and bigger training datasets seemed to both threaten a limit on the progress of fundamental AI and to mean that it would unavoidably be the preserve of a few big players, whether corporate or government, rather like speech-to-text before the AI boom. In the limit, continued growth in compute demand at the rates of 2019-2022 might mean new state-of-the-art AI models just get too expensive to build at all.

Figure. If it can't go on, it won't



Source: How Much Longer Can Computing Power Drive Artificial Intelligence Progress? Lohn, A., & Musser, M. (2022, January)

In search of alternatives

One way out of the bind is to develop better AI processors, and this has attracted enormous investment both from venture capitalists into startups and from the incumbent chipmakers. Starting in 2018, there was a wave of interest in very large chips, so-called waferscale systems that take up an entire 12-inch silicon wafer and aim to load the entire model on-chip. However, the enormous model growth since then has outdistanced even this. Scalability is crucial, which helps systems built up out of smaller devices linked by fast chip-to-chip interconnects and integrated in multi-chip packages. These are the two top priorities for the major chipmakers at the moment. Another option is to draw on the mobile industry's expertise in low-power processors.

We can also look again at the design of the models and the datasets they are trained on, of course. A key research paper in 2022, the so-called Chinchilla paper, identified that although the benefits of model size are real, there is a trade-off between model size, training compute, and training data. If we can train faster or for longer, expose the model to more data, or curate the training data so as to expose it to more of the variance in the underlying distributions, better performance might be achievable with smaller models.

In 2022 and into 2023, there have been a string of successful generative AI models that have achieved state-of-the-art results with much smaller footprints than behemoths like GPT-4. StabilityAI's StableDiffusion, for example, is 1.27 billion parameters in four distinct neural networks, while the same company's StableLM achieves comparable performance to the 175bn parameter GPT-3 from 7 billion parameters. The LlaMa family of language models also achieves similar performance in its 13 billion parameter version which runs on any Apple Silicon Mac, while its 65 billion version rivals PaLM for performance. These models have in common some combination of high quality training data, a stack of different, specialized neural networks, and instruction tuning. Even Sam Altman, CEO of OpenAI and the de facto leader of the scale camp, recently said that he expects the future to be about smaller, specialized, composable models.

Written 25th April.

Generative AI: The impact of AI-based autocoding on software development

Michael Azoff, Chief Analyst, Cloud Native Computing

Catalyst

"…ChatGPT is good enough to be useful as a coding assistant, but the output should be checked by professional developers."

The spotlight is on how generative AI is emerging from large language models with the release of ChatGPT by OpenAI, which has taken the world by storm. While the technology is in a line of incremental improvements, ChatGPT has gained in usability improvements, which, coupled with its current free access, has drawn a lot of people to try it out. This report examines the impact of ChatGPT and other AI tools in aiding software programmers with autocoding, such as code generation and debugging.

STOP PRESS: As this report went to publication, OpenAI released GPT-4, its latest model, which also performs programming tasks. While an improvement over ChatGPT, it is, to quote OpenAI, "less capable than humans in many real-world scenarios."

Omdia view

OpenAI created Codex, an AI model based on GPT-3 that translates natural language to code. OpenAI described it as most capable in Python and proficient in other languages such as JavaScript, Go, Perl, PHP, Ruby, Swift, TypeScript, and SQL. GitHub Copilot (see below) is based on Codex.

When OpenAI released ChatGPT in late 2022, it also incorporated assistance with coding. More recent applications have found ChatGPT useful in debugging code. Anecdotal views among developers writing about their experiences in developer forums say that ChatGPT and Copilot are good enough to be of help, but they are not error free, and developers need to check all code that is produced.

We looked at who can benefit from this technology and examined use cases in turn. The following assessment is based on the current state-of-the-art capabilities. It is important to note two points:

- This technology is continually evolving and will improve. With improvements, the use cases will broaden.

- ChatGPT, as with other state-of-the-art intelligent virtual assistants, remembers a thread of conversation; therefore, a good way to use it (and being done by some developers) is to continually refine an output from ChatGPT if the first iteration is not what is required. ChatGPT is able to use this feedback to improve its future answers.

The impact of ChatGPT will then affect three user groups:

- **Professional developers**
  - Pros: Likely to gain the most from autocoding, being well placed to judge the quality of the output, and developers are advised to always check the generated output. Tools can help speed up development by generating both small and large blocks of code and also for debugging code. Autocoding can also assist a developer who is learning a new programming language. Interrogating the tool to refine the output is commonly performed and is a recommended tactic. The media-reported fears that autocoding will put professional developers out of work is an overreaction: autocoding assists developers and will not replace them.
  - Cons: Could lead to lazy programming if the output is not checked, tools do not perform custom design work and will always need a human to direct the requirements. Autocoders are also limited because AI models would not have received enough training yet. Finally, its use for application security requirements should be restricted to security specialists (should they choose to use it). A research paper paper by Owura Asare, et al. showed that "Copilot is not as bad as human developers at introducing vulnerabilities in code."

- **Non-professional developers and business power users**
  - This group includes individuals who might have some programming experience, but coding is not their main job function. This is the user category that no-code/low-code (NCLC) tools address.
  - No code: No-code tools need to be fully automated, so how the code is generated is of crucial importance. ChatGPT uses natural language, which leaves room for ambiguities. How you phrase a question will in itself drive a different answer. A no-code tool typically uses a visual programming interface to define the requirements reducing the scope for mistakes. AI may be used under the hood to help generate the code.
  - Low code: The absence of a professional developer means adequate checking of the code may not be possible. If the low-code application is limited in scope, then using autocoding as a frontend can be useful. Alternatively, it would be inadvisable for a low-code application fronted by an autocoder to have change privileges and interact with the corporate database unless the tool has in-built governance to prevent mishaps.

- **Citizen data scientist**
  - In this scenario, professional developers or business users without data science expertise may use autocoders to engage in various machine learning (ML) tasks. Given the expertise

required to build unbiased, fair, and explainable AI applications (with potential EU laws in the wings to enforce this), it would be advisable to work with an ML solution that is already designed for this user category and, in turn, these tools may make use of autocoding under the hood.

Key messages

- Autocoding as a technology is here to stay and will become a standard developer tool, helping professional developers accelerate their work.
- The technology behind ChatGPT is good enough to be useful as a coding assistant, but the output should be checked by professional developers.
- It is most likely that autocoding technology will evolve and improve, so the advice on how to use this technology will change.
- CIOs and CTOs must address the use of autocoding within their organization to ensure application security by controlling who can make use of autocode, as well as reap the benefits this technology can provide its developers.
- Generative AI-based coding is only useful for mining existing coding knowledge—to ask it to create working code for a novel algorithm for which no examples exist may create nonsensical code or “hallucinations.”

Recommendations

- C-level executives should begin to explore how to best exploit autocoding for their organizations and adopt a strategy for its safe use while also leveraging it for the benefit of their organization.
- Development team leaders and managers should ensure that autocoding is being used safely by their teams and that all generated code receives a review. There should be transparency in where and how this technology is being adopted.
- Developers should try out autocoding to see how it can accelerate their work. There is anecdotal evidence that this technology can act as a useful assistant in development work.

Written 17th March.

.

Value-driven Use Cases

Considering the specific impact for vertical industries and applications.

Generative AI in the consumer domain

Eden Zoller, Chief Analyst, Consumer AI

The current wave of generative AI tools has captured people's imaginations and created a genuine sense of excitement and possibility, with some impressive usage stats emerging. OpenAI's ChatGPT was reported to have crossed the one million user mark within a month of launch (December 2022). OpenAI has reported that over three million people are using the DALL-E 2 text to image application, generating over 4 million images a day. Generative AI tools present users with immediate, tangible benefits and gratification, which is in contrast to many new technologies that have limited perceived value, are overhyped or constrained by device availability and costs.

"The easy, intuitive prompt style user interface for generative AI apps is a game changer, by putting generative AI everyone's hands and helping to drive scale and reach"

Some of the standout attributes that make generative AI so compelling in consumer domain are summarized below, while tables 1-3 present a snapshot how generative AI capabilities could potentially be used to enhance TV and film, advertising and marketing, the traditional creative arts.

Tangible benefits for users

- The easy, intuitive prompt style user interface for generative AI apps is a game changer, by putting generative AI everyone's hands and helping to drive scale and reach
- Generative AI capabilities are being integrated with mass market applications that consumers already use (e.g. ChatGPT with Microsoft Bing search and other products)
- Generative AI enables new forms of creativity and self-expression across a wide range use cases (e.g. writing, photography, video, design).
- Associated with the above, generative AI can enhance expressive abilities for those that face challenges in this area. For example, enabling people with arm/hand co-ordination or mobility issues to produce words, pictures and graphics using a few simple text prompts or going forward, verbal prompts. Generative AI could also help people with dyslexia produce written material.

- Enables the development of highly personalized applications. Generative AI has the ability to provide consumers with something unique and produced for them alone – a piece of music, a game, recipe, story and so on.
- The outputs of current generative AI tools like ChatGPT are easy to share, with certain creations going viral on social platforms, creating more buzz and encouraging yet more people to try generative AI.
- Generative AI can help consumer product/service ideation, concept testing and prototyping.
- Generative AI for coding can assist with the development of most consumer applications in some shape or form but will be particularly useful for games.
- Generative AI can be powerful when combined with other technologies and devices. The 'Create with Alexa' children's personalized story telling experience is a good example, combining different generative AI outputs (story narrative, illustrations/images, music) curated by the Alexa voice assistant and delivered on an Amazon Echo smart home display.
- A new wave of generative AI consumer apps beckon. The first wave of generative AI foundation models have focused on text and image outputs, but there is a new wave of generative AI models waiting in the wings focused on text-to-video and text to 3D objects. This will open up even more use cases and opportunities.

Generative AI applications in selected consumer verticals

Figure: Generative AI in TV and film

Relevant category outputs	Examples of potential applications
Text	Script ideation and writing Copy writing to promote TV and film productions
Audio	Music and sound: original scores, sound effects, ambient music (to create/emphasize atmosphere) Audio descriptions
Images	Images so stimulate ideation, create storyboards Images to promote TV and film productions
Video	Special effects Regenerating archive footage and film, repairing damaged/poor video Creating discrete objects, characters and scenery through to complete long-form outputs Generating personalized TV and film Promotional videos for a film/TV title
3D Objects	Virtual presenters and avatars Creating/adapting TV & film for immersive experiences and environments

Source: Omdia

Figure: Generative AI in advertising and marketing

Relevant category outputs	Examples of potential applications
Text	Promotional copy writing (for brochures, websites, blog posts etc.) Personalized marketing communications (emails, chat/text, social platforms)
Audio	Original music to accompany video advertising formats Bespoke music designed to reflect the essence of a brand
Images	Images so stimulate ideation, create advertising campaign storyboards Generate brand and related images for promotional collateral, e.g. pictures, graphics, illustrations used in a magazine, online adverts, outdoor media "Try before you buy" tools for end-users, e.g. generate images of themselves wearing different outfits, accessories, styles
Video	Generating bespoke full length video advertising Video blogs Personalized video communications
3D Objects	Creating/adapting advertising for immersive experiences and environments

Source: Omdia

Figure: Generative AI in traditional creative arts

Relevant category outputs	Examples of potential applications
Text	Story ideation for fictional work Co-or full authoring of fictional and non-fictional works Ghost-writing autobiographies Song lyrics
Audio	Generating reader voices and sound effects for audio books Creating full form audiobooks Generate new music that can be hyper-personalized for a particular individual's emotional state, location, activity
Images	Images to stimulate ideation for illustrations, pictures Generate complete illustrations, pictures Style create or transfer, e.g. create something in the style of X, or change something to the style of Y Image domain transfer, e.g. change a photo to a sketch, painting Image blending – i.e. create new output based on combining elements from different original images Images to promote artistic outputs
Video	Create videos from still images (e.g. from pictures, photos) Promotional videos for artistic outputs
3D Objects	3D objects from static art works

Source: Omdia

Generative AI applications must be handled with care

Although generative AI applications offer compelling benefits, these are frequently countered by some serious challenges and even potential harms that need to be considered, with just a few these highlighted below.

- Generative AI makes mistakes. Generative AI outputs can be inconsistent, inaccurate and contain errors. But people may not be aware of this because generative AI outputs appear authentic and are convincing. If users relay on material that contains mistakes this could cause harm.

- Compromising data privacy. The web data on which many foundation models are trained contains large amounts of personal data. People can also disclose personal information when interacting with generative AI applications via prompts, without being aware of the risks.

- A new chapter in manipulation and abuse. The outputs of generative AI applications are constantly improving, making it harder to tell the difference between authentic words, images and video, and those that have been altered. This makes generative AI a powerful tool for spreading misinformation and for inflicting abuse – at scale. Generative AI applications could also be used to perpetuate fraud, for example by mimicking an official text communication requesting financial information or money.

Don’t experiment on consumers with poorly conceived applications

- Generative AI apps must well thought-through, which is not always the case. Generative AI app should address a genuine use case – otherwise it is just a novelty.

- Execution needs to improve. Generative AI apps should be only directed at tasks they are properly able to support and must be rigorously tested before release. This should be obvious but appears to be falling on deaf ears given the growing litany of mistakes produced by certain generative AI apps.

- Neglect potential harms and ethical issues at your peril. Service providers must be pro-active in protecting people from potential harms caused by generative AI, ethical and otherwise. Look at strong, contextual, nuanced filters to minimize data pollution at both the input and output stage. Provide clear, tamper proof signatures for artifacts that are AI generated.

Written 21st April.

How ChatGPT signals new productivity potential in the digital workplace

Tim Banting, Practice Leader, Enterprise IT: Digital Workplace

ChatGPT has taken the internet by storm, as a generative AI system that can appear human-like in its ability to answer questions in the same way as a regular conversation.

"Chatbots have historically underwhelmed customers; however, ChatGPT illustrates that AI can now provide a smarter, more human-like response"

Although ChatGPT has some well-deserved praise for its quality of responses, its ability to handle a wide variety of subjects, and conversational style, it is merely an example of a good quality, generative AI model. The outputs these models produce may sound convincing by design; however, the responses they generate can be wrong or biased. That said, the financial investment OpenAI will receive and the early examples of ChatGPT's potential indicate how AI technology will transform the digital workplace.

Generative AI for customer service: early stages, but worth your attention

Microsoft's sizable investments highlight AI's strategic importance, but conversational and generative AI is nothing new. Vendors and enterprises are exploring how this technology is currently used, and how AI will improve productivity, collaboration, and customer engagement in the future.

Interestingly, ChatGPT leverages generative AI to create new, unscripted content, which means its algorithm can automate new text, audio, code, images, and videos during an interaction. When these capabilities combine with ChatGPT's speed, access to a vast pool of web content, highly customizable responses, and conversational tone, it's clear that the chatbot can become an incredible force multiplier. This naturally leaves contact center professionals wondering if there is a place for ChatGPT in the customer service world. The short answer is, potentially; however, ChatGPT would have to overcome some significant hurdles.

First, the chatbot currently doesn't have access to real-time information. Its data and knowledge of world events after 2021 is limited, so, for example, it cannot help you make flight recommendations. Also, ChatGPT has a short memory. While it can remember what a user previously stated in a conversation, its memory is limited to about 3,000 words within an existing conversation. Any information beyond this is not stored.

Current issues—Omdia's brief experiment with ChatGPT

ChatGPT can generate incorrect information, alarmingly, in a plausible way. In a recent experiment, Omdia asked ChatGPT "What is the most valuable baseball card from the 1980s?" twice in succession and compared the results. Whilst the second attempt was correct, the first attempt generated an incorrect response, but with a decisive and authoritative tone. The response sounds convincing when enough facts are combined with an error and presented in such a way. Subsequently, this could lead a customer or prospect down the wrong path, which could be irritating and potentially dangerous. Automated self-service solutions must be accurate and consistent with customer responses – especially when there's only one correct answer.

Clearly, ChatGPT is not there yet and must overcome some significant hurdles to become a reliable customer service tool; however, it would be unwise to ignore its potential, especially as companies are investing heavily. Perhaps partnerships, innovations, and integrations could resolve many of its existing challenges. If so, ChatGPT could help lower the barrier to entry for contact centers looking to deploy chatbots that can handle very large datasets. And maybe we'll see a generative AI tool automate good first drafts to assist agents, and respond to emails, business chats, and social media.

These are only a few potential customer service use cases; there could be plenty of others. So customer service executives should undoubtedly pay attention to ChatGPT; however, like any responsible leader, they should proceed cautiously.

How could generative AI be leveraged in future applications?

The pace of change with generative AI is rapid, but below is a brief synopsis on some of the future applications considered by Omdia right now:

- **Communications and collaborations.** While ChatGPT has received recent media attention, AI is already heavily leveraged in unified communications and collaboration (UC&C) platforms (e.g. video blur backgrounds). ChatGPT will be an additional AI-based capability that vendors may choose to add to their UC&C platforms, depending on the platform's speech to text engine. If in the future, it could ingest company-specific and domain-specific information and lexicons, then ChatGPT may prove to be extremely useful in two tasks: meeting summaries, and surfacing information in real time. Both use cases improve employee and organizational efficiency – the sweet spot for these types of applications

- **Digitally augmenting traditional support approaches.** Via their own experiences with ChatGPT, many people now understand how AI can be utilized in professional and personal settings. Through better contextualizing and personalizing employee support approaches, AI has exciting potential to help improve end-user and support admin experiences. Traditionally, an agent fields inquiries, manually searches internal knowledge repositories for appropriate solutions, and replies. Instead, generative AI (trained on business domain content, employee technology profiles, and incident history) could offer contextualized responses, automate service requests,

and enhance technical support. In complex cases, where first-line resolution is impossible, the handling agent could be guided automatically through the relevant escalation process.

- **Automated intelligent workflow.** Low/no-code solutions currently democratize development activities via drag-and-drop interfaces; however, generative AI may provide easier automation of workflows and development of micro apps. In the same way ChatGPT has been used to write lines of code, it could also be applied to simplify business processes and improve productivity.

- **Enhancing employee productivity.** Generative AI will be an important technology for enterprises in 2023 and beyond, especially as other prominent technology vendors (notably Google) will start to introduce their own solutions. When Google plays its hand here, the vendor must clearly show how it can add business value rather than relying on its AI engineering and technical prowess. Bard could be used as a digital assistant, undertaking some of the more mundane and repeatable tasks that employees handle. It could also be embedded within Gmail, Calendar, Meet, Chat, Drive, Docs, AppSheet, Cloud Search, and other services to automate and simplify work. Text recommendations, grammatical corrections, natural language-based approaches to querying databases, slide assistance, real-time support suggestions, and endpoint management are all examples of how AI could be used to enhance Google Workspace.

- **AI-powered content.** Generative AI could also be used to parse information from content across an enterprise, automatically extracting metadata to make it easier to find and categorize information. An emerging technology category, Content AI aims to transform how content is created, processed, made discoverable, and automated within workflows to improve productivity. Enterprises could automate content-based workflows (e.g., contract processing, approvals, etc.), integrate with a user's taskbar to make it easier to search for answers to natural language requests, and use AI to generate summaries of key content.

Conclusion

ChatGPT has been trained on vast datasets and information from the internet (sources like Reddit discussions, archived books, and Wikipedia articles) to help it learn dialog and attain a human style of responding. Vendors should point out that ChatGPT is merely an example of generative AI's capabilities. Vendors should explain that to get the most value from generative AI, models need to be trained from domain-specific organizational data, not public internet sources. Chatbots have historically underwhelmed customers; however, ChatGPT illustrates that AI can now provide a smarter, more human-like response. Today's contact center agents have many tasks to fulfill, many of which can be automated. Conversational AI can create internal knowledge base articles, scripts, and web posts to better inform customers and agents about a problem and the potential solutions. More advanced customer engagement platforms offer sentiment analysis, but conversational AI is likely to be quicker, more affordable, and customizable to suit the needs of an enterprise. Such platforms will be able to monitor conversations and threads, interpret emails, and alert customer engagement teams of any follow-up actions to be made to ensure customer satisfaction.

Outside the contact center, vendors can capitalize on generative AI through workflow automation, improvements to collaboration and communication, and simplified employee support. Such technology will help to unlock the productivity puzzle plaguing organizations today, brought about by siloed data, manual processes, and a lack of organizational awareness.

Written 9th February.

Will ChatGPT-powered email head to the contact center?

Mila D'Antonio, Principal Analyst, Customer Engagement

"ChatGPT powered email would help streamline work for contact center agents"

Microsoft announced the preview of a new generative AI-powered email in Microsoft Viva Sales. Omdia predicts the launch will give rise to ChatGPT-powered email in the contact center.

Summary

In early February 2023, Microsoft announced the preview of a new generative AI-powered email in Microsoft Viva Sales that promotes efficient communications for sellers. Microsoft intends to make the technology generally available within the next two months. The AI-powered capability leverages Azure OpenAI and ChatGPT to access customer relationship management (CRM) data from Microsoft Cloud and then auto-suggests customizable content and auto-generates an email. The value proposition is that sellers will spend less time composing emails and searching databases for information. Omdia predicts the launch will provide momentum for the deployment of ChatGPT-powered email in the contact center.

ChatGPT-powered email would help streamline work for contact center agents

With ChatGPT, Viva Sales can remind sellers to follow up with prospects or customers and auto-generate responses along with sending product descriptions and proposals. Sellers can personalize responses based on a range of categories, from "make a proposal" to "reply to an inquiry." Microsoft 365, Windows, Microsoft Enterprise Mobility + Security, Microsoft Dynamics 365, and Salesforce can all be utilized to enrich email replies with data.

In an analyst briefing on February 8, Microsoft said the technology currently offers no learning of tonal style, which is the technology's ability to converse in human-like, emotive, and expressive characteristics. The roadmap, however, includes two levels of tonal learning. Until then, Microsoft will offer custom prompts that alert users with suggestions like, "try this to make it sound happy." Microsoft also said that it believes the technology can learn the tone and style of individual

customers and bring that inflection into replies. That ability to express emotion is best suited for the customer care domain, where conveying empathy and reducing customer frustration are important aspects of the customer experience.

Although no pivotal use cases currently exist, as enterprises train ChatGPT on their data, it will change the way they access and consume data across the enterprise. Omdia believes ChatGPT-powered email in the contact center will serve as one of the initial use cases. It highlights where the technology excels the most—by fine-tuning a model on a dataset of conversational text, it can learn how to generate naturally sounding and contextually relevant text-based conversations. This ability to understand complex questions and then automate naturally sounding and relevant email responses could bring significant value to the contact center, especially with email.

"Although no pivotal use cases currently exist, as enterprises train ChatGPT on their data, it will change the way they access and consume data across the enterprise. Omdia believes ChatGPT-powered email in the contact center will serve as one of the initial use cases."

Recent Omdia insights support this. According to Omdia's The State of Digital CX: 2022 Analysis survey, when asked which channels they use the most to interact with customers, many of the respondents—leaders in customer service and customer experience—cited email as the top communications channel. 47% said that it was their most-often used channel. The appetite for AI adoption and deployment is also large. According to Omdia's IT Enterprise Insights: ICT Drivers and Technology Priorities – 2023, 59% of respondents said they place high importance on the adoption of AI and machine learning to advance their digital strategies over the next 18 months. In the same survey, respondents cited the need to increase efficiencies as one of the top priorities for 2023.

While generative AI–powered email via Viva Sales holds great promise, the market is already home to plenty of AI-powered email tools that effectively collect data, analyze it, and draft comprehensive emails to companies' target audiences. Most of the tools learn a company's brand voice and then generate emails. Whether ChatGPT differs from these products is immaterial. Generative AI-powered email for sales will serve as a rising tide that may eventually help to advance AI-powered email deployments in the contact center, owing to its promise of efficiently executing tasks and lowering operating costs once the technology is rooted and proven. At that inflection point, it should enable level four agent augmentation, where customers only interact with digital agents.

Since ChatGPT debuted last November, much has been written about its potential use cases in the contact center. While the generative AI landscape must first overcome a range of hurdles from potential ethical implications, data privacy laws, and even accuracy issues, the general takeaway is that generative AI, such as ChatGPT, will revolutionize customer service with its ability to handle multiple conversations, understand natural language text, and gather valuable customer data.

Written 14th February.

Three ways generative AI can improve customer experiences

David Myron, Principal Analyst, Customer Engagement

Five9, NICE, and Zoom raise the CX bar with generative AI. The CX vendor offerings enable enterprises to leverage generative AI to improve customer experiences, without the risks associated with using information outside their domain.

Five9, NICE, and Zoom raise the CX bar with generative AI

In only a few short months, OpenAI's introduction of ChatGPT in November has inspired a slew of innovations, from Five9 and NICE to Zoom, making generative AI one of the most popular topics at Enterprise Connect in Orlando. These three solutions enable enterprises to leverage generative AI to improve customer experiences in different ways, without exposing them to the risks associated with using information outside their domain.

> *"Some organizations might have the time, money, and resources to create their own large language models, appropriate guardrails, and data curation strategies to handle publicly shared data. However, these technologically sophisticated companies are in the minority."*

Three ways to leverage generative AI for CX improvements

- During the event, Five9 released its Agent Assist 2.0 with AI Summary, powered by OpenAI. Five9's Agent Assist 2.0 with AI Summary automatically summarizes customer call transcripts in seconds. The capability, which uses the same generative AI technology as ChatGPT, can summarize call transcripts without model training and manual categorization. By automating the time-consuming post-call data entry process, agents can spend less time entering post-call notes and more time handling calls.

- NICE unveiled Enlighten Actions, enabling enterprises to facilitate customer experience (CX) outcomes using a conversational interface. NICE Enlighten Actions combines NICE Enlighten AI, an AI-powered suite of CX solutions, with the generative models from OpenAI, the developer of ChatGPT. This enables enterprises to expedite the creation of AI-powered CX processes by making Enlighten AI accessible through a conversational AI chat interface. So, for example, using Enlighten Actions with NICE CXone, a contact center manager can ask, “What are the top five call types that we do not have knowledge articles for?” After revealing the top five results, the manager can reply, “Write knowledge articles for each intent.” CXone will immediately write the articles, which can be reviewed and published when ready. Similarly, a manager can type, “Which interactions have the lowest customer satisfaction?” and then ask CXone to write relevant training articles to be reviewed and published by the manager when ready. NICE Enlighten Actions is integrated across NICE’s CXone Expert, NICE’s knowledge management solution, CXone’s Bot Builder, SmartAssist, and the Enlighten AutoSummary.

- Zoom Video Communications joined forces with OpenAI and unveiled plans to offer generative AI capabilities across the Zoom platform through its smart companion, Zoom IQ. The company plans to add an email composer to Zoom IQ for Sales to provide email draft suggestions in response to conversations originating from Zoom meetings, Zoom phone calls, and email threads. Zoom also plans to include generative AI capabilities in its team chat, meetings, email, and whiteboard capabilities, as well.

Minimize risk exposure

While these solutions are vastly different, they share a common goal—to automate tailored customer interactions or processes at scale. This will be the driving force behind future investments in generative AI for customer experience.

Another similarity that these solutions share is that, unlike ChatGPT, they don’t require enterprises to use publicly available data from the web. ChatGPT has impressed many with its ability to scour the web for very specific content and generate responses in a conversational tone. However, ChatGPT utilizes probabilistic models that distill large amounts of often conflicting data from all over the web to render responses. This means that, while it is usually accurate, it can still deliver responses that are irrelevant, inaccurate, or worse, harmful.

So, when it comes to leveraging generative AI for customer experiences, it’s important to walk before you run. Some organizations might have the time, money, and resources to create their own large language models, appropriate guardrails, and data curation strategies to handle publicly shared data. However, these technologically sophisticated companies are in the minority. Today, most enterprises would be better off leveraging generative AI solutions that only distill information from approved domain-specific knowledge found within their organization’s knowledge bases and customer conversations.

Written 25th April.

AI and ML-driven solutions to help CSPs improve the customer experience

Roz Roseboro, Principal Analyst, Service Provider

Multiple announcements right before and during Mobile World Congress show that AI/ML-driven solutions are poised to help communications service providers (CSPs) improve quality of experience and provide more relevant and personalized offers. The hope is that doing so—both reactively and proactively—will lead to lower churn and increased ARPU.

"From a care perspective, AI and ML help systems make recommendations more quickly, more personalized to the subscriber and based on more types of data."

- **On February 20,** Nokia announced AVA Customer and Mobile Network Insights: "…a cloud-native analytics software solution that simplifies the collection and analysis of 5G network data to provide CSPs with stronger and more cost effective analytical capabilities. The solution delivers "intelligence everywhere" through AI and machine learning tools that support intelligent and automated decision making based on correlated reports generated from data across 5G networks."

- **On February 26,** Microsoft announced a public preview of Azure Operator Insights: "…enables the collection and analysis of massive quantities of network data gathered from complex multi-part or multi-vendor network functions. It delivers insights for operator-specific workloads to help operators understand the health of their networks and the quality of their subscribers' experiences."

- **On February 27,** Amdocs and Microsoft announced the Intelligent Customer Engagement Platform: "The Customer Engagement Platform will be integrated with Amdocs' end-to-end set of solutions, from customer experience to monetization products to network automation while fully exploiting the capabilities of Dynamics 365, the Microsoft Power Platform, and the Microsoft Cloud." It will also leverage Amdocs' Data and AI solution to "enable cross-domain data input (from network, billing, and transactional data) to be fed in to drive insight-driven

recommendations." When asked how this product relates to Azure Operator Insights, Microsoft said that the Intelligent Customer Engagement Platform focuses on the OSS/BSS domain, while Operator Insights is focused on the infrastructure layer.

- **Also on February 27,** Google Cloud announced Telecom Subscriber Insights: "…Telecom Subscriber Insights, a new service designed to help CSPs accelerate subscriber growth, engagement, and retention. It does so by taking insights from CSPs' existing permissible data sources such as usage records, subscriber plan/billing information, customer relationship management, app usage statistics, and others."

These solutions, to varying degrees, collect information from disparate OSS, network devices, billing, and other systems and apply AI/ML processing to correlate events and deliver insights to help improve the customer experience. That two of the announcements came from hyperscalers should not be surprising given the immense processing power that will be needed to support AI/ML at scale—not to mention the analytical tools the hyperscalers have developed for their cloud businesses.

The significance of these announcements

Even without widespread 5G, CSPs are awash with data. More troubling is that even with 5G, revenues are not growing commensurate with investment. All of the mature markets are saturated, so growth mostly comes from poaching others' customers. Better to extract more from existing customers. Best way to do that is to improve the quality of experience.

CSPs have long had systems to help Customer Service Representatives (CSRs) upsell customers and billing departments manage churn. What's changed more recently is AI/ML. From a care perspective, AI and ML help systems make recommendations more quickly, more personalized to the subscriber and based on more types of data. AI/ML allow CSPs to do more sophisticated correlation of network and service performance and predictive/proactive maintenance to minimize or altogether avoid service interruptions or degradation.

The challenges of implementing AI/ML-solutions

As with most data-related initiatives, the challenges with implementing AI/ML-driven customer experience solutions relate to data integrity and organizational issues. CSPs have accumulated petabtyes of data over the years, and few, if any, have a good sense for how accurate and complete the data is. The introduction of real-time data into the mix only exacerbates the issue. In some cases, CSPs will ingest data from third party sources, which again, adds a layer of complexity and risk. It almost goes without saying that privacy and data sovereignty issues must be considered anytime subscriber data is in play. And while it is ideal to collect data from numerous sources from across the organization, getting the different departments to agree to share their data is often difficult, and must be mandated from the most senior levels to bring everyone on-board. Indeed, such initiatives need executive sponsorship to ensure compliance and participation. Without that, these projects are unlikely to reach their full potential and deliver maximum value to the business. The journey to a data-driven organization is fraught with challenges and pitfalls, but if planned and managed correctly, it will be more than worth it in the end. The solutions announced at MWC can play an important role in enabling CSPs to reach their goals.

Written 3rd March.

Meet the Analysts

Now you've read their insights, why not click through to see more of our expert analyst's opinions, articles and reports?

Natalia Modjeska

Bradley Shimmin

Lian Jye Su

Andrew Brosnan

Alexander Harrowell

Ketaki Borade

Michael Azoff

Eden Zoller

Tim Banting

Mila D'Antonio

David Myron

Roz Roseboro

.

Learn More

Enjoyed this report?

If so, you'll love our other Generative AI content. **Claim your 10% discount on GAI reports here.**

BUILD BRAND. EARN TRUST. DRIVE DEMAND.

Informa Tech's Applied Intelligence Group

Across the Internet of Things (IoT), Artificial Intelligence (AI), Quantum Computing, and Data Science, we provide integrated research, consulting, media, training, and events, that empower our customers to use tech as a driver for positive change.

If you're looking for leading market intelligence, to make meaningful connections, or to tell your story to a world of engaged technology decision makers, then we need to talk.

Visit appliedintelligence.com to explore more and get in touch.

EVENTS

THE AI SUMMIT SERIES

APPLIED INTELLIGENCE LIVE!

RESEARCH AND CONSULTING

OMDIA

MEDIA SITES

AI BUSINESS

ENTER QUANTUM

IOT WORLD TODAY

Brought to you by Informa Tech

Contact Us

omdia.com |askananalyst@omdia.com

Copyright notice and disclaimer