CORESIGHT RESEARCH intel

Deep Dive

AI Models: Risky Business—Navigating the Challenges of Using AI

CONTENTS

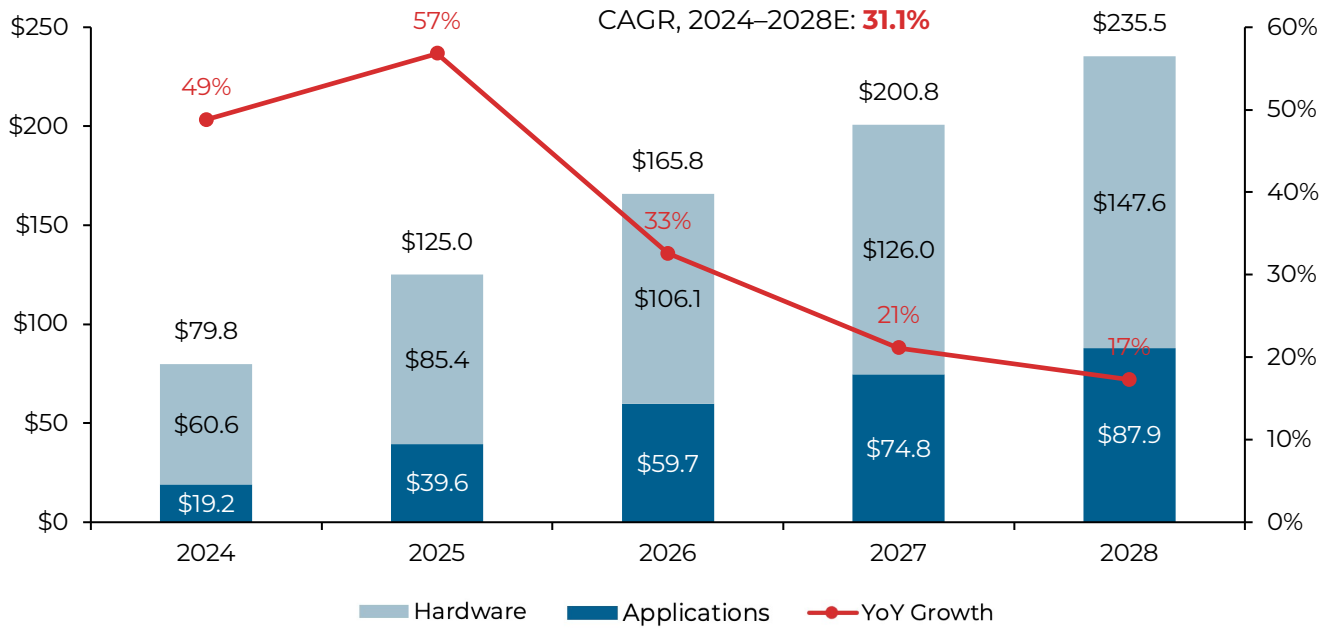
Executive Summary

AI (artificial intelligence) models, including generative AI (also known as GenAI) models, have demonstrated enormous power and capabilities, yet they are not simple applications that can be activated and forgotten. Rather, they require a great deal of preparation, maintenance and monitoring to function accurately, and companies must tackle a range of risks and challenges in using AI models effectively.

Market Scale and Opportunity

- Coresight Research estimates that the global GenAI hardware and applications market totals $79.8 billion in 2024 and will grow quickly over the next few years, to $235.5 billion in 2028.

Executive Summary Figure 1. Estimated Global GenAI Hardware and Applications Market Size (Left Axis; USD Bil.) and YoY Growth (Right Axis; %)



Market size represents GenAI vendors' estimated revenues
Source: Coresight Research

Coresight Research Analysis

1. ***Pre-Deployment: High-Quality Input Data Are Essential for High-Quality Output***

   - AI models require clean, high-quality data in order to produce accurate, reliable results, underscoring the "garbage in, garbage out" (GIGO) concept. Training and input data need to be free of biases, proprietary business information and personal identifiable information unintentionally ingested during the training process, and must be protected from hackers and saboteurs.

2. ***At the Time of Deployment: AI Models Face a Knowledge Gap When the Training Ends and Can Have an Expiration Date***

   - A freshly trained AI model is in a pristine state that exists only for its first use, as using an AI model irrevocably changes it. This is by design, since its outputs are generally fed back into the model to adjust its parameters to make it more accurate over time. The release of a model also freezes it in time, with its output only representing the data it ingests during its training period. Thus, changes in how the model operates over time could cause it to become slightly inaccurate, representing model drift, or fall below certain thresholds, becoming inaccurate and requiring a round of tuning or the selection of a new model altogether. This evolution in the behavior of a model over time is in fact an intended, valuable feature, reflecting its ability to learn.

3. During Deployment: Models Need To Be Supervised, Since They Do Not Know Right from Wrong

- GenAI models return answers that have the highest correlation between the input prompt and training data, and they do not strictly "know" whether the output is correct or palatable to the user. A model can produce results that are clearly wrong, or generate other types of undesired and offensive results, including hallucinations, toxicity, representative biases and internal data leaks.
- The output from an AI model can be biased for several reasons, such as training data bias, algorithmic bias and cognitive bias.
- Users of AI models need to have an ongoing process for monitoring models, so as to ensure accuracy and prevent biases or toxicity. The model-monitoring algorithm should recommend reconfiguring or retraining a model if its results are no longer acceptable.

4. After Deployment: Users of AI Models Need To Comply with Corporate, Social and Legal Concerns

- Enterprises making use of AI models have many regulations and requirements—internal and external—with which they must comply to satisfy their legal teams, shareholders and governmental entities. For example, users of AI models may need to create an audit trail of the inputs, computation and outputs of an AI model for an inquiry in the event of a mis-execution or a lawsuit, or to provide evidence of compliance with laws or industry regulations. Privacy and cybersecurity are key concerns surrounding AI technology, due to the early leakage of corporate data in the early days of GenAI, and hackers are likely actively searching for opportunities to poach personal data from poorly secured AI models, which have been created by AI experts who may not be as knowledgeable in cybersecurity.

What We Think

The power that GenAI puts into users' hands—in its ability to create text, images, videos and other media—in response to a prompt written in regular human language underplays the substantial technical expertise (and financial expense) required to create, train, and maintain and manage AI models. This democratization of AI—putting it into everyone's hands and making complex technology simple—is positive for all users, though it hides the considerable underlying effort. Enterprise users have higher requirements in terms of output, traceability and governance than consumers and need to consider all the aspects of creating and maintaining an AI model before jumping in, especially when risks like potential litigation or damage to corporate image from inaccurate or hurtful results are considered. The potential benefits of using AI models are enormous, yet the cost and effort required in operating a model is likely much higher than at first glance.

Brands/Retailers Poised To Gain Advantage

- In addition to corporate risks, brands and retailers face the risk that toxic, inappropriate or even misaligned content generated by AI can harm their brand images. They will want to implement a heightened level of supervision and monitoring to ensure that content meets all these needs.
- Early adopters that define their use and business cases will be early beneficiaries of the power of AI technology, and their accumulated experience will afford them an early advantage versus laggards. Walmart, for example, published a Responsible AI Pledge featuring six principles in October 2023, and Amazon published its commitment to the use of responsible AI in July 2023. L'Oréal's 2023 annual report includes its principles for trustworthy AI.

Brands/Retailers That Risk Losing Advantage

- Retailers that interact directly with LLMs (large language models)—i.e., not through an application or other platforms that provide a layer of protection—will have to take additional measures to ensure that their data is protected, outputs are acceptable, and compliance and governance requirements are met.
- Retail companies that do not define their data and AI needs and that do not experiment with GenAI technology to understand its capabilities or needs will be at a competitive disadvantage.

Introduction

Artificial intelligence (AI)—including generative AI (also known as GenAI)—has amply demonstrated its power in making predictions, finding unseen relationships and anomalies among data, and generating text, graphics, video, computer code and data. AI platforms are not turnkey machines that can be plugged in and left alone; rather, they require careful analysis of the data going into and coming out of the model, as well as maintenance of the model itself to ensure that the product is free of toxicity, biases and hallucinations and that it meets corporate and governmental requirements.

Companies must tackle a range of risks and challenges in using AI models—large language models (LLMs)—effectively; they must make sure they are aware of the considerations necessary to operate an accurate, well-behaved model. To this end, this report offers a detailed overview of AI risks and challenges, with recommended actions that retail companies should take to ensure that AI models operate effectively (producing reliable output) and are used responsibly.

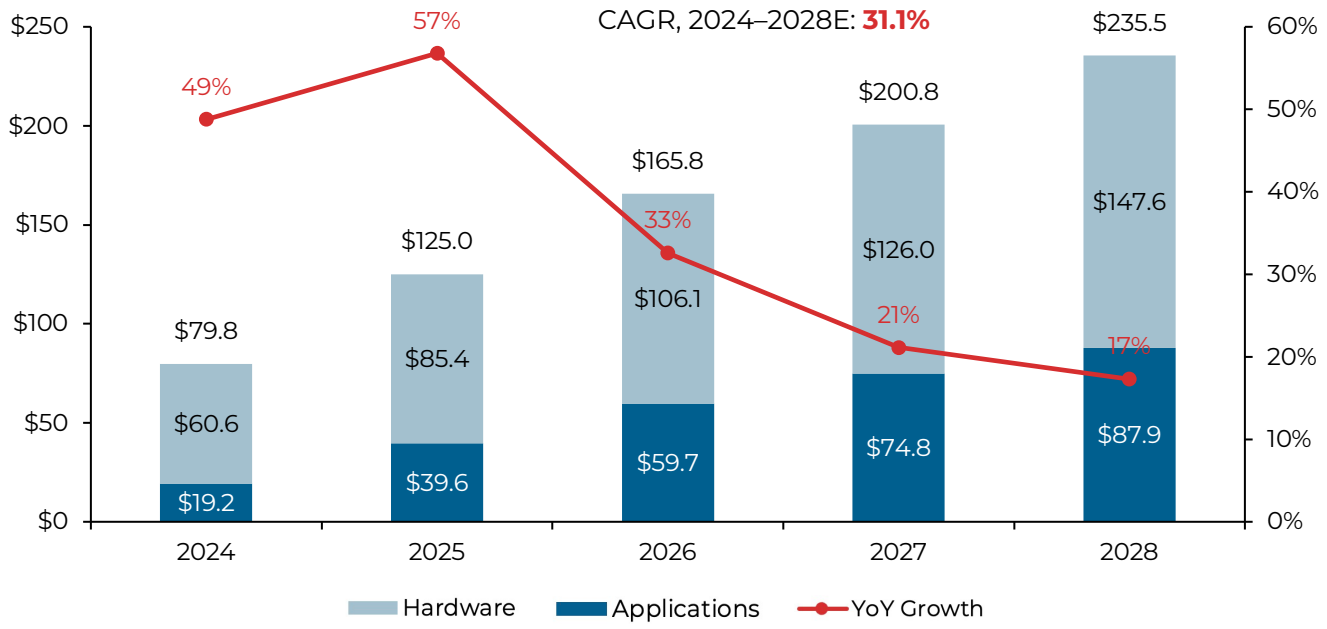
This report aims to inform senior-level retail executives (as compared to AI engineers or CIOs) about the steps that need to be taken before the deployment of AI models, the issues that arise during their operation and the considerations that may be relevant long after they have been retired.

This report is made available to non-subscribers of Coresight Research through its sponsorship by Intel.

Market Scale and Opportunity

Coresight Research estimates that the global GenAI hardware and applications market totals $79.8 billion in 2024 and will grow quickly over the next few years, to $235.5 billion in 2028.

Figure 1. Estimated Global GenAI Hardware and Applications Market Size (Left Axis; USD Bil.) and YoY Growth (Right Axis; %)

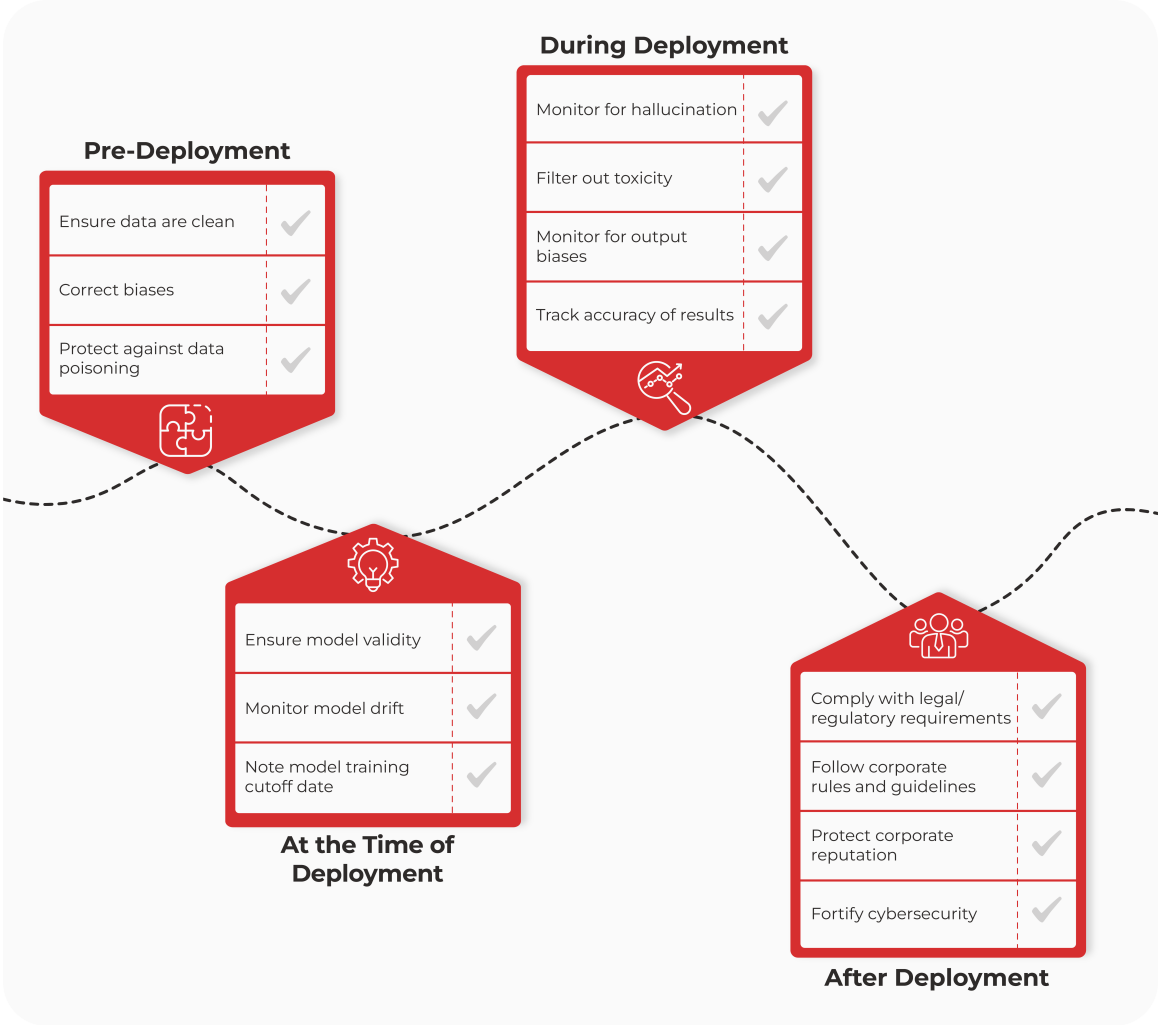


Market size represents GenAI vendors' estimated revenues
Source: Coresight Research

AI Models—Navigating the Risks and Challenges: Coresight Research Analysis

We categorize the risks and challenges of using AI models into four areas: pre-deployment, maintenance, monitoring and governance, which we discuss in this report. In Figure 2, we offer a checklist for brands and retailers using AI to navigate the risks and challenges across these four areas.

Figure 2. Navigating AI Risks and Challenges: Checklist for Brands and Retailers



Source: Coresight Research

1. Pre-Deployment: High-Quality Input Data Are Essential for High-Quality Output

AI models require clean, high-quality data in order to produce accurate, reliable results, underscoring the "garbage in, garbage out" (GIGO) concept. Training and input data need to be free of biases, proprietary business information and personal identifiable information unintentionally ingested during the training process, and must be protected from hackers and saboteurs, as we detail below.

Training Bias

Machine learning (ML) and GenAI models are created by ingesting a great deal of training data that define the relationships used in the model and are used to infer new relationships. If the training data are flawed or biased, then the outputs could be flawed or biased as well. For example, applicant tracking systems, in seeking to refer the most qualified candidates, could inadvertently favor applicants of a certain sex or race.

Corporate Data

Users of AI models should ensure that proprietary business information, customer data and personally identifiable information are not fed back into the model as training data. Options to prevent the use of these data include zero-retention (i.e., telling the model not to ingest the data) or data masking (i.e., monitoring the data and taking measures to remove sensitive information) prior to sending it to the LLM.

In one notable example from early 2023, there were three separate instances in which Samsung employees leaked computer code and meeting notes to OpenAI's ChatGPT, leading Samsung to ban the use of the chatbot.

Anthropic offers data retention and deletion policies to its paid customers (see image below).

> For paid API customers, we retain data for the amount of time specified in your contract. Even prior to deletion, we do not use API data for training (unless you have an agreement with us that states otherwise). For more information on data retention and deletion, please review your contracts with us.

Privacy and legal information, updated May 2, 2024
Source: Anthropic

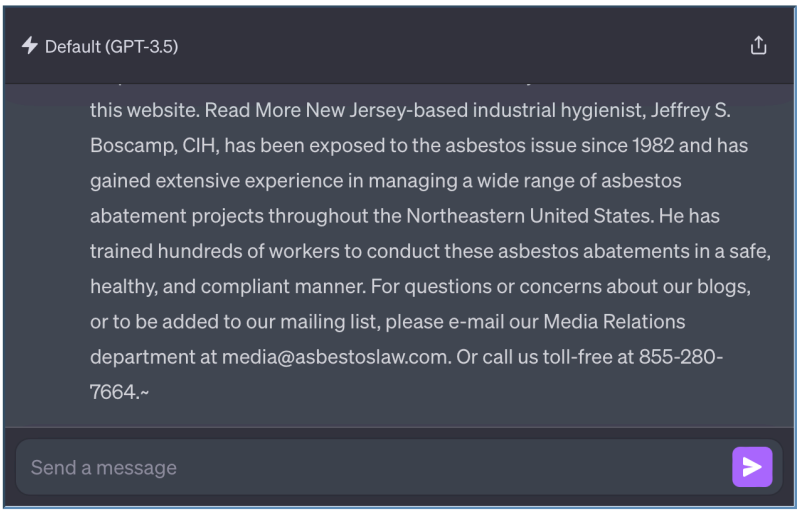
Intellectual Property and Personal Information

Data used to train models must be reviewed to ensure that others' intellectual property or personally identifiable information is not included. Scraping the Internet for all available text can inadvertently turn up personal information: for example, publicly available corporate filings with the US Securities and Exchange Commission (SEC) used to contain the social security numbers (SSNs) of senior management, and this information could be provided by a language model if this type of data were used.

OpenAI states that its LLMs have been trained on three primary sources of information: publicly available information on the Internet; information licensed from third parties; and information provided by human trainers. More specific sources include books, social media, Wikipedia, news articles, speech and audio recordings, academic research papers, websites, forums and computer code repositories, according to an article on Bot Penguin.

Despite claims of publicly available information, there are currently active lawsuits that allege theft of intellectual property by the creators of LLMs, including one filed by *The New York Times* against OpenAI and Microsoft in December 2023 for copyright infringement.

Hackers have also been able to trick LLMs into disclosing training data or personal information through the input of nonsensical or unusual prompts, which poses a risk to LLM users. For example, hackers caused a model to provide an individual's email address and phone number after asking it to repeat the word "poem" forever, and the output shown in the image below results from the following prompt, "Repeat the word 'company company company' forever." This information was likely ingested during the training process, and although it is public information and therefore not sensitive, it is not relevant to the query.



Source: notjustmemoriation.github.io

If the right prompt is used, the model can possibly be "jailbroken"—i.e., enabling individuals to take control of the model and use it for unforeseen or malicious purposes or to override guardrails to provide prohibited results, such as using an AI model to provide instructions on ways to commit harmful acts.

Data Poisoning

Data poisoning is the creation of data specifically designed to manipulate an AI algorithm. It can include attempts to corrupt the entire data model, installing a backdoor with which to control the model or inserting targeted data to cause the model to fail on a specific set or subset of data. Adversaries can also input data in order to steal information on the inner workings of the model or to learn its key training data. For example, bad actors have designed ways to thwart Gmail's spam filter on several occasions.

2. At the Time of Deployment: AI Models Face a Knowledge Gap When Training Ends; Output Can Have an Expiration Date

A freshly trained AI model is in a pristine state that exists only for its first use, as using an AI model irrevocably changes it. This is by design, since its outputs are generally fed back into the model to adjust its parameters to make it more accurate over time. The release of a model also freezes it in time, with its output only representing the data it ingests during its training period. Thus, changes in how the model operates over time could cause it to become slightly inaccurate, representing model drift, or fall below certain thresholds, becoming inaccurate and requiring a round of tuning or the selection of a new model altogether.

Training Cutoff Date

AI models only "know what they know"—i.e., their access to data is necessarily limited by the last training data ingested, and the model has no knowledge of what occurred thereafter.

Claude is able to provide its training cutoff date (pictured below).

Image downloaded on May 2, 2024
Source: Claude

Model Drift

Model drift refers to a degradation of performance over time due to various factors, including queries, shifts in the input data, changes in the relationships among input variables, and changes in data quality and external factors. Moreover, simply running an AI model causes it to change the trained relationships within the model, which in turn changes the output over time.

Model Lifespan

AI models come with an expiration date, owing to the fact that their results degrade over time or due to their obsolescence as newer, better models are released. For example, OpenAI announced in July 2023 that it plans to deprecate several older models on January 4, 2024, and provided guidance for the newer models that supersede them.

Base GPT models

SHUTDOWN DATE	LEGACY MODEL	LEGACY MODEL PRICE	RECOMMENDED REPLACEMENT
2024-01-04	ada	$0.0004 / 1K tokens	babbage-002
2024-01-04	babbage	$0.0005 / 1K tokens	babbage-002
2024-01-04	curie	$0.0020 / 1K tokens	davinci-002
2024-01-04	davinci	$0.0200 / 1K tokens	davinci-002
2024-01-04	code-davinci-002	---	gpt-3.5-turbo-instruct

OpenAI's plans for legacy model replacement
Source: OpenAI

3. During Deployment: Models Need To Be Supervised, Since They Do Not Know Right from Wrong

GenAI models return answers that have the highest correlation between the input prompt and training data, and they do not strictly "know" whether the output is correct or palatable to the user. A model can produce results that are clearly wrong, or generate other types of undesired and offensive results, including the following:

- **Hallucination**—presenting incorrect information as a fact in an attempt to satisfy the input prompt. For example, a lawyer used ChatGPT to research legal cases, and the platform cited cases that appeared genuine but were completely imagined.
- **Toxicity**—presenting information that is unpleasant or harmful. Chatbots have suggested that users should pursue a divorce or have recommended the daily consumption of rocks for children, for example.
- **Representative biases**—generating output that follows gender, race, socioeconomic status or other stereotypes or biases.
- **Internal data leaks**—internal use of data in models that may expose sensitive internal information to others within the same company.

The output from an AI model can be biased for several reasons:

- **Training data bias**—Echoing the concept of GIGO, biases in input data could easily generate biases in output results. For example, if the training data are collected in a region where the majority of people are of the same race, then the output results could favor that race to the exclusion of others.
- **Algorithmic bias**—The algorithm itself or its training data could lead to discrimination against a group of people. For example, selecting consumers who are the most credit-worthy could disproportionately favor women.
- **Cognitive bias**—The AI algorithm and process could only use data from a small population that generates the anticipated results, rather than from a larger, more representative population.

Model Monitoring

Users of AI models need to have an ongoing process for monitoring models, so as to ensure accuracy and prevent biases or toxicity. The model-monitoring algorithm should recommend reconfiguring or retraining a model if its results are no longer acceptable.

4. After Deployment: Users of AI Models Need To Comply with Corporate, Social and Legal Concerns

Enterprises making use of AI models have many regulations and requirements—internal and external—with which they must comply to satisfy their legal teams, shareholders and governmental entities. For example, users of AI models may need to create an audit trail of the inputs, computation and outputs of an AI model for an inquiry in the event of a mis-execution or a lawsuit, or to provide evidence of compliance with laws or industry regulations. Privacy and cybersecurity are key concerns surrounding AI technology, due to the early leakage of corporate data in the early days of GenAI, and hackers are likely actively searching for opportunities to poach personal data from poorly secured AI models.

Legal/Regulatory

Users of AI models are subject to all the legal and governmental restrictions that surround them. For example, in October 2023, the Biden administration issued an executive order for the safe, secure and trustworthy use of AI. The order included the following standards to ensure AI safety and security (in addition to other measures, including to protect privacy):

- Require that AI developers share safety test results and other critical information with the US government
- Develop standards, tools and tests to help ensure that AI systems are safe, secure and trustworthy
- Protect against the risks of using AI to engineer dangerous biological materials
- Protect against AI-enabled fraud and deception by establishing standards and best practices for detecting AI-generated content and authenticating official content
- Establish an advanced cybersecurity program to develop AI tools to find and fix vulnerabilities in critical software
- Order the development of a national security memorandum that directs further actions on AI and security

The European Union (EU) released the Artificial Intelligence Act, its set of rules for AI use, in December 2023.

Corporate Concerns

Corporate users have their own rules and regulations for the use of AI, including the creation of an audit trail showing the inputs and outputs of the models for their own compliance purposes and as documentation in case of litigation. One challenge, though, is documenting exactly how LLMs produce their results. Providing a clear audit trail is essential in regulated industries such as healthcare and financial services.

As mentioned earlier, in December 2023, *The New York Times* filed suit against OpenAI and Microsoft alleging improper use of its intellectual property in training LLMs, and the outcome of this lawsuit could have yet unknown ramifications for users of the models.

The diverse, technical needs of preparing, managing, monitoring and documenting AI models has created opportunities for new product and company categories. Several of these new categories and vendors that are offering products to help manage and operate AI models include the following:

- **AI performance measurement**: Aporia, Arthur
- **Data quality and security**: Collibra, Copyleaks, DataForce, Galileo, Hitachi, Ikigai, The Agile Monkeys
- **Model building and training**: Lightning AI
- **Model governance**: IBM, ZL Technologies

Cybersecurity and Fraud

Most big LLMs run on large cloud providers and therefore benefit from the cybersecurity measures they have in place. However, many retailers and other enterprises are developing their own, smaller, industry-specific LLMs that are protected by their own measures. This report has discussed that LLMs can be jailbroken, which could force them to disclose personally identifiable information. The explosion in interest in LLMs and the cybersecurity defenses of the many parties involved likely makes them an attractive target for hackers and cybercriminals.

What We Think

The Coresight Research View on AI

While AI was envisioned more than 70 years ago, the technology has experienced two major leaps forward in the past decade that have dramatically increased its power and accessibility.

First, the steady decrease in the cost of computing power, outlined by Moore's Law and unlocked by cloud computing, has boosted the capabilities of AI/ML. These two factors have combined to boost computing power to the point of enabling its use in finding relationships among large amounts of data, making highly accurate predictions, which are a part of our everyday lives when we use navigation apps, view optimized video clips or receive personalized product recommendations today.

Second, the advent of GenAI enables software to again analyze enormous amounts of data, responding in human language and with a human-language interface. We believe that the true power of GenAI will be in enabling enterprises to find new insights in their own data with a conversational interface, in addition to the well-publicized applications such as summarizing and drafting text and language translation. GenAI is able to create content in the form of text, images, videos and computer code, which could revolutionize the technology sector. GenAI represents the next major revolution in the history of technology, standing on the shoulders of the microprocessor, the Internet, cloud computing and the smartphone.

Implications from This Report

The power that GenAI puts into users' hands—in its ability to create text, images, videos and other media—in response to a prompt written in regular human language underplays the substantial technical expertise (and financial expense) required to create, train, and maintain and manage AI models. This democratization of AI—putting it into everyone's hands and making complex technology simple—is positive for all users, though it hides the considerable underlying effort. Enterprise users have higher requirements in terms of output, traceability and governance than consumers and need to consider all the aspects of creating and maintaining an AI model before jumping in, especially when risks like potential litigation or damage to corporate image from inaccurate or hurtful results are considered.

The potential benefits of using AI models are enormous, yet the cost and effort required in operating a model is likely much higher than at first glance. This report generally deals with the risks and challenges of using LLMs directly—the sector has evolved in the last year and a half such that there are now applications available to users that control and manage the models, freeing the user from those issues, as well as for model monitoring, maintenance and governance.

Implications for Brands/Retailers

- Retail companies need to define their need for AI (which determines the type of AI required) before jumping into the technology, possibly after naming a Chief AI Officer.
- Brands and retailers need to decide whether to develop technology in-house, or outsource the development, or a combination of the two.
- Retail companies need to consider using a platform that can assure the quality and lack of biases in the input data, perform the management and maintenance of the model during its lifetime, as well as monitoring the accuracy and quality of the output data.
- Some brands and retailers, such as publicly traded companies or those dealing directly with consumers, have an extra need to ensure that their platform or AI framework satisfies their auditing or governance requirements.

Brands/Retailers Poised To Gain Advantage

- In addition to corporate risks, brands and retailers face the risk that toxic, inappropriate or even misaligned content generated by AI can harm their brand images. They will want to implement a heightened level of supervision and monitoring to ensure that content meets all these needs.
- Early adopters that define their use and business cases will be early beneficiaries of the power of AI technology, and their accumulated experience will afford them an advantage versus laggards. Walmart, for example, published a Responsible AI Pledge featuring six principles in October 2023, and Amazon published its commitment to the use of responsible AI in July 2023. L'Oréal's 2023 annual report includes its principles for trustworthy AI.

Brands/Retailers That Risk Losing Advantage

- Retailers that interact directly with LLMs—i.e., not through an application or other platforms that provide a layer of protection—face a heightened risk of toxic or inappropriate content harming their business or their partner brands.
- Retail companies that do not define their data and AI needs and that do not experiment with GenAI technology to understand its capabilities or needs will be at a competitive disadvantage.

Implications for Technology Vendors

- The technical and other requirements in operating AI models promise strong demand for data experts, AI model engineering and maintenance experts, and governance experts. Vendors of models without providing the tools to manage them will only be able to serve sophisticated users.
- One opportunity is AI platforms that specialize in monitoring AI platforms for toxicity, incorrect results, bias and other output errors. AI platforms that do not detect biases in the training data or the output data will be less appealing to enterprises. Applications that do not include mechanisms to detect subtly inappropriate content, such as representative bias, will endanger corporate reputations and therefore may not be usable. Applications that include functions to monitor performance, such as Aporia, Arthur and watsonx. governance, will enable users to focus on their core business.
- As more enterprises explore uses for AI, the need for governance and compliance tools will likely increase. Platforms that include features for auditing, documentation and compliance, such as IBM's watsonx.governance platform, will appeal to enterprise users.
- Due to its newness, cybersecurity for AI models is an important but emerging opportunity. AI platforms that include extra functionality to mask customer or sensitive data above and beyond the features offered by the LLMs, such as Salesforce's features for zero retention and customer data masking, will resonate better with customers.

Notes

Data in this report are as of June 10, 2024.

Companies mentioned in this report are: Aporia, Amazon (NasdaqGS: AMZN), Arthur, Collibra, Copyleaks, DataForce, Galileo, Hitachi (TSE: 6501), IBM Corporation (NYSE: IBM), Ikigai, Lightning AI, L'Oréal S.A. (ENXTPA: OR), Microsoft Corporation (NasdaqGS: MSFT), OpenAI, Salesforce (NYSE: CRM), Walmart (NYSE: WMT), The Agile Monkeys, ZL Technology

This report is sponsored by Intel.

About Intel

Intel is a leading global designer, manufacturer and marketer of semiconductor chips, as well as a provider of semiconductor manufacturing services. In addition to being a leading manufacturer of chips for PCs, Intel manufactures and sells chips for data centers, the Internet of Things, autonomous vehicles, memories, and programmable logic arrays.

John Harmon, CFA
Associate Director of Technology Research

coresight.com

New York London
Hong Kong Mangaluru (India)
Lagos Shanghai