Innovation Guide for Generative AI Models

Published 5 January 2024 - ID G00800194 - 17 min read

By Analyst(s): Arun Chandrasekaran, Arnold Gao, Ben Yan

Initiatives: Artificial Intelligence; Digital Future; Generative AI Resource Center

Due to their pretraining and wide use-case applicability, generative AI models are a major AI advance, but the risks they pose aren't fully understood. Data and analytics leaders should evaluate the benefits, risks and opportunities of these models to tap their business value and minimize risk.

Overview

Key Findings

- Generative AI models represent a huge change in the field of AI due to their massive pretraining and versatility across a variety of tasks.

- Generative AI model deployments pose several risks, such as unintended loss of intellectual property, hallucinations, lack of explainability, legal risks and potential for misuse. While it is hard to completely eliminate these risks, best practices are emerging to mitigate them.

- Although early adoption of generative AI models is within the natural language processing (NLP) domain, adoption is growing in the computer vision, software engineering and rich media domains. Large technology research labs are extending them to deep sciences in aid of molecular generation, drug discovery and chemical formulation.

- While a few large technology vendors dominate most of the recent advances in this area, a growing ecosystem of startups and vibrant open-source projects are gaining traction with enterprises.

Recommendations

- Be objective about the adequate balance between accuracy, costs, security and privacy, and time to value when selecting generative AI models to determine the appropriate model needed. Not all use cases require the largest or the highest-performing models.

- Select vendors that promote responsible training and deployment of AI models with transparent training processes, stringent privacy SLAs and enforceable legal indemnifications.

- Adopt a platform-centric approach with centralized AI engineering tools and applications where models can be swapped with low exit barriers, given the tumultuous and rapidly evolving model landscape.

- Avoid expensive model customizations, which can raise exit costs and prevent you from leveraging state-of-the-art and updated models.

Strategic Planning Assumption(s)

- By 2027, over half of the generative AI models used by enterprises will be domain-specific (industry or business function), up from 1% today.

- By 2028, open-source generative AI models will underpin more than 50% of enterprise generative AI use cases, up from less than 10% today.

- By 2028, more than 50% of developed countries will have enacted regulations to govern generative AI, up from less than 1% today.

- By 2028, more than 50% of enterprises that have built their own models from scratch will abandon their efforts due to costs, complexity and technical debt in their deployments.

- By 2028, one-third of interactions with generative AI services will invoke autonomous agents to achieve tasks, up from less than 1% today.

Contribute to Beta Research

The following research is a work in progress that does not represent our final position. We invite you to provide constructive feedback to help shape this research as it evolves. All relevant updates and feedback will be incorporated into the final research.

Use these jump links to navigate the document:

Market Definition

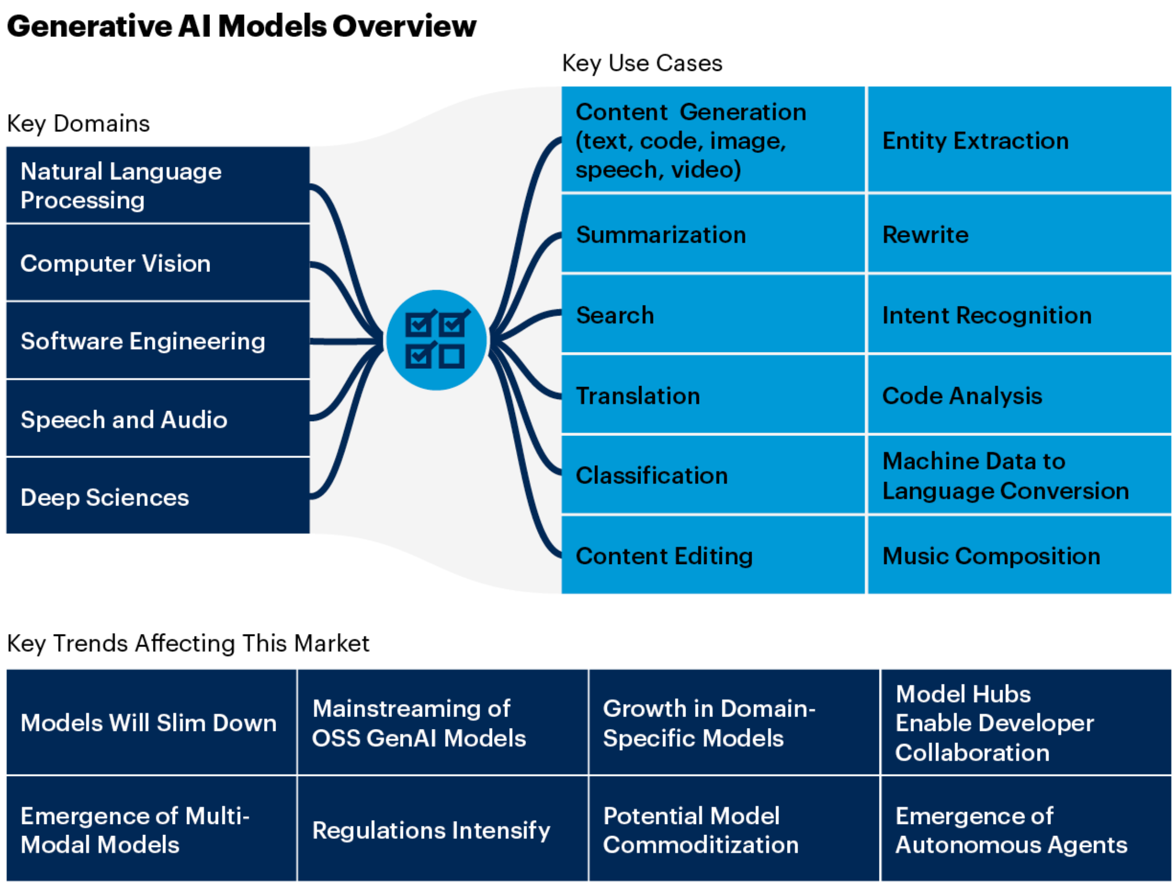
Back to top

Generative AI models are mid- to large-parameter models trained on a broad gamut of datasets in a self-supervised manner. They are mostly based on transformers, for large language models (LLMs), or diffusion, for computer vision, deep neural network (DNN) architectures and will incorporate a wide variety of modalities (multimodal) in the near future. Generative AI models are critically important due to their pretraining and broad applicability to a wide variety of downstream use cases. Generative AI models are a subset of our broader research, Innovation Guide for Generative AI Technologies.

Market Map

Back to top

Figure 1: Generative AI Models Overview



Generative AI Models Overview

Key Domains

- Natural Language Processing
- Computer Vision
- Software Engineering
- Speech and Audio
- Deep Sciences

Key Use Cases

Content Generation (text, code, image, speech, video)	Entity Extraction
Summarization	Rewrite
Search	Intent Recognition
Translation	Code Analysis
Classification	Machine Data to Language Conversion
Content Editing	Music Composition

Key Trends Affecting This Market

Models Will Slim Down	Mainstreaming of OSS GenAI Models	Growth in Domain-Specific Models	Model Hubs Enable Developer Collaboration
Emergence of Multi-Modal Models	Regulations Intensify	Potential Model Commoditization	Emergence of Autonomous Agents

Source: Gartner
774602_C

Gartner

Market Dynamics

Back to top

Generative AI models are the foundation for innovation in AI today. They have quickly gained massive interest due to:

- **Scale and performance —** The key benefit of the scale of these models is that they can be effective in zero-shot or few-shot scenarios, where little domain-specific training data is available. Generative AI models offer a higher level of performance than from-scratch approaches. This is possible by leveraging pretrained models and training them further on instruction datasets in a supervised learning fashion. This often achieves better performance for specific use cases with a limited amount of additional data.The increasing scale of the models has resulted in better performance for several natural language tasks, such as reading comprehension, sentiment analysis, content generation and fact checking. Meanwhile, the efficacy (of increasing scale) for reasoning-oriented use cases is still unproven. However, the model sizes are expected to have slower growth or even plateau as cost and simplicity become important criteria for buyers.

- **Ease of access —** Cloud computing has enabled providers to train, iterate and distribute models as APIs faster. The sheer volume of models being provided as cloud services is lowering their barrier to entry into the enterprise, although the inference costs are becoming a topic of focus as buyers try to scale generative AI efforts economically.

- **Domain adaptation —** With fine-tuning, generative AI models can be adapted for several domains. Hence, they not only are useful in language or NLP tasks but also can be used in the banking or legal space for anomaly detection, in healthcare to analyze drug interactions or medical histories, in government and education, and so on.

Closed-source models dominate the generative AI model landscape today. The viral adoption of ChatGPT and OpenAI's close partnership with Microsoft enabled OpenAI and Microsoft to be the early leaders in this space. However, other large technology companies have been aggressively competing and are starting to make inroads, with Google, AWS, Salesforce, NVIDIA, Databricks and IBM being the prominent ones. The steep training costs, access to cutting-edge researchers and cloud-based distribution are competitive moats that large technology companies are wielding to gain early customer traction.

Generative AI models are evolving to accommodate modalities beyond text, such as DALL-E (OpenAI), Imagen (Google) for images, Codex (OpenAI) for code, and MusicLM (Google) and GODIVA (Microsoft) for audio and video. For most early pilots and deployments for generative AI, end-user clients are consuming these models as an API in the cloud and steering them via simple prompt engineering (zero-shot or few-shot learning) or steering these models via retrieval augmented generation (RAG) architectures for prompt augmentation.

Market Evolution

Back to top

- **Models slim down:** While the past few months have seen the launch of massively large generative AI models (such as GPT-4), the future might see more use-case-specific and midsize to smaller-size models. The obvious merits of these use-case-specific models are their use-case affinity, smaller size, curated datasets and lower cost. Vendors such as AWS, IBM, Databricks and Salesforce, and even OpenAI, Microsoft and Google, are recognizing and driving this industry shift. Edge generative AI models are still emerging but can potentially reshape our interactions by adding intelligence at the edge, in smartphones, mobile devices, field robots and Internet of Things (IoT) devices in buildings, vehicles, factories and other locales.

- **Open-source models will challenge "state-of-the-art" closed-source models:** Open-source models are rising in prominence and challenging closed-source models. While there is still a gap in performance accuracy between state-of-the-art closed-source models and open-source models, this gap might narrow in the future due to investments by deep-pocketed entities such as Meta and Databricks, and startups such as Mistral AI, in this ecosystem. As AI regulations increase, customers may more closely consider open-source models due to deployment flexibility, customizability and better control over security and privacy.

- **Domain-specific models:** We will see the advent of more vertical-domain-specific models, particularly in healthcare, life sciences, financial services and legal. Most of these models will be built on top of AI foundation models but with domain-specific data. Early examples include BioGPT from Microsoft, BioMedLM from Stanford, Med-PaLM 2 from Google and BloombergGPT from Bloomberg.

- **Potential model commoditization:** Models may become commoditized quite rapidly as more open-source models emerge and as model builders employ similar web crawling techniques to build these models. However, enterprise buyers might still be willing to pay a premium for models with better safety and governance, agile software releases and a better price-to-performance ratio. Indeed, this will turn into a buyer's market from a seller's market.

- **Model hubs enable developer agility:** Model marketplaces such as Hugging Face will become more influential, gaining mind share with developers and disseminating a wide range of pretrained models mapped to various use cases and domains, datasets and tools to operationalize them.

- **Regulations will alter model training and release process:** Regulations specifically around generative AI are lacking today. This is likely to change, particularly with China and the EU drafting stringent regulations on model training and governance. This will impact the cost and frequency of the model training process and will raise the barriers for entry for model creators.

- **Multimodal models expand potential use cases:** Multimodal models will evolve to deliver a bigger boost in model performance. Such a combination of modalities will open up new cases in generative AI as well as alter user interfaces (toward modalities such as voice) and experience. For example, in healthcare, medical diagnosis will require both text (e.g., doctor's notes) and images (e.g., X-ray, MRI scans) for accurate results. For more, see Innovation Insight: Multimodal AI Explained.

- **Autonomous agents:** Generative AI models require extensive human intervention and skills such as prompt engineering to steer them. The ability to create autonomous agents, which can act in an automated manner with limited human intervention, is an exciting possibility, although the "black box" nature of these agents poses unknown risks. Similarly, action transformers can learn from human actions to complete tasks, extending the automation a notch above to complete complex tasks.

Business Benefits (Use Cases)

Back to top

Generative AI models have several prominent use cases. We see client interest in the following key areas.

Natural Language Processing (NLP)

Unsurprisingly, NLP has enormous client interest, given that most of these models are LLMs. Prominent use cases within this space include:

- **Text generation** — Generative AI models are good at text generation due to advances in prompt engineering. They can generate original and coherent text in a highly contextualized way, which lends itself to use cases such as writing headlines and paragraphs, creating product descriptions, generating conversational responses and completing text.

- **Summarization** — Models can summarize a full document, enabling rapid readability of emails, documents, meeting notes, transcripts and so on.

- **Rewrite** — Rewriting can be particularly useful for writing in non-native-language, moving from informal to formal writing and correcting grammar in voice-to-text conversions.

- **Entity extraction** — For certain entity extraction tasks, generative AI models can greatly improve entity recognition, often in tandem with more-structured entity extraction models.

- **Search** — Through text embedding, and when paired with vector databases, generative AI models can deliver effective semantic searches. This enables use cases such as recommendations and product searches.

- **Translation** — Multilingual generative models can improve machine translation tasks and are moving closer to transcreation, which is a culturally aware recreation of the text in its original meaning. This can be effective in dealing with texts that have traditionally been hard for machines to translate, such as idioms, allegories, pop culture references, wordplay and visual depictions.

- **Classification** — Generative AI models can output predicted classes for each input and the accompanying confidence level values. These can be useful in a variety of natural language understanding use cases, such as toxic content moderation, intent classification and sentiment analysis.

- **Intent recognition improvements** — Generative AI models can improve intent recognition, which is the most common way that conversational AI interprets what user phrases mean. This is done either by numerical representation of words to enhance the value of training phrases or by automatically generating additional training phrases with the same meaning to aid supervised training.

Beyond the NLP use cases, generative AI models are making advances in several other fields, with computer vision and software engineering being the most prominent domains. The key use cases include the following.

Computer Vision

- **Image generation** — Fine-tuned diffusion models (e.g., DALL-E 3, MidJourney, Adobe Firefly and Stable Diffusion) can create images from text. This has transformative implications for industries such as media and entertainment as well as gaming, where new business models are being invented to leverage this innovation. This can also be extremely useful for synthetic data generation for scenarios in which real data is scarce or unavailable in several industries such as automotive, retail, healthcare, energy and manufacturing.

- **Image completion** — Generative models can also learn and complete the distribution of images. This could propel the data-centric AI use case that organizations are aiming toward.

- **Video generation** — Models from providers such as Runway can create short-form videos from text, which could revolutionize art, entertainment and human creativity.

- **Video editing and summarization** — Generative AI models can enable video editing and summarization via text, which can be useful for editing webinars, podcasts and meeting recordings and for summarizing long videos.

- **VR and AR** — The future of virtual reality (VR) and augmented reality (AR) combined with generative AI extends beyond productivity and training. In automotive, generative AI can simulate test drives and showcase futuristic car designs, while educational institutions can use AR glasses to overlay 3D models and interactive content, revolutionizing learning for students.

Software Engineering

- **Natural language to code** — Generative models are being fine-tuned to augment developers by translating natural language to code.

- **Code completion** — Make suggestions for code, functions and so on.

- **Documentation generation** — Analyze the code to autocreate documentation.

- **Test data generation** — Generate sample test data tests to aid software development workflows.

- **Others** — Other use cases include application migrations, code analysis for quality and vulnerability, aiding in effective code search, generating infrastructure as code and so on.

Speech and Music

- Several use cases are emerging within the speech and music domain that leverage generative AI models' capabilities of text to speech, voice cloning, generating audiobooks from text or speech, composing original music, conversational AI avatars and creating authentic sound effects.

General Sciences and Others

- **Generative healthcare applications —** In healthcare, life sciences and pharmaceuticals, generative models are being used to create new drugs and decipher genome sequences for classification of lung cancer.
- **Robotics —** Generative models can be the basis for facilitating better human and robot interactions.
- **Safety and security —** Generative models are also being used for violence detection. They can be used effectively by homeland security and law enforcement agencies.

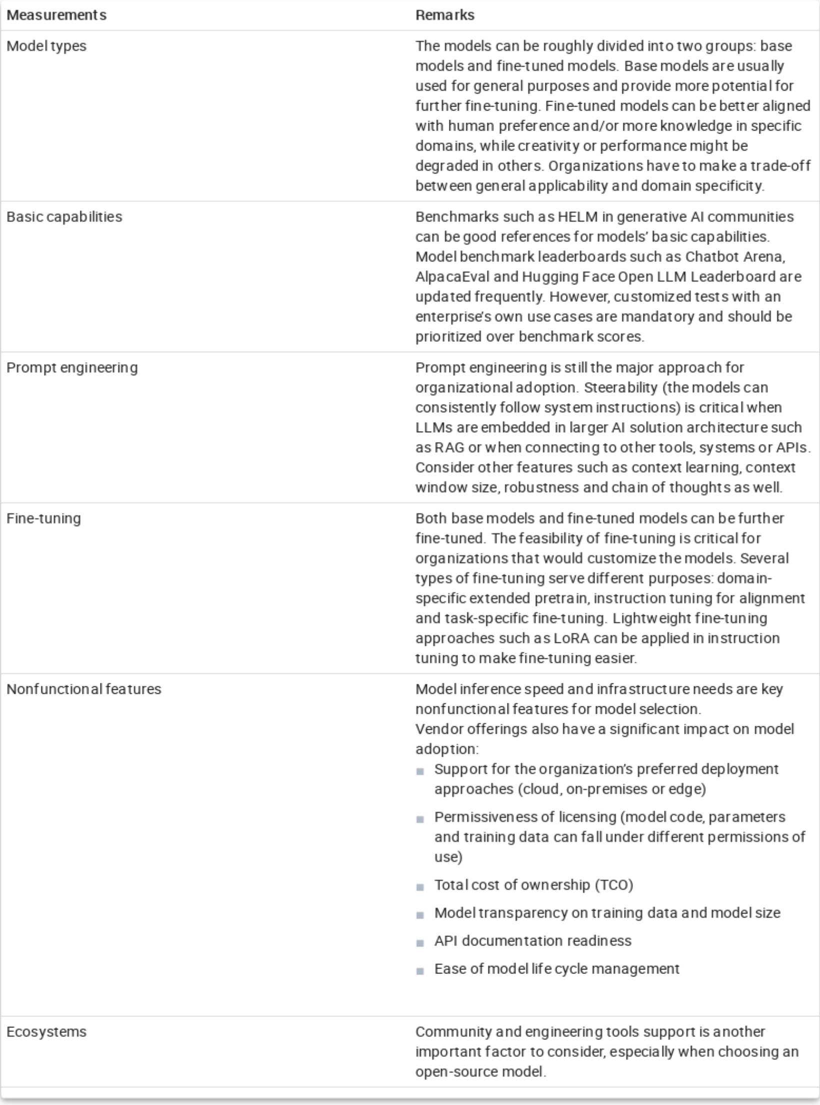
Piloting and Evaluating Vendors

Back to top

Evaluating and choosing generative AI models and vendors is not only about picking up the best ones in benchmarks such as Holistic Evaluation of Language Models (HELM), Chatbot Arena or AlpacaEval. Data and analytics leaders need to look at these models through different lenses and select the suitable ones for their enterprises (see Table 1).

Table 1: Measurements of Generative AI Models and Vendors

(Enlarged table in Appendix)



Measurements	Remarks
Model types	The models can be roughly divided into two groups: base models and fine-tuned models. Base models are usually used for general purposes and provide more potential for further fine-tuning. Fine-tuned models can be better aligned with human preference and/or more knowledge in specific domains, while creativity or performance might be degraded in others. Organizations have to make a trade-off between general applicability and domain specificity.
Basic capabilities	Benchmarks such as HELM in generative AI communities can be good references for models' basic capabilities. Model benchmark leaderboards such as Chatbot Arena, AlpacaEval and Hugging Face Open LLM Leaderboard are updated frequently. However, customized tests with an enterprise's own use cases are mandatory and should be prioritized over benchmark scores.
Prompt engineering	Prompt engineering is still the major approach for organizational adoption. Steerability (the models can consistently follow system instructions) is critical when LLMs are embedded in larger AI solution architecture such as RAG or when connecting to other tools, systems or APIs. Consider other features such as context learning, context window size, robustness and chain of thoughts as well.
Fine-tuning	Both base models and fine-tuned models can be further fine-tuned. The feasibility of fine-tuning is critical for organizations that would customize the models. Several types of fine-tuning serve different purposes: domain-specific extended pretrain, instruction tuning for alignment and task-specific fine-tuning. Lightweight fine-tuning approaches such as LoRA can be applied in instruction tuning to make fine-tuning easier.
Nonfunctional features	Model inference speed and infrastructure needs are key nonfunctional features for model selection. Vendor offerings also have a significant impact on model adoption: ▪ Support for the organization's preferred deployment approaches (cloud, on-premises or edge) ▪ Permissiveness of licensing (model code, parameters and training data can fall under different permissions of use) ▪ Total cost of ownership (TCO) ▪ Model transparency on training data and model size ▪ API documentation readiness ▪ Ease of model life cycle management
Ecosystems	Community and engineering tools support is another important factor to consider, especially when choosing an open-source model.

Source: Gartner

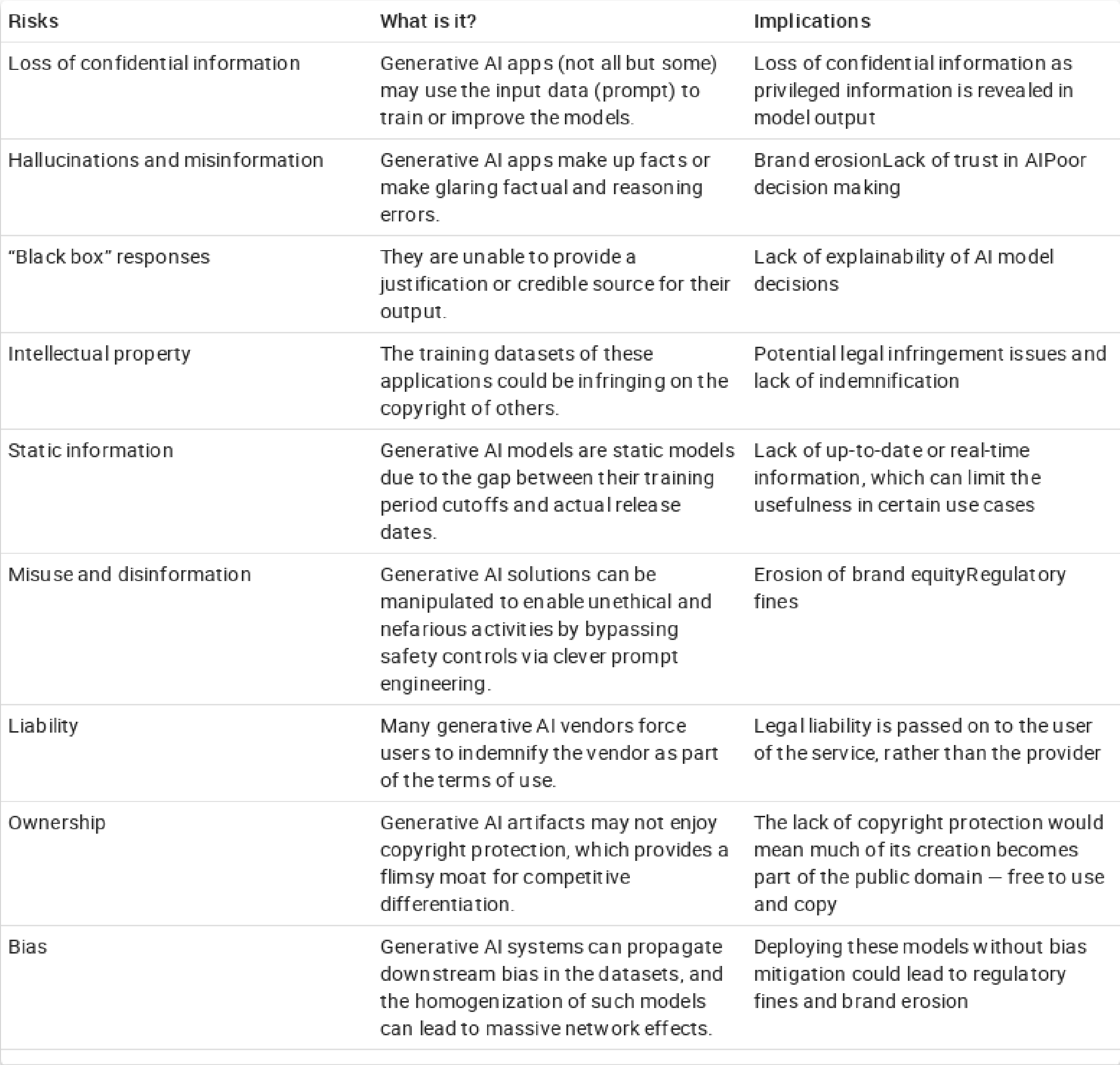
Managing Risks

Back to top

The field of generative AI has been progressing at breakneck speed. This has enabled organizations to experiment rapidly with newer AI models such as GPT-4, DALL-E, Stable Diffusion and applications such as ChatGPT and GitHub Copilot. Given the high stakes, this also creates an environment where technology vendors are rushing generative AI capabilities to market, becoming more secretive about their architectures and not taking adequate steps to mitigate the risks or potential misuse of these highly powerful services. Organizations must examine and mitigate both internal and external risks caused by generative AI (see Table 2).

Table 2: Risks and Implications of Generative AI Models

(Enlarged table in Appendix)



Risks	What is it?	Implications
Loss of confidential information	Generative AI apps (not all but some) may use the input data (prompt) to train or improve the models.	Loss of confidential information as privileged information is revealed in model output
Hallucinations and misinformation	Generative AI apps make up facts or make glaring factual and reasoning errors.	Brand erosionLack of trust in AIPoor decision making
“Black box” responses	They are unable to provide a justification or credible source for their output.	Lack of explainability of AI model decisions
Intellectual property	The training datasets of these applications could be infringing on the copyright of others.	Potential legal infringement issues and lack of indemnification
Static information	Generative AI models are static models due to the gap between their training period cutoffs and actual release dates.	Lack of up-to-date or real-time information, which can limit the usefulness in certain use cases
Misuse and disinformation	Generative AI solutions can be manipulated to enable unethical and nefarious activities by bypassing safety controls via clever prompt engineering.	Erosion of brand equityRegulatory fines
Liability	Many generative AI vendors force users to indemnify the vendor as part of the terms of use.	Legal liability is passed on to the user of the service, rather than the provider
Ownership	Generative AI artifacts may not enjoy copyright protection, which provides a flimsy moat for competitive differentiation.	The lack of copyright protection would mean much of its creation becomes part of the public domain — free to use and copy
Bias	Generative AI systems can propagate downstream bias in the datasets, and the homogenization of such models can lead to massive network effects.	Deploying these models without bias mitigation could lead to regulatory fines and brand erosion

Source: Gartner

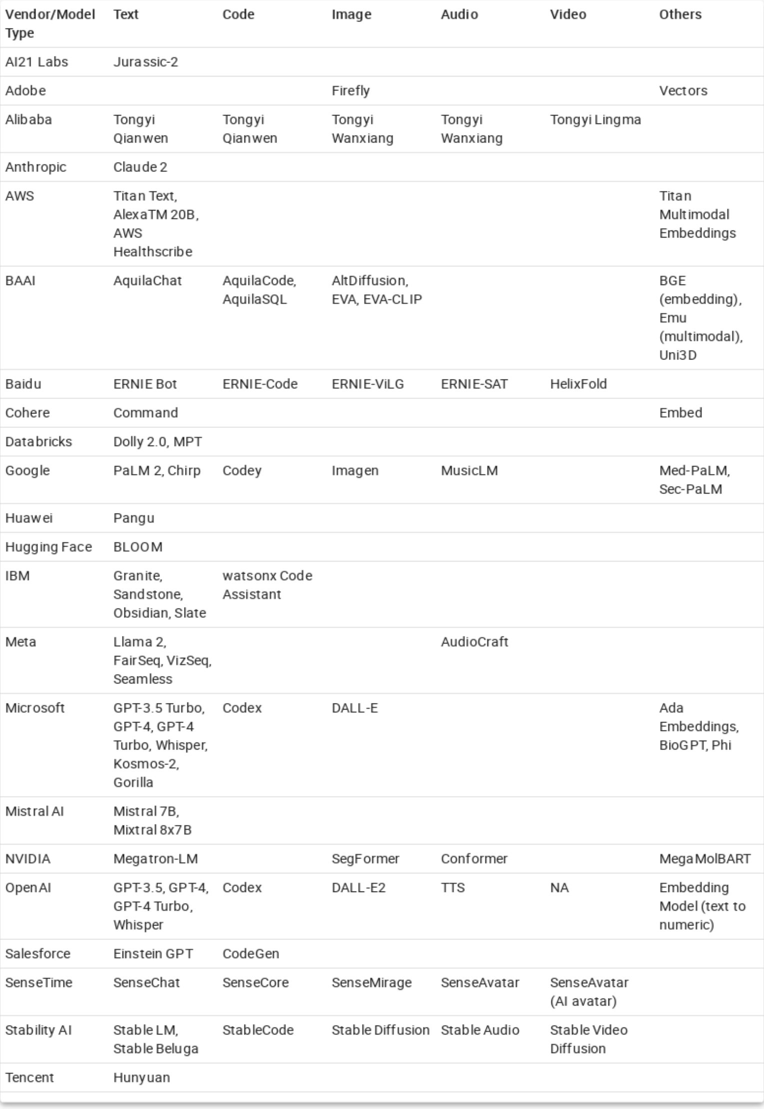
For risk mitigation strategies, see Innovation Guide for Generative AI in Trust, Risk and Security Management.

Representative Vendors

Back to top

Table 3: Vendors

(Enlarged table in Appendix)



Vendor/Model Type	Text	Code	Image	Audio	Video	Others
AI21 Labs	Jurassic-2					
Adobe			Firefly			Vectors
Alibaba	Tongyi Qianwen	Tongyi Qianwen	Tongyi Wanxiang	Tongyi Wanxiang	Tongyi Lingma	
Anthropic	Claude 2					
AWS	Titan Text, AlexaTM 20B, AWS Healthscribe					Titan Multimodal Embeddings
BAAI	AquilaChat	AquilaCode, AquilaSQL	AltDiffusion, EVA, EVA-CLIP			BGE (embedding), Emu (multimodal), Uni3D
Baidu	ERNIE Bot	ERNIE-Code	ERNIE-ViLG	ERNIE-SAT	HelixFold	
Cohere	Command					Embed
Databricks	Dolly 2.0, MPT					
Google	PaLM 2, Chirp	Codey	Imagen	MusicLM		Med-PaLM, Sec-PaLM
Huawei	Pangu					
Hugging Face	BLOOM					
IBM	Granite, Sandstone, Obsidian, Slate	watsonx Code Assistant				
Meta	Llama 2, FairSeq, VizSeq, Seamless			AudioCraft		
Microsoft	GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, Whisper, Kosmos-2, Gorilla	Codex	DALL-E			Ada Embeddings, BioGPT, Phi
Mistral AI	Mistral 7B, Mixtral 8x7B					
NVIDIA	Megatron-LM		SegFormer	Conformer		MegaMolBART
OpenAI	GPT-3.5, GPT-4, GPT-4 Turbo, Whisper	Codex	DALL-E2	TTS	NA	Embedding Model (text to numeric)
Salesforce	Einstein GPT	CodeGen				
SenseTime	SenseChat	SenseCore	SenseMirage	SenseAvatar	SenseAvatar (AI avatar)	
Stability AI	Stable LM, Stable Beluga	StableCode	Stable Diffusion	Stable Audio	Stable Video Diffusion	
Tencent	Hunyuan					

Source: Gartner (January 2024)

Version History

Version	Update Summary
v1.1	Minor update as of 5 January 2024 to include new startups and new models released in general availability by vendors.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Innovation Guide for Generative AI Technologies

How to Choose an Approach for Deploying Generative AI

4 Ways Generative AI Will Impact CISOs and Their Teams

How to Pilot Generative AI

Quick Answer: What Are the Pros and Cons of Open-Source Generative AI Models?

AI Design Patterns for Large Language Models

Generative AI: 4 Decisions to Make When Creating a Policy

Innovation Guide for Generative AI in Computer Vision

Quick Answer: How Will the Generative AI Plug-In Market Evolve?

Table 1: Measurements of Generative AI Models and Vendors

Measurements	Remarks
Model types	The models can be roughly divided into two groups: base models and fine-tuned models. Base models are usually used for general purposes and provide more potential for further fine-tuning. Fine-tuned models can be better aligned with human preference and/or more knowledge in specific domains, while creativity or performance might be degraded in others. Organizations have to make a trade-off between general applicability and domain specificity.
Basic capabilities	Benchmarks such as HELM in generative AI communities can be good references for models' basic capabilities. Model benchmark leaderboards such as Chatbot Arena, AlpacaEval and Hugging Face Open LLM Leaderboard are updated frequently. However, customized tests with an enterprise's own use cases are mandatory and should be prioritized over benchmark scores.
Prompt engineering	Prompt engineering is still the major approach for organizational adoption. Steerability (the models can consistently follow system instructions) is critical when LLMs are embedded in larger AI solution architecture such as RAG or when connecting to other tools, systems or APIs. Consider other features such as context learning, context window size, robustness and chain of thoughts as well.
Fine-tuning	Both base models and fine-tuned models can be further fine-tuned. The feasibility of fine-tuning is critical for organizations that would customize the models. Several types of fine-tuning serve different purposes: domain-specific

	extended pretrain, instruction tuning for alignment and task-specific fine-tuning. Lightweight fine-tuning approaches such as LoRA can be applied in instruction tuning to make fine-tuning easier.
Nonfunctional features	Model inference speed and infrastructure needs are key nonfunctional features for model selection. Vendor offerings also have a significant impact on model adoption: ■ Support for the organization's preferred deployment approaches (cloud, on-premises or edge) ■ Permissiveness of licensing (model code, parameters and training data can fall under different permissions of use) ■ Total cost of ownership (TCO) ■ Model transparency on training data and model size ■ API documentation readiness ■ Ease of model life cycle management
Ecosystems	Community and engineering tools support is another important factor to consider, especially when choosing an open-source model.

Source: Gartner

Table 2: Risks and Implications of Generative AI Models

Risks	What is it?	Implications
Loss of confidential information	Generative AI apps (not all but some) may use the input data (prompt) to train or improve the models.	Loss of confidential information as privileged information is revealed in model output
Hallucinations and misinformation	Generative AI apps make up facts or make glaring factual and reasoning errors.	Brand erosionLack of trust in AIPoor decision making
“Black box” responses	They are unable to provide a justification or credible source for their output.	Lack of explainability of AI model decisions
Intellectual property	The training datasets of these applications could be infringing on the copyright of others.	Potential legal infringement issues and lack of indemnification
Static information	Generative AI models are static models due to the gap between their training period cutoffs and actual release dates.	Lack of up-to-date or real-time information, which can limit the usefulness in certain use cases
Misuse and disinformation	Generative AI solutions can be manipulated to enable unethical and nefarious activities by bypassing safety controls via clever prompt engineering.	Erosion of brand equityRegulatory fines
Liability	Many generative AI vendors force users to indemnify the vendor as part of the terms of use.	Legal liability is passed on to the user of the service, rather than the provider
Ownership	Generative AI artifacts may not enjoy copyright protection, which provides a flimsy moat for competitive differentiation.	The lack of copyright protection would mean much of its creation becomes part of the public domain — free to use and copy

Bias	Generative AI systems can propagate downstream bias in the datasets, and the homogenization of such models can lead to massive network effects.	Deploying these models without bias mitigation could lead to regulatory fines and brand erosion

Source: Gartner

Table 3: Vendors

Vendor/Model Type	Text	Code	Image	Audio	Video	Others
AI21 Labs	Jurassic-2					
Adobe			Firefly			Vectors
Alibaba	Tongyi Qianwen	Tongyi Qianwen	Tongyi Wanxiang	Tongyi Wanxiang	Tongyi Lingma	
Anthropic	Claude 2					
AWS	Titan Text, AlexaTM 20B, AWS Healthscribe					Titan Multimodal Embeddings
BAAI	AquilaChat	AquilaCode, AquilaSQL	AltDiffusion, EVA, EVA-CLIP			BGE (embedding), Emu (multimodal), Uni3D
Baidu	ERNIE Bot	ERNIE-Code	ERNIE-ViLG	ERNIE-SAT	HelixFold	
Cohere	Command					Embed
Databricks	Dolly 2.0, MPT					
Google	PaLM 2, Chirp	Codey	Imagen	MusicLM		Med-PaLM, Sec-PaLM
Huawei	Pangu					
Hugging Face	BLOOM					

IBM	Granite, Sandstone, Obsidian, Slate	watsonx Code Assistant				
Meta	Llama 2, FairSeq, VizSeq, Seamless			AudioCraft		
Microsoft	GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, Whisper, Kosmos-2, Gorilla	Codex	DALL-E			Ada Embeddings, BioGPT, Phi
Mistral AI	Mistral 7B, Mixtral 8x7B					
NVIDIA	Megatron-LM		SegFormer	Conformer		MegaMolBART
OpenAI	GPT-3.5, GPT-4, GPT-4 Turbo, Whisper	Codex	DALL-E2	TTS	NA	Embedding Model (text to numeric)
Salesforce	Einstein GPT	CodeGen				
SenseTime	SenseChat	SenseCore	SenseMirage	SenseAvatar	SenseAvatar (AI avatar)	
Stability AI	Stable LM, Stable Beluga	StableCode	Stable Diffusion	Stable Audio	Stable Video Diffusion	
Tencent	Hunyuan					

Source: Gartner (January 2024)

Version	Update Summary
v1.1	Minor update as of 5 January 2024 to include new startups and new models released in general availability by vendors.