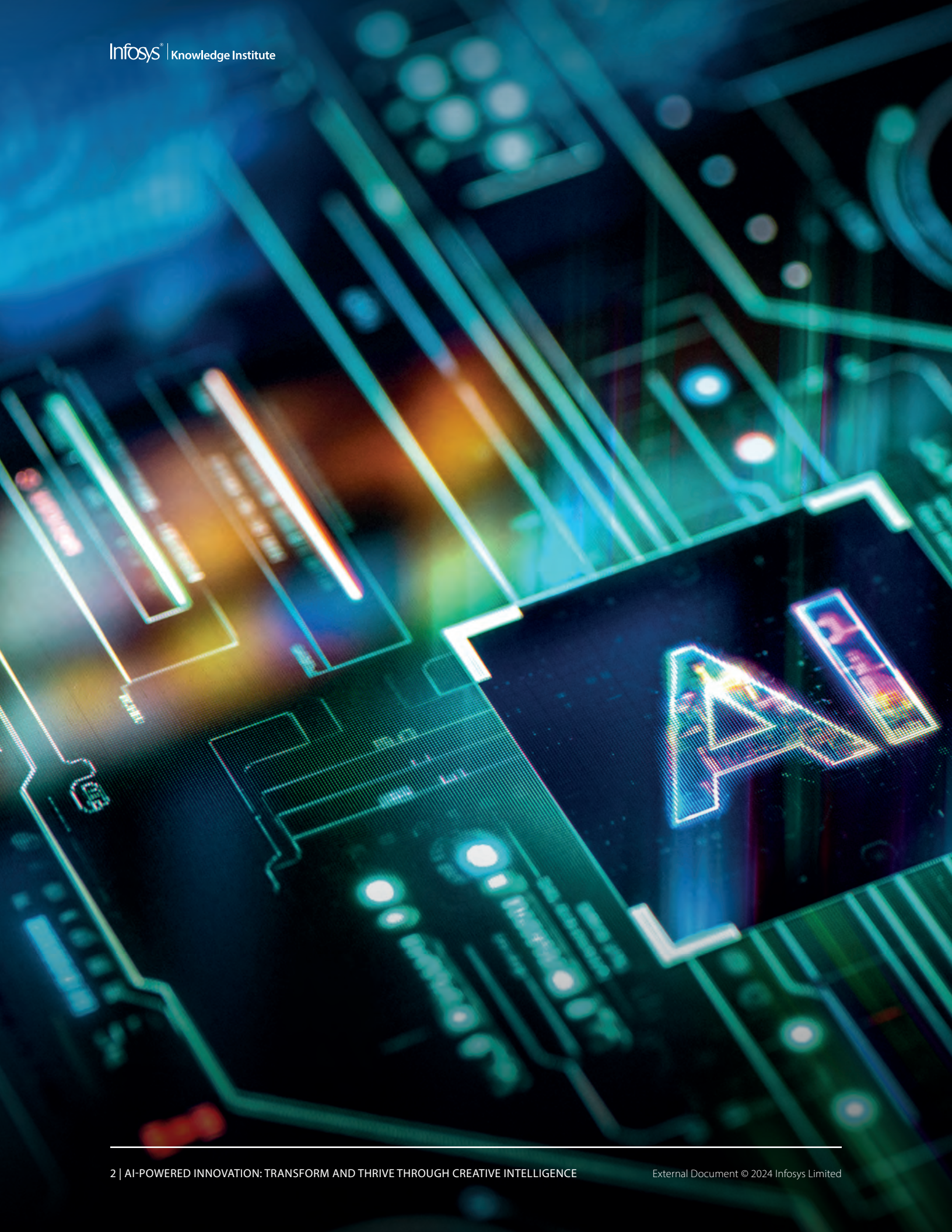


AI-POWERED
INNOVATION:
TRANSFORM AND
THRIVE THROUGH
CREATIVE
INTELLIGENCE



Contents

Shift toward pervasive AI	6
Generative AI	9
Natural language processing	12
Speech	14
Computer vision	16
Edge computing	18
Data engineering	20
Responsible AI	22
AI platforms	24
Advisory council	26
Contributors	26

Artificial intelligence powers forward through groundbreaking advances in machine learning, predictive analytics, deep learning, and the game-changing transformer architecture. The expanding influence of generative AI signals a seismic shift. Businesses break free from traditional AI roles and reimagine customer experiences and service delivery. The future is not just intelligent, it's creatively unpredictable.



Generative AI blurs the line between human and machine ingenuity. Over 85 large language models (LLMs) and [billions in global investments](#) propel this transformative progress. Businesses that integrate artificial intelligence (AI) in their digital and cloud investments gain significant competitive advantages, reshaping industries and transforming professions.

[AI-first enterprises](#) must develop strategic frameworks to identify experiences and processes that drive maximum benefit. Constructing a robust AI infrastructure, with the right tools, automation, and skilled talent, is essential to unlock the full potential

of emerging models. Equally critical is a conducive [product-centric operating model](#). Perhaps, the most important AI models are secure, private, and fail-safe. Given the potential pitfalls in AI deployment, a responsible AI strategy integrated into a platform reference architecture will significantly contribute to AI's rapid and fruitful scalability, whether generative or not.

In time, organizations will amplify human capabilities, foster innovation, unlock efficiencies at scale, accelerate growth, and cultivate a connected ecosystem.



SHIFT TOWARD PERVASIVE AI



Enterprises progress through the below three horizons for AI transformation:

Horizon 1 (H1) systems enhance existing systems with fragmented intelligence and leverage classical machine learning (ML) techniques. This boosts capabilities in predictive modeling (including classification, regression, and prediction tasks) and decision support (through recommendation engines and rule- or expression-based systems).

H2 systems are more complex, requiring higher-order generalization, accuracy, and learning capabilities. They rely on powerful deep learning tools, including

neural machine translation, conversational insights, object detection, facial recognition, and transfer learning.

H3 systems transcend the capabilities of H1 and H2, operating with an unprecedented level of sophistication and autonomy. They explore the potential of both closed and open LLMs for diverse applications. These systems are characterized by groundbreaking features, including multimodal processing, multilingual capabilities, responsible by design, unprecedented autonomy and collaboration via AI agents, and scalability and efficiency through model shrinking and data parallelism.

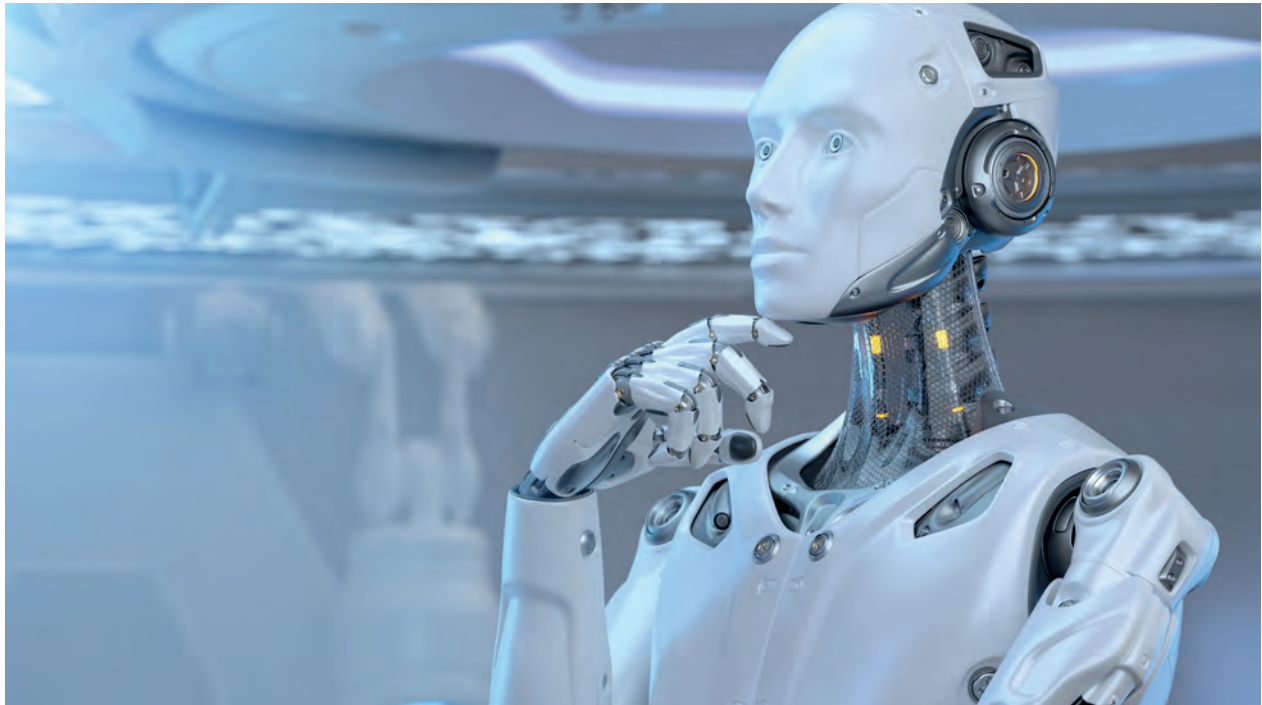
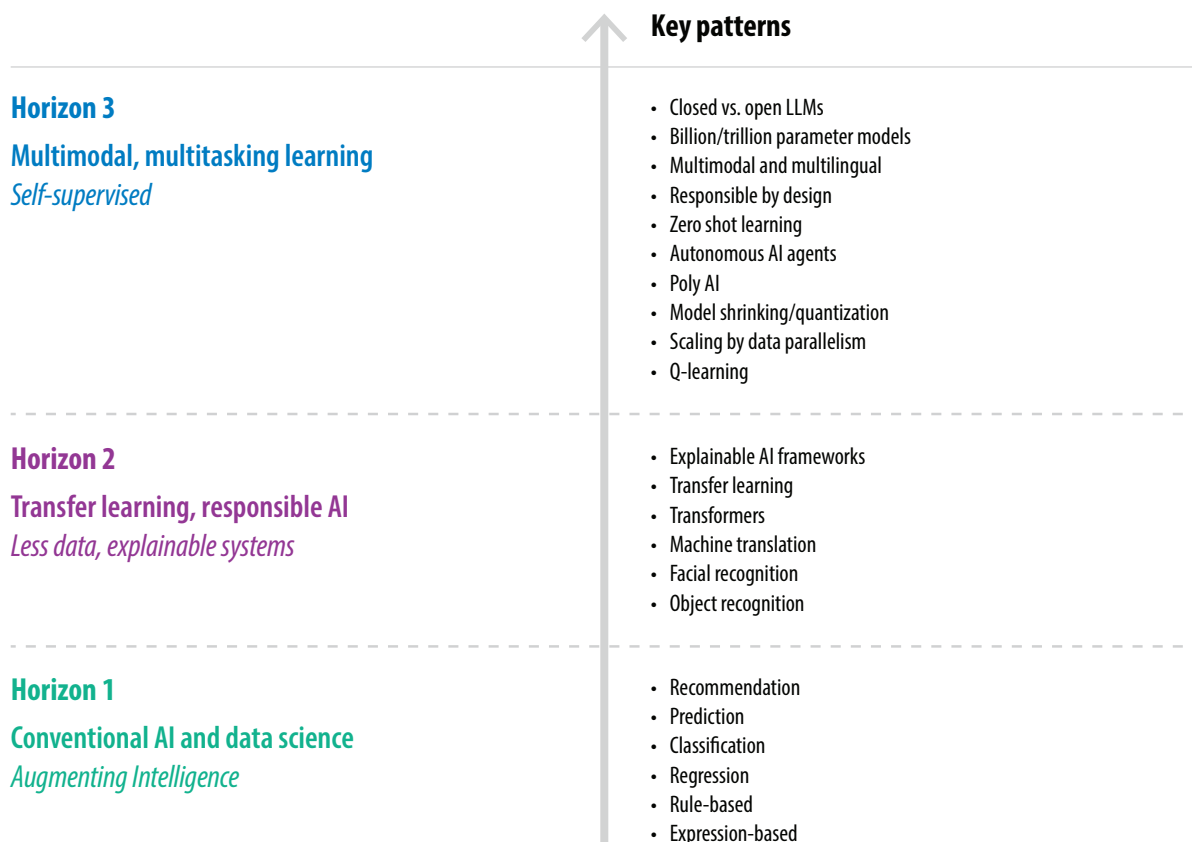








Figure 1. Market dynamics across the three horizons



Source: Infosys

Figure 2. Key trends across AI subdomains

 <p>Generative AI</p>	<p>Trend 1. Closed-source foundational models gain wider acceptance</p> <p>Trend 2. Organizations leverage LLMs for seamless code migration</p> <p>Trend 3. LLMs optimize knowledge management and semantic search</p> <p>Trend 4. LLMs improve productivity across the software development life cycle</p> <p>Trend 5. Autonomous agents shape the future of generative AI</p>
 <p>Natural language processing</p>	<p>Trend 6. Generative multimodal models become popular</p> <p>Trend 7. Instruction-tuned LLM multitasking facilitates fluid and natural conversations</p>
 <p>Speech</p>	<p>Trend 8. Language-neutral audio processing breaks language barriers</p> <p>Trend 9. Hyper-realistic speech generation and multimodal fusion transform AI experiences</p>
 <p>Computer vision</p>	<p>Trend 10. Compact multimodal intelligence revolutionizes industries</p> <p>Trend 11. Key point detection elevates retail experience</p>
 <p>Edge computing</p>	<p>Trend 12. Edge AI enhances efficiency and power across industries</p> <p>Trend 13. Generative AI shapes the future of edge</p>
 <p>Data engineering</p>	<p>Trend 14. Enterprise AI and decision making intensify with synthetic data</p> <p>Trend 15. Industrialized AI enhances data scientists' experience</p>
 <p>Responsible AI</p>	<p>Trend 16. AI security emerges as the bedrock of enterprise resilience</p> <p>Trend 17. Responsible AI guardrails uphold ethical AI</p>
 <p>AI platforms</p>	<p>Trend 18. AI democratization unlocks business potential</p> <p>Trend 19. Enterprise-level perspective for generative AI</p>

Source: Infosys

GENERATIVE AI



Closed-source LLMs are key in legacy modernization, knowledge management, and semantic search. Demand for GitHub Copilot and LLMs in software tasks has also increased. Still, open-source models remain popular, and benchmarking performance advancement in these technologies is important. Tools like LangChain, HuggingFace, Haystack, and LlamaIndex streamline LLM training, deployment, and usage, making them go-to for generative AI applications, even for nonspecialists. Even autonomous task execution and operational management workflows use LLMs like Baby Llama 2 (open source), BabyAGI, and AutoGPT. However, LLMs are prone to hallucinations and lack factual accuracy. To overcome, enterprises adopt generative models for coding tasks free from real-world knowledge influence. Forward-thinking firms employ retrieval augmented generation (RAG) for text tasks such as semantic search, as the facts are extracted and rephrased and not generated. RAG accesses external knowledge sources to complete tasks, enabling more factual consistency. It improves reliability of generated responses, reducing hallucinations.

Trend 1: Closed-source foundational models gain wider acceptance

Closed-source large foundational models like GPT-3.5/4/Turbo, Anthropic, PaLM, and Claude thrive on minimum infrastructure, which boosts their use cases across diverse segments. Enterprises can use closed models for industry use cases even without fine tuning, aided by retrieval augmented methods (which improves AI trust and quality), such as semantic search. They leverage a secure cloud ecosystem, trained on extensive public data, ensuring robust security and privacy controls.

These generic models perform multiple tasks, such as generating human-like language and using complex pattern identification to perform reasoning tasks. This broad capability facilitates applications such as document summarization and enterprise knowledge management.

A major tech company collaborated with Infosys to enhance its content moderation system. They implemented an AI model using supervised transfer learning for vision and text, enabling the identification and isolation of toxic content in user-uploaded forms.

Trend 2: Organizations leverage LLMs for seamless code migration

LLMs like GPT-4, trained in diverse codes and programming languages, excel in contextual inference, pattern recognition, and generalization. They interpret legacy languages and generate code in modern languages, facilitating migration from languages like assembly, Smalltalk, and C/C++. For legacy COBOL applications, crucial in many enterprise systems, maintenance and documentation challenges arise.

Manual processes are time consuming costly, particularly with a retiring cohort of COBOL experts. Generative AI systems proficiently analyze COBOL codebases, generating precise, readable documentation through advanced pattern recognition and natural language understanding. Variables, functions, and program flows are accurately described and defined.

A major American airline transitioned from a commercial integration and complex event processing platform to open-source tooling. This required migrating a large swathe of code written in a proprietary language to Java using a state-of-the-art LLM. This saved about 30% of the firm's effort.

Trend 3: LLMs optimize knowledge management and semantic search

Transformer models, incorporating embedding generation like OpenAI's GPT-3.5 and Microsoft's

E5 Large, precisely navigate industry contexts. Embeddings capture data's principal components, forming low-dimensional vectors that faithfully represent the original data. When combined with generative models (GPT-3.5/4, Llama, Falcon, Mistral-MoE, Flan-T5), they generate accurate, context-aware answers for end users, considering the enterprise context.

Retrieval augmented generation (RAG) is a prevalent strategy for use cases like knowledge retrieval, Q&A-based chatbots, summarization, and similar search experiences. Many organizations migrate their knowledge repositories to embedding-based vector databases, enhancing industry context semantic search processes and improving business and IT operation user experiences. For instance, in aviation, this approach trims 17% off aircraft repair design efforts, cutting airport downtime.

Trend 4: LLMs improve productivity across the software development life cycle

Companies increasingly explore LLMs to improve overall productivity across the software development life cycle, customizing user experiences through fine-tuned generative AI models. As our [Tech Navigator: Building the AI-first organization](#) discusses, data scientists are encouraged to use their preferred tools, combining open and closed AI models based on the enterprise use case.

The rising demand for GitHub Copilot and other LLM applications spans tasks like software requirement elicitation, code generation, documentation, unit test case creation, and general test case generation. In-editor experiences significantly boost both developer and tester productivity, elevating overall outcome quality. According to the annual [State of AI](#) report, using GitHub Copilot led to substantial productivity gains, with less experienced users benefiting the most — a productivity gain of approximately 32%. In a recent statement to Infosys, Nvidia CEO Jensen Huang said, "while some worry that AI will take their jobs, someone who is an expert in AI will certainly do so."

A large telecom company implemented fine-tuned open-sourced LLMs to generate Java, Python, Angular JS, .Net, and Shell scripting for developers. They used custom plugins for code editors to assist in code completion, documentation, and unit test case generation tasks, all while safeguarding sensitive information within the company network.

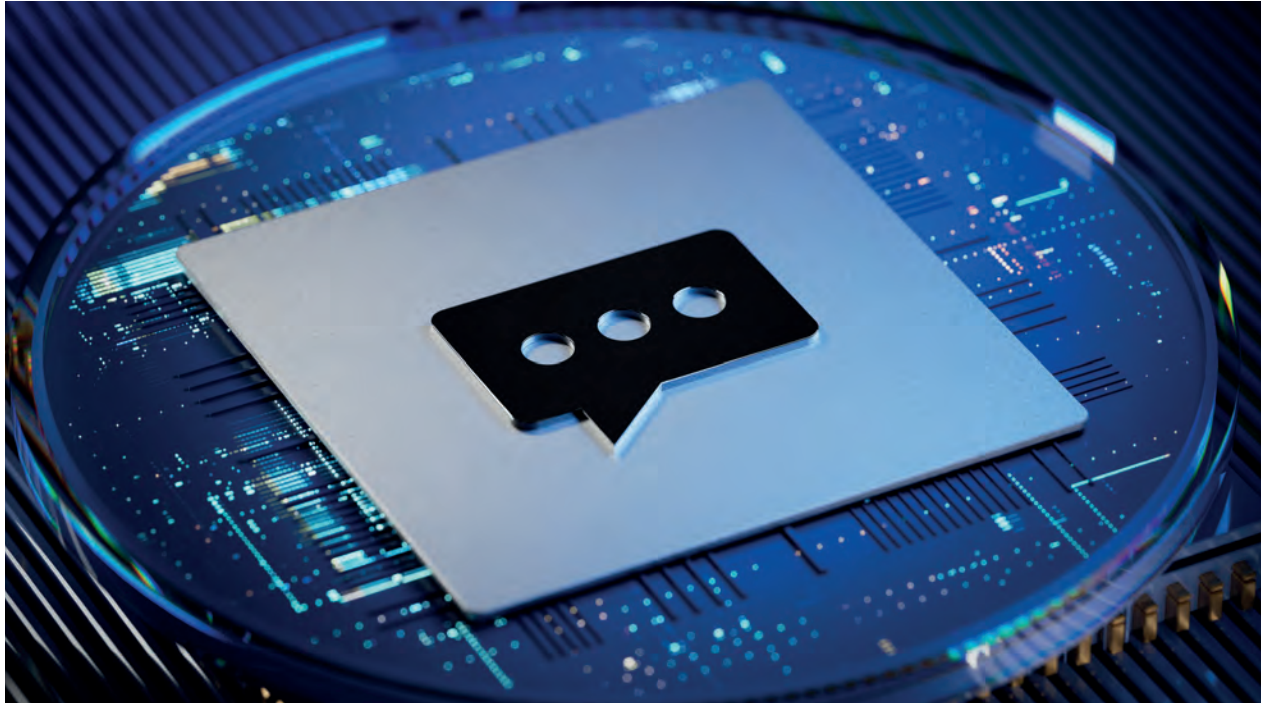
Trend 5: Autonomous agents shape the future of generative AI

AI agents are autonomous workers who independently execute tasks aligned with predefined goals and parameters. They are behind-the-scenes workhorses, independently accomplishing tasks — a contrast to tools like Copilot, which primarily

serve to aid and guide. In the realm of autonomous agents, the underlying LLMs rely on a model to plan and execute workflows, leveraging the capacity to identify and execute steps essential for complex tasks. This involves the execution of code using APIs to retrieve information from ERP systems. Autonomous agents find applications in tasks like invoice processing, involving data extraction from invoice documents, matching/validation against ERP systems, and subsequent invoice processing. Examples like Autogen, BabyAGI, and AutoGPT represent autonomous agents that serve as valuable buddies. They enhance roles such as IT support and function as knowledge assistants, leveraging the capabilities of LLMs, information extraction, and existing data tools within the organization. However, there are risks too. Autonomous agents are self-evolving, undergoing rigorous changes that may be hard to estimate and control. However, used carefully, autonomous agents are the next phase of generative AI, and will truly transform existing experiences and increase productivity.



NATURAL LANGUAGE PROCESSING



Natural language processing (NLP) is dismantling language barriers, enabling machines to understand and respond like intuitive partners. Beyond seamless interactions, NLP unlocks hidden insights from data, empowering smarter decisions and bridging information gaps. This is not just a tech upgrade; it's a paradigm shift, laying the foundation for truly human-like AI. NLP isn't replacing us; it's augmenting our intelligence, paving the way for a future where technology amplifies our potential.

NLP is undergoing a thrilling revolution, driven by two powerful trends: multimodality and multitasking. As technology evolves and data diversifies, these approaches play a crucial role in unlocking the true potential of NLP, enabling machines to understand and interact with the world more akin to human capabilities.

Trend 6: Generative multimodal models become popular

Initially, NLP focused on bag-of-words representation, emphasizing individual words with hot encoding-based sparse vectors. Then came models like GloVe and Word2Vec that addressed word similarity, but lacked contextual information, leading to biases.

Long-term short-term memory architectures (LSTMs) improved by capturing long-range dependencies but had slow processing times. In 2017, Google engineers introduced transformers, laying the groundwork for foundation models that excelled in diverse tasks. Attention mechanisms allowed LLMs to selectively focus on relevant parts of the input, dynamically allocating processing power to crucial information. This enabled them to handle complex and nuanced language, i.e., capturing long-range dependencies and context, going beyond surface-level analysis, thus surpassing limitations of earlier NLP approaches, crucial for tasks like question answering and summarization.

LLMs now excel across a range of modalities, such as image, audio, and video. Multimodal LLMs create rich, contextual, and highly accurate descriptions of multimedia content. These models comprehend sentiment in different media, with thought given to tone, emotion, and underlying implications used in prompts. State-of-the-art models like ImageBind link multiple modalities into a single embedding space, with Meta's ImageBind combining six modalities simultaneously: images, text, audio, depth, thermal, and inertial measurement unit data.

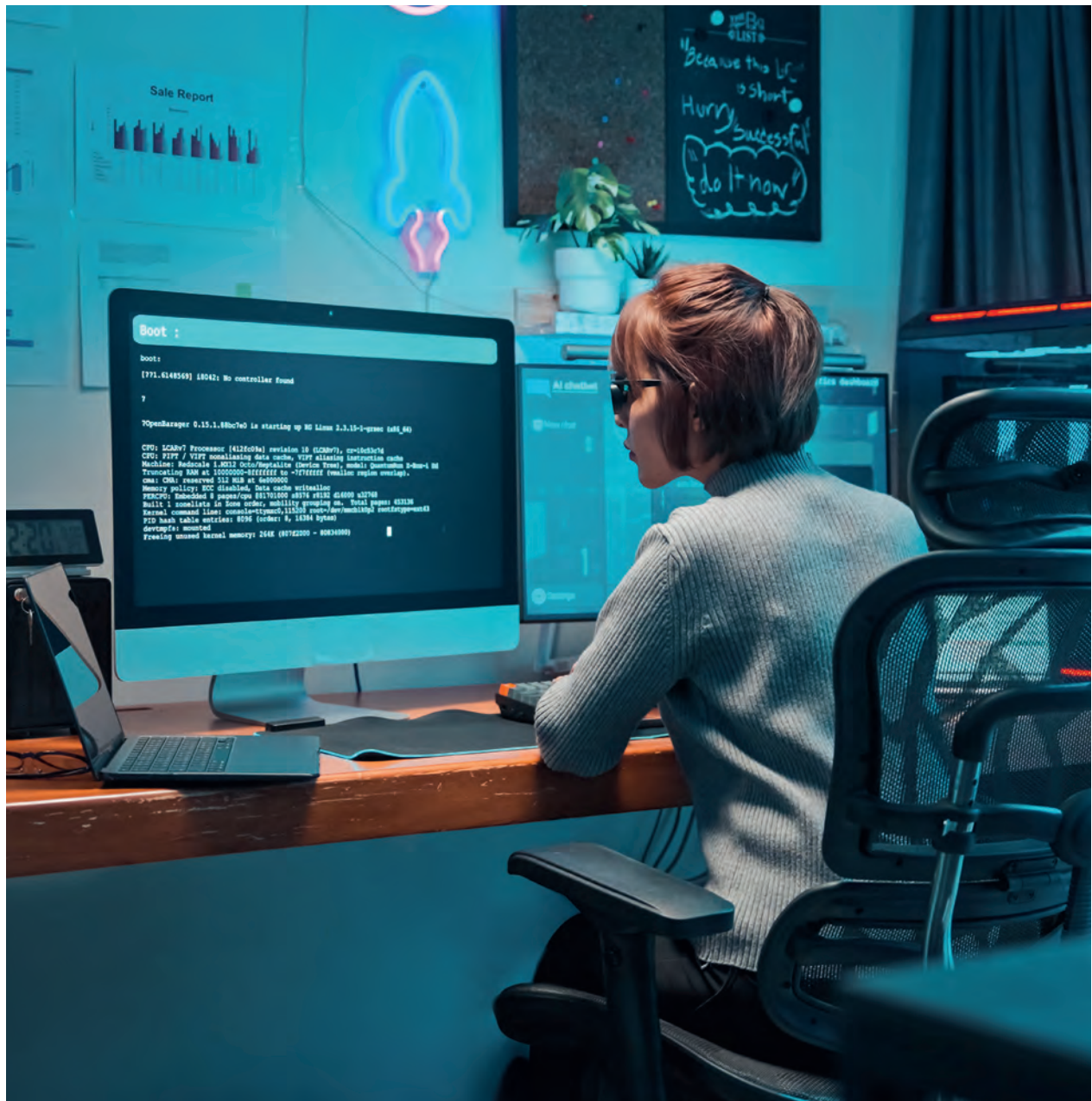
Trend 7: Instruction-tuned LLM multitasking facilitates fluid and natural conversations

LLMs, trained without specific labels and through reinforcement learning, effortlessly grasp insights from extensive text, smoothly adapting to new tasks and scenarios for enhanced flexibility.

However, instruction tuning enables LLMs to learn multiple tasks simultaneously. This method fine tunes pretrained LLMs with massive instructions

from multiple tasks. These trained LLMs (denoted as instructed LLMs) solve various unseen tasks in a zero-shot scenario without further fine tuning or demonstration examples.

While this approach enhances generalization skills, it incurs a computational cost due to extensive parameters. However, the result is more fluid conversations and potential operational cost reduction by minimizing resource-intensive API calls during inference.



SPEECH



Enormous real-time conversational data holds immense potential to derive intelligence and improve business offerings. It can be used for faster customer issue resolution, product feedback acquisition, cost reduction, workforce training, and much more.

Voice-based unstructured conversations emerge as a major intelligence source for enterprises, yet face technical challenges such as multilingualism, diverse vocabulary, accents, ambient noise, and varied recording channels.

Major players such as Microsoft, Google, and IBM have proprietary speech-based models that provide high-quality output, but they involve privacy concerns due to cloud-based operations and limited customization options.

Then, other open-source engines, including Kaldi, CMU Sphinx, Mozilla DeepSpeech, and Meta's wav2Letter, offer better customization. However, they involve different control levels, granularity of training data, effort requirements, and output accuracy. Altogether, the landscape has evolved, with open-source models becoming increasingly sophisticated and proficient in recent years.

Trend 8: Language-neutral audio processing breaks language barriers

Audio communication grappled with language barriers initially. Language-neutral audio processing aims to surpass spoken language limitations, making audio a universally understood exchange of information and emotion. Through real-time language conversion, it effortlessly bridges language gaps, featuring advanced technologies such as:

- Simultaneous audio translation: Leverages powerful audio models and neural machine translation models to convert spoken language into another language in real time, enabling fluid cross-lingual conversations.
- Universal speech recognition: Understands and transcribes spoken words across languages with ease.
- Multilingual voice assistants: Empowers voice assistants to handle multiple languages, catering to diverse user bases and creating a truly global voice interaction experience.

Generative AI drives this trend in large audio LLMs, neural machine translation, and automatic speech recognition. AI models trained on massive multilingual datasets constantly improve their ability to translate nuances and adapt to different conversational contexts. Businesses swiftly adopt these innovations, incorporating real-time language neutralization solutions into their products and services. They explore possibilities for multilingual voice interfaces, chatbots, and transcription services to reach a global audience. Additionally, they create content using universal audio symbols or nonverbal cues to effectively communicate with diverse audiences.

Trend 9: Hyper-realistic speech generation and multimodal fusion transform AI experiences

From robotic mimics to hyper-realistic speech generation (HRSG) — a breakthrough that infuses voices with life, producing near-perfect replicas rich in emotion, nuance, and individuality. Personalized voice cloning replicates voice with uncanny accuracy, making use cases for narrating audiobooks, guiding AI assistants, and creating virtual avatars that speak

specific words. AI now analyzes text for emotional cues and translates them into subtle variations using speech parameters like pitch, rhythm, and prosody. Built on speech algorithms that analyze and synthesize unique vocal characteristics, from pitch and timbre to microinflections and emotional nuances, HRSG creates indistinguishable digital twins. Businesses can use it to narrate stories with warmth, deliver presentations with authority, or convey specific emotions like joy, anger, or sadness. Imagine personalized narration for educational platforms, realistic customer service interactions, or even resurrecting the voices of historical figures.

AI, once confined to isolated domains processing speech, text, and visuals, has evolved with the ascent of multimodal fusion. Context-aware AI assistants understand surroundings, read text messages, and even sense emotional tone in voice. They analyze audio, visual, and sensor data, providing deeper context for superior customer service, personalized education, and enhanced healthcare. To harness these trends responsibly, businesses must train AI with diverse datasets and establish ethical guidelines for multimodal contexts.



COMPUTER VISION



Recent years saw a remarkable shift in computer vision, from traditional, CNN-only to transformer-based models, with foundational, diffusion, and language-guided methodologies.

Advancements in image synthesis, video processing, 3D reconstruction, and multimodality signal a transformative phase with implications for robotics. Anticipated breakthroughs extend to edge computing, driven by drones, UAVs, IoT, and the demand for lighter models in remote locations.

Major chip manufacturers and hyperscalers lead in developing specialized AI cores, mirrored in the proliferation of dedicated model optimization packages. This transformative wave not only shapes research but also increases practical applications across diverse domains, underlining the field's adaptability and potential to redefine machine perception.

Trend 10: Compact multimodal intelligence revolutionizes industries

Retail, logistics, healthcare, and manufacturing currently harness pretrained foundation models for computer vision tasks such as image classification, object detection, and segmentation. While these models offer rapid customization, they are large and often require substantial data for fine tuning.

Compact, task-agnostic models are poised to replace their data-hungry counterparts. They promise faster adaptation, increased accuracy, and less reliance on extensive data, making AI solutions more accessible and efficient across industries.

Integrating computer vision with other modalities, such as language processing, opens new horizons. Merging computer vision with robotics and human interactions holds potential to revolutionize

healthcare, autonomous vehicles, and manufacturing, creating intelligent systems that redefine industry standards and enhance our daily lives.

A US telecom giant partnered with Infosys to create an advanced object detection model on Android devices using computer vision. It enabled field engineers to efficiently evaluate installation or repair tasks, saving \$150,000 annually on repairs and gaining 900 hours per year. The optimized operational expenses and improved customer experience.

Trend 11: Key point detection elevates retail experience

Key point detection identifies and localizes specific points of interest in an image, including body pose, to analyze ongoing human behavior. This form of computer vision can be used by businesses to analyze interactions with products and provide actionable insights for improved customer engagement.

Current trends embrace convolution-based models, utilizing both top-down and bottom-up approaches. Despite challenges with viewing different angles in training datasets, the forthcoming integration

of next-generation foundation models promises to increase accuracy further. Advancements in the field are already being used to enhance ergonomic assessments in health and safety, transforming interactive gaming experiences. Future advancements will include integrating innovative algorithms such as 'track anything' and time-series forecasting, promising further accuracy in key point detection. Combining generative AI with these forward-looking algorithms promises a future with even more refined precision for human activity analysis, opening new avenues in health, safety, gaming, and more. As industries continue to harness computer vision technologies, this synergy of algorithms promises to redefine standards and applications, paving the way for a more accurate and versatile key point detection landscape.

Infosys' Retail Lab employs advanced body pose key point detection for a seamless shopping experience. By identifying key body movements like elbows, wrists, and fingers, firms gain in-depth insights into customer behavior and product interactions. This facilitates actionable insights and robust analytics, ultimately elevating retail experience.

EDGE COMPUTING



Edge computing minimizes latency, enhances efficiency, and reduces dependency on centralized cloud resources. It enriches applications such as 5G-enabled multiaccess edge computing, autonomous vehicles, aerial imagery analysis, biometrics, access control, defect inspection, and even smart spaces. It is also employed in generative AI applications, elevating personalization with a focus on latency, security, and reliability.

When is edge processing better than cloud-based inference? Consider the dynamic interplay of bandwidth, latency, economy, reliability, and privacy (in short, BLERP).

Edge is ideal with no/low connection and low latency needs like autonomous vehicles or AR applications. Edge reduces cloud computing expenses, especially if the customer is willing to invest in a capable edge device or if bandwidth is free. For critical reliability, as in a robotic medical assistant, edge processing becomes imperative. In privacy-sensitive applications, edge facilitates inference/prediction and training/learning securely, benefiting healthcare, manufacturing, and financial services.

Trend 12: Edge AI enhances efficiency and power across industries

Edge AI transforms AI deployment by prioritizing on-device processing, reducing reliance on central servers for instant decisions. This optimization significantly improves latency, response times, security, and bandwidth efficiency — crucial in applications like autonomous vehicles. Edge computing also minimizes data transmission, enhancing privacy in computer vision.

A delicate balance between model complexity and computational efficiency is crucial, especially for compact, portable edge devices like mobile phones, IoT, and drones. Private video analysis through real-time edge AI processing generates less but high-quality data.

Edge AI in remote healthcare and manufacturing eliminates the need for human operators, enables 'always-on computer vision,' and empowers firms to avoid sending video streams to the cloud, amplifying its potency.

Edge AI will process inputs from multiple modalities (vision, audio, text) and emerge as a transformative force in our interconnected world. This imminent progression promises heightened potency on the edge, paving the way for more efficient and powerful edge AI models across diverse industries.

A remote power generation entity partnered with Infosys to build a drone system featuring nimble, low-latency computer vision models. Tailored for vast power plants in hard-to-reach terrains, these models seamlessly integrate with drones' edge computing devices. Equipped with cameras and deployed computer vision models, drones deliver real-time data, empowering firefighting teams and authorities with swift, informed decision-making in remote and challenging locations.

Trend 13: Generative AI shapes the future of edge

The next development in edge computing is generative AI use cases. Earlier LLMs like GPT-3.5 or GPT-4 were too big for edge deployment. To optimize

performance, smaller models tailored for specific tasks emerged, making generative AI feasible for edge applications such as sentiment analysis, Q&A, and language translation.

With the rise of more powerful LLMs, there is a growing need for full-scale generative AI at the edge. Quadric, a California-based edge AI chip provider, is one such example. It recently unveiled compatibility between its neural processing unit IP core (Chimera) and Meta's Llama 2 model. Quadric is a key player in the chip provider landscape, particularly due to its support for LLMs. Anticipated proliferation of LLM implementations is driving advancements in mission-critical industries. For instance, autonomous decision-making and suggestions in warehouse environments in healthcare (privacy-preserving medical monitoring) and manufacturing.

In collaboration with Infosys, a manufacturing company developed a generative AI solution on edge for automated review of digitized engineering documents. The solution reads title block information, which helps associate metadata with digitized copies and fastens search and retrieval of relevant drawings from scanned data.

DATA ENGINEERING



Data engineering accelerates analytical decision-making, operationalizes business value through MLOps, and drives accurate decision-making on reliable data. As much as 80% of an AI project involves data cleansing, preparation and data engineering activities, making data engineering crucial for AI (sentience) and automation in data-centric services. In fact, a lack of effective data engineering projects can seriously curtail advancement in AI, including generative AI.

According to our [Data + AI Radar 2022](#), only 26% of executives are satisfied with their AI initiatives, largely due to a lack of enterprise readiness for data and AI. Firms must give data analysts and data scientists access to clean, high-velocity data that is fingerprinted for metadata and brought into compliance with applicable laws. One way of doing this is to create synthetic data using algorithms that create new data, which mimics real-world data. Also, they should provide data scientists access to self-service tools and frameworks through an internal development platform (IDP) to increase efficiency and speed in data engineering projects.

Trend 14: Enterprise AI and decision making intensify with synthetic data

Enterprise AI intensifies the reliance on data for decision-making, but finding suitable training data remains a challenge. Enter synthetic data — computer-generated, algorithmically crafted datasets filling gaps where real data is scarce, sensitive, biased, or poses privacy risks. It powers generative AI, robotics, metaverse, and 5G, extending into mission-critical scenarios like healthcare.

Some research even shows that synthetic data will completely overshadow real-world data usage by 2030. It enables robust ML algorithms, upholds GDPR standards, and respects cross-border data flows. However, human-centric considerations, including value, privacy, ethics, and sustainability, demand careful attention. Infosys advises building a synthetic data center of excellence to use synthetic data responsibly and effectively.

A prominent medical manufacturer and supplier partnered with Infosys to build synthetic data sets for product development and AI-based predictive analytics. These synthetic data sets can be shared and reused beyond the scope of initial collection, an essential component to accelerate research and product development.

The shift toward an industrialized ecosystem sees increased adoption of automated advisory for faster feature engineering and improved quality analysis. The emergence of industrial data scientists, empowered with increased data access and AI/ML tools, is transforming the landscape.

Key factors for AI empowerment include simplified infrastructure, deployment, domain expertise collaboration, and access to generative AI programming tools like OpenAI's Codex.

Trend 15: Industrialized AI enhances data scientists' experience

Data scientists, traditionally involved in manual data analysis and cleansing, lack standardized tools for wrangling, analytics, feature engineering, and model experimentation.



RESPONSIBLE AI



As AI integrates into daily lives, human wellness, fairness, transparency, and accountability become crucial. [Responsible AI](#) is not just a trend, it's a crucial aspect of AI development and deployment that is gaining attention from governments, organizations, and the research community.

While AI systems introduce biases and ethical challenges, they offer tools and technologies to address these issues. Responsible AI involves leveraging AI to design, deploy, and govern AI systems in alignment with ethical principles and societal values — an ongoing process that leverages AI to solve problems it may have inadvertently created.

In a legal scandal, lawyers found guilty of [submitting fake court citations](#) raise questions about the integrity of the legal profession and underscore the necessity for higher ethical standards in AI.

Trend 16: AI security emerges as the bedrock of enterprise resilience

Responsible AI is not only an ethical imperative but also a strategic advantage for companies looking to thrive in an increasingly AI-driven world. Rules and

regulations balance the benefits and risks of AI. They guide responsible AI development and deployment for a safer, fairer, and ethical AI ecosystem.

Staying informed about associated compliance requirements is vital for organizations and individuals, especially amid industry lawsuits related to generative AI. Novelist John Grisham has taken OpenAI to court, and programmers allege Codex crawl their work but don't [treat attribution, copyright notices, and license terms as legally essential](#). Meanwhile, visual artists have targeted Stability AI, Midjourney, and DeviantArt for copyright infringement. Regulations to significantly impact the industry's trajectory, shaping growth, innovation, market entry, and ethical practices. Businesses must proactively respond to evolving AI regulations, incorporating them into a strategy for compliance, trust, and long-term success.

AI's dual role presents a dichotomy: it can automate hacking, sidestepping traditional security, yet simultaneously bolster security through anomaly detection, threat prediction, and real-time monitoring. This interplay necessitates continuous adaptation to stay ahead of risks. AI security is vital for the integrity

of applications across domains, shaping the industry's future by addressing critical concerns like data privacy, integrity, and fairness.

Businesses should approach AI security with a multifaceted strategy by implementing robust security measures, regularly updating AI models, conducting risk assessments, educating employees, using AI-based security tools, addressing ethical considerations, and collaborating with experts. Businesses can also leverage AI-driven security tools for advanced threat detection, anomaly detection, and real-time monitoring, augmenting their overall security posture. Lastly, collaborating with AI security experts and staying updated on emerging threats and best practices is essential for proactive adaptation to the dynamic security landscape. These comprehensive measures collectively fortify AI security, instilling trust and resilience in AI initiatives while safeguarding systems and sensitive data.

Trend 17: Responsible AI guardrails uphold ethical AI

Responsible AI guardrails encompass a set of rules and best practices designed to ensure the ethical and responsible development, deployment, and utilization of AI systems. These guidelines address ethical, legal, and societal concerns associated with AI, aiming to mitigate potential risks and harms. The relationship between AI and responsible AI guardrails

is dynamic, with AI influencing and being influenced by advancements in technical guardrails and ethical guidelines.

In response to advancements in guardrails on the usage of AI, businesses should prioritize responsible AI practices that build trust, mitigate risks, and uphold ethical standards in AI development and deployment. Key practices include clear ethical guidelines, robust training, risk assessments, transparent AI models, strong data privacy, fairness considerations, human oversight, regulatory compliance, continuous monitoring, ethical AI committees, and public engagement, all contributing to responsible AI efforts.

Lex, Infosys' in-house learning platform, leverages generative AI with responsible by design. When users submit queries, a vigilant AI moderation layer swiftly and effectively enforces responsible learning practices, promptly addressing any deviations or infringements. This approach not only enhances the platform's responsiveness but also underscores its commitment to fostering a responsible and ethical learning environment.

AI PLATFORMS



Growing AI pervasiveness has redirected enterprises from a function-specific to a platform-based approach, ensuring unified development aligned with organizational requirements. Ethical AI use, compliance with regulations, and features like explainability are vital in AI platforms. Following a layered development principle, where each function is built independently, ensures speed and agility in onboarding new AI models. Scaling is achieved by embedding most of the data processing, consumption, and governance patterns into the platform for autonomous management. A poly AI approach allows transparent and consistent use of various tools and processes across multiple hyperscalers. This strategy creates a robust foundation for responsible and widespread AI adoption.

Trend 18: AI democratization unlocks business potential

Over two-fifths of APAC firms have implemented generative AI without realizing business value, compared with just [three-tenths in North America](#). Challenges in talent, expertise, poor data, lacking AI strategy contribute much. AI democratization

maximizes business value when the whole firm has access to AI tools, scripts, and frameworks. Low code, no code (LCNC) platforms, with visual interfaces, empower nontechnical users to build and deploy applications without complex coding.

Key AI platforms propel advancements in enterprise adoption. LCNC tools with intuitive visual interfaces democratize AI, fueling market growth from \$17.7 billion in 2021 to a projected \$125 billion by 2027. Prebuilt models for tasks like image recognition and automated workflows streamline implementation and reduce technical barriers. Cloud-based infrastructure enhances affordability and scalability, eliminating the need for costly hardware investments. Collaborative features facilitate secure cross-functional teamwork, fostering innovation. Additionally, explainable AI tools promote transparency, build trust, and empower users to make informed decisions based on AI insights.

A global firm partnered with Infosys to develop an LCNC solution tailored for its oil segment, automating more than 200 processes to streamline IT operations. For another client, Infosys built over 300 processes in just 10 months through Infosys IP and FastApp.

Trend 19: Enterprise-level perspective for generative AI

Generative AI requires a platform-based approach for enterprise-scale deployment, utilizing agile techniques to abstract engineering complexity. The platform should embody a forward-looking approach with a layered architecture, data readiness, and in-flow AI; embrace democratic elements such as unified visibility, self-service, and crowd-sourcing; and demonstrate scalability through cloud-native design, rapid adoption, and self-governance features.

The platform should capture evolving business needs, embracing responsible by design principles, with safety, bias, security, explainability, and privacy throughout the AI life cycle. This builds trust, ensures regulatory compliance, and addresses legal considerations. A poly AI approach ensures various tooling and processes are transparent, measured, and monitored homogeneously across multiple

hyperscalers. In a layered architecture strategy, each layer functions as an independent application with distinct user personas, interfaces, technology, services, and deployment.

A major US telecom company tackled fragmented AI development by implementing an enterprise-wide self-service AI platform. This platform fosters collaboration among data scientists, engineers, business analysts, and others, breaking down silos and enabling the development of cross-functional AI solutions across sales, customer service, and finance functions.



Advisory Council

Mohammed Rafee Tarafdar
EVP and Chief Technology Officer

Prasad Joshi
SVP, General Manager

Balakrishna DR
EVP, Service Offering Head

Satish HC
EVP, Co-head of Delivery

Dinesh HR
EVP, Co-head of Delivery

Shyam Kumar Doddavula
VP, Principal Product Architect

Thirumala A
SVP, Head - Education, Training, and Assessment

Venkata Seshu G
VP, Delivery Head

Rajeshwari Ganesan
Distinguished Technologist, STG

Kamalkumar Rathinasamy
Distinguished Technologist, STG

Contributors

Kaushal Desai

Dhiraj Dhake

Deepak Palasamudram

Akshatha Miyal Kamath

Pramit Saha

Amit Kumar

Vishal Manchanda

Dr. Puranjoy Bhattacharya

Shyam Kumar Doddavula

Rajeshwari Ganesan

Kamalkumar Rathinasamy

Swaminathan Natarajan

Sidharth Subhash Ghag

Jagadamba Krovvidi

Syed Ahmed

Bhumika Mahajan

Karthik Andhiyur Nagarajan

Manjula Natarajan

Ramjee R

Producers

Ramesh N
Infosys Knowledge Institute

Harry Keir Hughes
Infosys Knowledge Institute

Pragya Rai
Infosys Knowledge Institute

About Infosys Knowledge Institute

The Infosys Knowledge Institute helps industry leaders develop a deeper understanding of business and technology trends through compelling thought leadership. Our researchers and subject matter experts provide a fact base that aids decision-making on critical business and technology issues.

To view our research, visit Infosys Knowledge Institute at infosys.com/IKI or email us at iki@infosys.com.

For more information, contact askus@infosys.com



© 2024 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and / or any named intellectual property rights holders under this document.

