IBM Technology Atlas

AI roadmap

Large-scale, self-supervised neural networks, which are known as foundation models, multiply the productivity and the multimodal capabilities of AI. More general forms of AI emerge to support reasoning and commonsense knowledge.

Updated July 2024

- ✅ completed
- ⊖ pushed to next year
- ↻ on target

	2023	2024	2025	2026	2028	2030+
AI journey	✅ ***Extend foundation models beyond natural language processing.***	↻ ***Build modular and multimodal transformers for new enterprise applications.***	***Alter the scaling of generative AI with neural architectures beyond transformers.***	***Bring robust, strategic reasoning and commonsense knowledge to AI.***	***Develop broadly intelligent agents that learn autonomously.***	***Build adaptable and generalist AI for effective human-machine collaboration.***
Strategy overview	✅ In 2023, we will expand enterprise foundation model use cases beyond natural language processing (NLP). ⊖ 100B+ parameter models will be operationalized for bespoke, targeted use cases, opening the door for broader enterprise adoption.	↻ We will deploy enterprise AI assistants and applications using advanced transformers and developer-friendly frameworks to facilitate processing richer contextual information and enhanced control and monitoring of generative AI.	We will use a diverse selection of neural architectures beyond, and including, transformers that are co-optimized with purpose-built AI accelerators to fundamentally alter the scaling of generative AI.	We will support faster learning and the ability to provide explanations through better introspection, retrospection, and different forms of reasoning.	We will build autonomous AI that learns reliably and efficiently from its environment and responds to previously unseen situations through broad generalizations. These AI systems will start exhibiting aspects of biological intelligence.	Our AI models will be composed of modules with different cognitive abilities (e.g., perception, memory, emotion, reasoning, and action), enabling them to exhibit behavioral norms for social interactions and mutual theory of mind.
Why this matters to our clients and the world	✅ The expansion of AI foundation models will lower the barrier for entry, broaden the use cases, reduce labeling requirements for training by 10-100x, and provide greater efficiencies through reuse of models across use cases.	↻ LLM applications will broaden their applicability by integrating more easily with the core enterprise systems. ↻ Agentic assistants and applications will tremendously boost enterprise productivity.	Use case-driven, end-to-end optimizations, from transistors to neurons, will make a vast range of trade-offs available for energy consumption, cost, and deployment form-factors of AI, unlocking its potential at an unprecedented scale.	AI systems capable of fact checking and reflective thinking are faster and more accurate learners and planners. They will earn trust in real-world situations via demonstration of cognitive capabilities.	AI will be capable of continually and efficiently learning from multimodal input about how the world works. Those systems will learn to operate effectively even under uncertainty and develop problem-solving skills.	By being able to predict, act, plan, and adapt to new situations and environments, these unified neural architectures will enable a broad variety of use cases that require effective human-machine collaboration.
The technology or innovations that will make this possible	✅ Prebuilt models, workflows, toolchains, and multimodal neural architectures will leverage foundation models over diverse domain-specific data such as code, IT, security, geospatial, and materials. ✅ OpenShift-based cloud-native middleware will help scale foundation model workloads to thousands of GPUs.	↻ Transformer architectures will become modular and multimodal. They will be augmented with episodic memory for efficient knowledge updates. ↻ We will develop LLM-oriented orchestration and compositional frameworks along with modules for AI alignment, trust guardrails, and generative-AI-specific monitoring and risk assessment.	Novel neural building blocks will supplement traditional transformer blocks and create more efficient neural architectures. We will develop an open and community-governed foundation model software stack capable of exploiting accelerator-specific innovations for more efficient generative AI deployments.	Advances in reasoning-focused architectures will be integrated with learning modules. Slow-learning systems will be combined and controlled with world models that rely on fast learning such as an episodic memory module. Advances in planning techniques will enable the evolution of AI systems towards their end goals.	Agents augmented with multiple memory systems and multiple neural mechanisms will interact autonomously with each other. Hybrid neural architectures will rationalize over constantly evolving information about the world. They will learn to refine their multi-scale world model and develop generalizable skills for complex problem-solving.	Memory encodings of different sensory perceptions (e.g., visual, olfactory) will make AI weigh rewards and threats, safely interact with the world, and find optimal ways to achieve goals. Algorithms will be combined with hardware to natively support heterogeneity in neurons and neural connections in a similar fashion to biological intelligence.
How these advancements will be delivered to IBM clients and partners	✅ Watsonx will be launched with three elements: watsonx.data, watsonx.ai, watsonx.governance. ✅ The infrastructure will include resource- and topology-aware OpenShift clusters, and advanced networking between nodes and GPUs within a node.	↻ Watsonx will introduce more advanced modular LLMs along with developer-friendly application builders and governance features to accelerate development and deployment of AI applications. ↻ Watsonx assistants will seamlessly integrate code and language to provide out-of-the-box productivity tools.	Watsonx assistants will incorporate multiple AI agents targeted for different tasks and corresponding data modalities. Watsonx will support a variety of cost-effective accelerators in its deployments.	Watsonx will display cognitive characteristics, broadening its deployment to scenarios that require high trust in systems.	Watsonx will display characteristics of combined cognitive and emotional intelligence. Watsonx will support autonomous and broadly intelligent agents with appropriate trust guardrails.	Watsonx will support effective human-machine and machine-machine collaboration.