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 January 2024

✓ completed
→ pushed to next year
↻ on target

	2023	2024	2025	2026	2028	2030+
AI journey	✓ ***Extend foundation models beyond natural language processing.***	↻ ***Build multimodal, modular 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 autonomous and broadly intelligent agents.***	***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 assistants and enterprise applications using transformers that process richer context and large language model (LLM)-oriented frameworks which provide better 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 to mission-critical use cases and integrate more easily with the core enterprise systems. Generative AI 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.	More robust and explainable AI capable of fact-checking and reflective thinking is a faster and more accurate learner and planner. It earns trust in real-world situations via demonstration of cognitive capabilities.	AI will be capable of continually learning how the world works in an efficient manner and operate effectively even amid uncertainty.	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 be improved to be multimodal and modular with decoupled memory. Larger (200B+) models will be trained on better quality and larger datasets. ↻ We will develop LLM-oriented orchestration and composition frameworks with modules for AI alignment, trust guardrails, and LLM-specific monitoring and risk assessment.	Novel neural building blocks will transcend traditional attention mechanisms in transformers. Our open foundation model software stack will be capable of exploiting accelerator-specific innovations for more efficient and capable AI. We will automate the composition and optimization of LLM applications based on user-specified criteria and constraints.	Advances in reasoning-focused architectures will be integrated with long-term memory modules. Neural systems will combine their world knowledge with reasoning and planning to strategically move toward their end goals.	Neural networks will be augmented with multiple memory systems (e.g., working, episodic, semantics), and multiple neural mechanisms will interact autonomously with each other. Neural modules will rationalize over the stored information and incoming data streams to flexibly improve their multi-scale world model.	Different streams of sensory information (e.g., visual, olfactory), memory encodings, and rationalization pathways will make AI weigh rewards and threats and interact with the world in precise ways to achieve goals. Algorithms will be combined with hardware to natively support heterogeneity in neurons and neural connections.
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 models along with new application enablement and governance features to accelerate development and deployment of AI applications. ↻ Watsonx assistants will seamlessly integrate code and language to provide out-of-box productivity tools.	Watsonx assistants will incorporate multiple AI agents targeted for different data modalities and tasks. Watsonx will support a variety of cost-effective devices 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.