IBM Technology Atlas

AI roadmap

Large-scale self-supervised neural networks, i.e., foundation models, multiply the productivity and the multi-modal capabilities of AI. More general forms of AI emerge to support reasoning and common-sense knowledge.

	→ on target 2023	→ on target 2024	2025	2027	2029	2030+
AI journey	***Foundation models extends beyond natural language processing***	***Governance and trust permeate AI***	***AI becomes more energy and cost efficient***	***Foundation models in production become scalable***	***Trustworthy and explainable AI starts to reason***	***Fully multi-modal AI gives enterprises unprecedented scale***
Strategy overview	2023 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 to broader enterprise adoption.	In 2024, we will integrate trust guardrails throughout the AI foundation models lifecycle and AI governance at the organizational level. Data representations will optimize across privacy, fairness, explainability, robustness, etc.	In 2025, we will improve the energy and cost efficiency of foundation model training and inference by 5x and bring 200B+ parameter foundation models to enterprises. It's all about making them more powerful, useful, and practical.	By 2027, we will be routinely doubling the number of foundation model parameters in production for the same energy envelope every 18 months. Training and inference will be 4x more energy efficient vs. 2025.	2029 will be an inflection point. AI will support diverse forms of reasoning with explainability and trust. Energy efficiency will increase 4x more and scalable, operationalized AI models will be routine in enterprises.	By 2030 and beyond, fully multi-modal architectures will learn diverse data representations, and developers will be able to manipulate them at multiple levels of abstraction to give enterprises competitive advantage.
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.	Governance and trust in AI will drive significant business value by enabling advanced AI in mission-critical use cases while increasing automation, improving the quality of AI regulatory compliance, and maintaining customer trust.	Reducing cost and improving power efficiency will make foundation models more practical for enterprise workloads and promotes deployment across many enterprise use cases.	Foundation models will scale with growing data volumes and model complexity. watsonx's full stack infrastructure will be able to support enterprise data warehouses based on foundation model representations.	2029 AI means robust and explainable NLP, reasoning over text, and generation of knowledge and trustworthy natural language to advance mission critical use cases. Tuning of domain-specific AI models will be 10x faster and cheaper.	By being able to predict, act, plan, and adapt to new situations and environments, these unified neural architectures — grounded in biological brains — will broaden use cases even further while reducing energy consumption.
The technology or innovations that will make this possible	Prebuilt models, workflows, toolchains, and multimodal neural architectures will leverage foundation models over enterprise data such as Industry 4.0 data, transactions, IT and security data, geospatial data, code, and materials. Serverless pipelines will accelerate development. OpenShift based cloud-native middleware will help scale to 1000s of GPUs.	watsonx's middleware will instrument diverse AI guardrails. Trust benchmarks, metrics, audits, and repairs for foundation models will be developed and included in watsonx.governance, with a framework for auditable generation of high-quality synthetic data for watsonx.data. Human-centered AI experiences will allow learning and communicating trust policies and parameters, and expertise.	An AI inference accelerator that leverages the IBM Research "Gen 3" core and supports 255 virtual functions will enable rapid changing between AI models. AI inference systems based on this accelerator will include innovations in power and cooling.	New foundation models and hybrid architectures will facilitate multi-modal representations and dual-type processing. Algorithmic innovations will make training, adaptation, and compression energy efficient. Network design and performance, system topology, and protocols will be advanced. Hardware and software will be codesigned and watsonx's full stack optimized.	Advances in reasoning-driven attention mechanisms and network architectures will support reasoning tasks (causal, symbolic, compositional, common-sense, etc.). Reasoning and learning architectures will be integrated through representations that enable seamless communication between learning and reasoning modules in a network.	Advances in selective attention mechanisms will form a flexible model representation. Full multi-modality will be enabled by novel memory encodings and rationalization pathways, topologically regularized training, and advances in hardware architectures that natively support heterogeneity in neurons and neural connections.
The platform or infrastructure will these advancements by delivered on	watsonx has three elements: watson.data, watson.ai, watson.governance to allow clients and partners to specialize and deploy models for enterprise use cases in different domains or build their own models. The infrastructure will include resource and topology-aware OpenShift clusters, and advanced networking between nodes and GPUs within a node.	Optimal servers will be developed with a focus on AI accelerators and rack designs that improve thermal efficiency. Serverless middleware will support the new AI accelerators in watsonx.ai. New watsonx OpenShift operators will improve workload performance, user productivity, and administrator flexibility.	watsonx will have new middleware capabilities supporting heterogenous AI infrastructure — including GPUs, AI accelerators, and CPUs. AI systems will be deployed in the cloud with novel AI accelerators and accessible through watsonx.ai. Systems and platform will be codesigned with each other to increase cost efficiency.	A more powerful watsonx.ai will be available with more advanced infrastructure incorporating new neural architectures codesigned with AI accelerators and the software stack to support more capable models within the same energy envelope.	watsonx will support emerging architectures based on the principles of biological neural computations.	watsonx will leverage new types of computations, including quantum computing.