IBM watsonx.ai

Buy What You Need

1. watsonx.ai	Small	Medium	Large
Inference	166M tokens/month Class 3 model ($0.005)	1.2B tokens/month Class 2 model ($0.0018)	3.9B tokens/month Class 2 model ($0.0018)
Tuning	No tuning for this example	5 hours of tuning / month ($24/hour)	12 hours of tuning / month ($24/hour)
Saas Plan Tier Fee	Essentials: No fee	Standard: $1050/month	Standard: $1050/month

2. watsonx.discovery	Small	Medium	Large
vCPU	2	4	8
RAM	4	12	24
Disk	100 GB	200 GB	600 GB

3. Object Storage	Small	Medium	Large
Size	1 TB	2 TB	4 TB

Value Proposition

01 **AI Model Development and Deployment:** watsonx.ai provides an integrated environment to build, train, and deploy AI models, whether you are working with traditional machine learning models, deep learning, or even large-scale foundation models (like generative AI). It supports seamless collaboration between data scientists, developers, and AI engineers.

02 **Pre-built and Custom AI Models:** watsonx.ai comes with a repository of pre-trained models for common use cases such as natural language processing (NLP), image recognition, and timeseries analysis.

03 **Open-Source Frameworks:** It integrates with popular open-source tools such as TensorFlow, PyTorch, and Scikit-learn, allowing data scientists to use familiar environments and libraries while benefiting from Watsonx.ai's robust infrastructure and scalability.

04 **Cloud-Native Scalability:** Built on a cloud-native architecture, watsonx.ai provides the ability to scale AI workloads up and down based on demand. This makes it ideal for handling large datasets and complex AI models without compromising performance.

05 **Foundation Model Support:** One of the standout features of watsonx.ai is its support for largescale foundation models (e.g., generative AI models like GPT and BERT), which can be finetuned and applied to various business-specific tasks, including content generation, customer service automation, and data analysis.

06 **Data Integration and Automation:** The platform integrates with data sources both on-premise and in the cloud, leveraging IBM's data tools like watsonx.data and DataStage for efficient data preparation and pipeline management.

Customer Experience and Operational Efficiency of Knowledge Workers remain a Key challenge in many organisations. AI with LLMs trained on public data may not be enough to have an efficient and reliable alternative since it would lack business context and might have outdated information.

Many models do not have mechanisms of enhancing the customer experience by tapping into internal and external knowledge base. Need for a reliable Generative AI which can leverage proprietary data effectively and enable Faster and accurate decision making.

Watsonx.ai is IBM's cutting-edge AI and data platform designed to help businesses rapidly build, train, and deploy AI models at scale. With integrated open-source tools, pre-built models, and advanced AI governance, it empowers organizations to harness real-time insights and drive smarter decisions.

Whether you're optimizing operations, improving customer experiences, or unlocking new opportunities, Watsonx.ai delivers scalable, trusted AI solutions that accelerate innovation across industries.

Streamline AI application development with watsonx.ai

Enterprise-grade AI studio helps AI builders innovate with APIs, models, tools and runtimes

A well-considered approach to AI can help you scale and operationalize faster and more effectively. Generative AI (gen AI) provides scalability through foundation models that are trained on unlabeled data. Traditional AI and machine learning (ML) techniques offer fine-tuning and customization, using labeled data for improved accuracy. But what's the best way to simultaneously use both of these approaches?

IBM® watsonx.ai™ is an enterprise-grade AI studio for builders— from the application developer to the data scientist and everyday line-of-business user. As a one-stop AI developer studio, watsonx.ai combines gen AI with traditional ML techniques to help builders innovate with all the APIs, models, tools and runtimes to simplify and scale the development and deployment of AI applications.

Within watsonx.ai, AI builders can support the rapid adoption of AI use cases, from data through deployment, by using a collection of foundation models. These include IBM Granite™ models; a Prompt Lab interface and APIs to support agentic and retrieval augmented generation (RAG)-based use cases with or without code; a data science toolset to build AI and ML models automatically as well as a collection of powerful visual data pipelines and flows; and synthetic data generation—all running on a scalable, open and trusted hybrid AI infrastructure.

Leverage an enterprise-grade developer AI toolkit

The comprehensive suite of developer-focused capabilities preconfigured software development kits (SDKs), APIs, agentic workflows, RAG frameworks and templates, advanced tuning methods and more helps bring your AI application development process together in natural language or code.

Manage the full AI lifecycle

Accelerate and manage the full AI model and application lifecycle with easy-to-use tools for model training and gen AI development. With watsonx.ai, AI builders have access to MLOps pipelines and AI runtimes, benchmarks and guardrails all in one place to build powerful gen AI applications with trusted data and built-in governance that they can effectively manage.

Bring traditional AI and ML into production faster

Build AI and ML models automatically with model training, development and visual modeling, and synthetic data generation. You can also develop predictive and prescriptive models, code in Python Notebooks or RStudio, or work directly in your integrated development environment (IDE) of choice.

By leveraging both gen AI and traditional AI and ML techniques watsonx.ai is changing the game of AI app development.

Request for a Proposal →