ORACLE
NETSUITE

BUSINESS GUIDE

Oracle and Generative AI

A Brief Overview of Generative AI Capabilities in Oracle Cloud

Grab a seat and enjoy.
Read Time: 7 minutes

Oracle and Generative AI

A Brief Overview of Generative AI Capabilities in Oracle Cloud

The massive public enthusiasm for ChatGPT clearly demonstrates that the hype cycle for generative AI is in full swing. Oracle has provided AI throughout our product portfolio for years and, like many technology providers, has a growing set of generative AI capabilities.

Oracle's AI Strategy

Oracle's AI strategy is to make artificial intelligence (AI) pervasive across its cloud applications and cloud infrastructure. We serve business users who want to improve business processes and outcomes through prebuilt AI capabilities, data scientists and developers who want complete control to build and deploy AI models of any kind, and independent software vendors (ISVs) who want the most performant and cost-effective platform to host their AI services.

Our AI strategy is to build an AI stack focused on enterprise success, and it has three principal tenets:

1. Transformative AI foundation: Gain superior performance and scalability for AI training with NVIDIA GPUs, nonblocking remote direct memory access (RDMA) networks, and locally attached solid-state drives.
2. End-to-end data life cycle: One modern data platform to collect, curate, and manage AI/ML data assets.
3. Oracle business apps integrations: Easily add AI to business processes with application resources to gain faster insights.

Generative AI

There are many use cases for AI and machine learning (ML), and Oracle has offered products and solutions for many years that cross the full spectrum of AI. This document, though, focuses on just one branch of AI referred to as "generative AI," which is a category of AI algorithms that use neural networks like large language models (LLM) and other models like generative adversarial networks (GAN) and transformers to generate new outputs based on the data they have been trained on. Unlike traditional AI systems that are designed to recognize patterns and make predictions, generative AI creates new content in the form of images, text, audio, code, and more. "Deep fakes" are an example of generative AI: images, video, or audio that are generated by the AI engine. Generative AI requires significantly more computing power and data than traditional AI because of the complex algorithms and models involved in creating new data.

Generative AI has received a lot of recent interest because of the release of applications in which it is embedded, such as ChatGPT, which generates new conversation text, and Dall-E, which generates images. In early 2023, Microsoft announced generative AI embedded in multiple software products, including Microsoft Dynamics, Office 365, and GitHub. Google quickly countered with announcements about new generative AI offerings. All of this indicates that generative AI is an important branch of AI and will have increasing impact for years to come.

It's important to point out that generative AI is just one branch of AI — it's not the only AI, and it's not "better" than other kinds of AI.

Traditional AI like those listed below will continue to provide useful services regardless of whether they use generative AI models like GANs or transformers.

- Chatbots*
- Document understanding*
- Computer vision*
- Speech recognition*
- Code generation*
- Medical imaging*
- Machine translation*
- UX personalization*
- Synthetic audio, images, video
- Text analysis*
- Language modeling*
- Sentiment analysis*

As of this writing, Oracle has not announced plans to offer consumer-oriented generative AI applications. Our primary focus in the generative AI space is on providing high-performance, low-cost infrastructure and tooling for developers and ISVs to train and deploy their generative AI models. We also have a longstanding commitment to embed AI across our product lines to deliver the same kinds of use cases that generative AI is being used for (whether or not a generative AI model is used).

Oracle offers:

- Infrastructure for generative AI model deployment and training
- Related AI using other models
  - AI cloud services for "generative AI adjacent" workloads
  - AI prebuilt into Oracle products

Examples are provided in the following sections.

Cloud Infrastructure and Tooling for Generative AI Model Deployment and Training

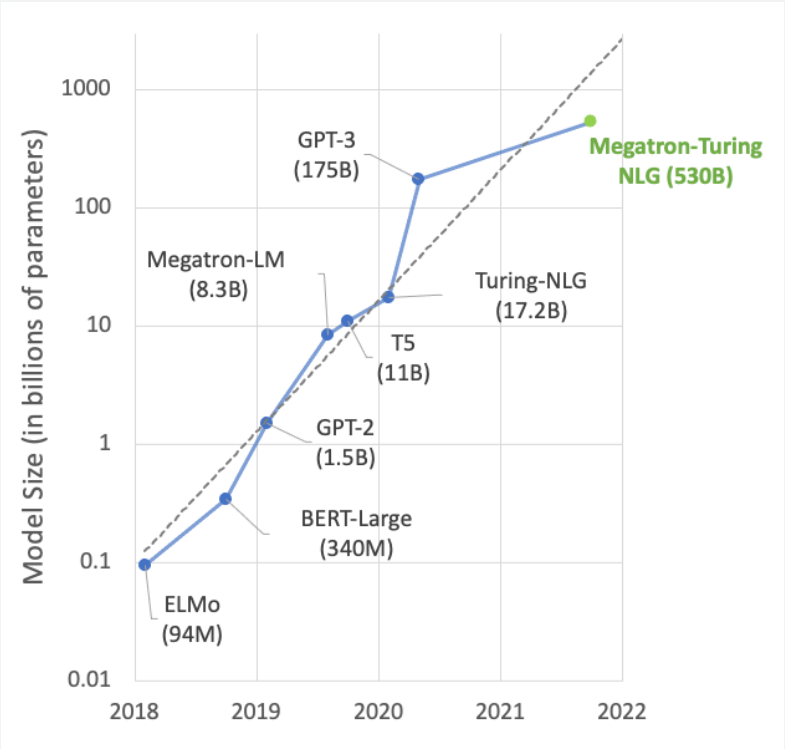
Oracle Cloud Infrastructure (OCI) is designed for all AI and ML workloads, and is well suited for generative AI workloads.

OCI's AI infrastructure rivals or betters the performance of dedicated, custom on-premises compute clusters while providing the elasticity and consumption-based costs of the cloud. Because of our unique Gen 2 Cloud architecture, we were the first to fine-tune GPT3-sized AI models with a single NVIDIA A100 GPU.

AI models continue to grow, and training them is becoming more expensive and more difficult. Many models no longer fit into the memory of a single GPU. For example, LLMs require hundreds to tens of thousands of GPUs. Dedicated high-performance networks are also required to interconnect these GPUs. Assembling and configuring cloud infrastructure for an LLM requires deep technical expertise from your own organization and from your infrastructure provider. Oracle has found a way to simplify that.

OCI Supercluster provides ultra-low latency cluster networking, HPC storage, and bare metal instances powered by NVIDIA GPUs for training LLMs, including conversational AI and diffusion models. Each supercluster can scale up to 32,768 NVIDIA A100 GPUs with 1600 Gb/sec of cluster networking bandwidth.

*AI services offered by Oracle Cloud



Large language models are growing in size and complexity. Graph source: NVIDIA.

Unlike other cloud service providers, Oracle Cloud offers bare metal GPU servers with performance, isolation, and control by using dedicated compute instances powered by NVIDIA H100, A100, and A10 Tensor Core GPUs, high core counts, large amounts of memory, and high bandwidth.

OCI provides 16x more internode bandwidth than Google Cloud Platform (GCP), 4x more internode bandwidth than Amazon Web Services (AWS), and better price-performance than all of the cloud megascalers.

Oracle has the highest performance, lowest cost cloud infrastructure for generative AI model training and deployment.

OCI is ideal for generative AI models because of its advanced networking, GPU shapes, clustering, and software options.

- OCI cluster networking: Training generative AI applications using LLMs requires powerful clusters of computing infrastructure that can process vast amounts of data. OCI provides RDMA with dedicated RDMA over Converged Ethernet (RoCE) v2 cluster networks with latencies as low as 1.5 microseconds and 1600 Gb/sec of internode bandwidth. These superclusters can contain up to 4,096 nodes (or 32,768 GPUs) connected by a high-performance RDMA network fabric.
- Open source libraries and frameworks: AI systems rely on complex ML models to generate new data. ML frameworks such as TensorFlow, PyTorch, and

CSP	Local SSD	Cluster Network Bandwidth (GBPS)	# VCPU	CPU Memory Per Node	Instance $/HR	Price/ Performance (Less Is Better)
OCI	27.2 TB	1600	256	2048 GB	$32.00	14.6
AWS	8 TiB	400	96	1152 GB	$40.97	74.8
Azure	6.4 TiB	1600	96	1900 GiB	$32.77	15.0
GCP	3 TB	100	96	1360 GB	$31.44	229.5

Keras provide a set of tools and APIs to build and train models, and they also provide a variety of prebuilt models for image, text, and music generation. Pandas is a popular ML library, typically used when preparing data for later use in ML frameworks.

OCI supports pandas, Dask, NumPy, Plotly, Matplotlib, TensorFlow, Keras, PyTorch and others, and offers a reference architecture for developers who want to set up an open source ML and AI environment. OCI Data Science users have access to natural language processing for GPU/CPU conda environments, including the Hugging Face transformer library, which is a leading open-access framework for LLMs. Data scientists can tune and deploy transformer models from that library, which includes the recent BLOOM and BLOOMZ models (including many natural languages and also some programming languages).

- Multiple GPU options: OCI provides bare metal and VM instances powered by NVIDIA GPUs for a variety of use cases, from mainstream graphics and videos to the most demanding AI training and HPC workloads.
  - NVIDIA H100 GPU, the latest generation GPU for LLM training, will be available on Oracle Cloud, and offers NVIDIA AI Enterprise, which includes essential processing engines for each step of the AI workflow, from data processing and AI model training to simulation and large-scale deployment.
  - NVIDIA A100 Tensor Core GPU is a high performance GPU that is used for several generative AI use cases because, among other things, it specializes in detecting and classifying JPEG images and segmenting them into their component parts.
  - NVIDIA A10 Tensor Core GPU is a versatile processor for graphics and video processing as well as AI inferencing. When combined with NVIDIA RTX Virtual Workstation (vWS) software, A10 is ideal for running professional visualization applications.
- OCI storage services deliver high-performance and low-cost cloud storage, including local, block, file, object, and archive storage. Users can deploy cluster file systems such as WEKA, BeeGFS, Lustre, Gluster, and IBM Spectrum Scale.
- In addition to the bare metal GPU servers mentioned earlier, OCI also offers virtual GPU instances that deliver secure and elastic compute capacity in the cloud for workloads ranging from small development projects to large-scale, global applications such as LLMs.
- Oracle Interconnect for Microsoft Azure is a joint project between Oracle and Microsoft that enables a multicloud AI approach. This low-latency, private connection between two leading cloud providers allows OCI-based AI projects to seamlessly and cost

effectively interact with Azure services like Azure OpenAI. Interconnect pricing is port-based, and there are no additional charges for bandwidth consumed.

Generative AI Proof Points

Because Oracle can run complex ML models more economically than AWS or GPC, Oracle Cloud Infrastructure has been chosen by many organizations and vendors for their generative AI projects, including the following ones:

- Adept AI Labs has deployed thousands of NVIDIA A100 GPUs in superclusters on OCI. Adept is building a new breed of digital assistant aimed at being useful for business. They describe it as "a universal AI collaborator for every knowledge worker."
- SoundHound has developed an independent voice AI platform on OCI that combines best-in-class voice AI with third-party generative AI models like ChatGPT. This allows businesses across industries to integrate conversational voice assistants into their products and services. SoundHound's voice AI platform has the ability to understand the complexity of human speech so that conversational experiences are more natural and intuitive.
- MosaicML is a generative AI foundry that provides an LLM development environment for AI developers. They deploy on OCI.
- Character.ai is hosted on OCI. It's a neural language model chatbot web application that can generate human-like text responses and participate in the contextual conversation.
- Lawrence J. Ellison Institute for Transformative Medicine of USC scientists have trained a neural network on OCI to spot different types of breast cancer on a small data set of less than 1,000 images. Instead of educating the AI system to distinguish between groups of samples, the researchers taught the network to recognize the visual "tissue fingerprint" of tumors so that it could work on much larger, unannotated data sets.

Oracle is the cloud of choice for many generative AI projects.

AI Cloud Services for Generative AI Adjacent Workloads

Oracle offers a set of Cloud AI services with models that can be custom trained with an organization's own data to improve model quality, making it easier for developers to adopt and use AI technology. These pretrained models use a variety of ML and AI models, though not LLMs. We refer to them here as "generative AI adjacent" in that they can be used in conjunction with generative AI projects.

- Oracle Digital Assistant is an AI service that offers prebuilt skills and templates to create conversational experiences for business applications and customers through text, chat, and voice interfaces.
- OCI Language makes it possible to perform sophisticated text analysis at scale. With pretrained models built in, developers don't need ML expertise to build sentiment analysis, key phrase extraction, text classification, translation capabilities, and more into their applications.
- OCI Speech uses automatic speech recognition (ASR) to convert speech to text. Built on the same AI models used for Oracle Digital Assistant, developers can use Oracle's acoustic and language models to provide highly accurate transcription for audio or video files across many languages.
- OCI Vision applies computer vision to analyze image-based content. Developers can easily integrate pretrained models into their applications with APIs or custom-train models to meet their specific use cases. Vision provides image analysis for object and scene-based images. Use cases include detecting visual anomalies in manufacturing and tagging items in images to count products or shipments.
- OCI Document Understanding is an AI service for extracting text, tables, and other key data from document files. Built on OCI Vision, it helps developers identify and locate objects, extract text, and identify tables, document types, and key-value pairs from business documents. Prebuilt models are available, and developers don't need ML expertise to work with them.
- OCI Data Labeling helps users build labeled datasets to train AI models for labeling of documents, images, and text. The labeled data sets can be exported and used for model development across many of Oracle's AI and data science services, including OCI Vision and OCI Data Science, for a consistent model-building experience.

These traditional Oracle AI services can be used in conjunction with generative AI projects.

AI Prebuilt Into Oracle Products

Oracle integrates AI-generated output into its existing applications, tools, and platforms. Over the course of 20 years, we've innovated AI-driven automation in our analytics, cloud platform offerings, cloud business applications, and database portfolio. As of this writing, these AI capabilities are not yet based on generative AI models.

- Oracle Digital Assistant uses AI to generate natural-language responses to chat inquiries and automate routine tasks. Examples include the generation of answer intents, auto-complete text, quick replies, and more. It includes a code generation feature called SQL Dialog that translates a user's natural language input into SQL queries, sends the queries to a back-end source, and displays the response.

- Oracle Analytics Cloud includes built-in integration with OCI AI services for use directly in analytics projects. For example, Oracle Analytics Cloud with OCI Vision enables object detection, image classification, and text detection from within Oracle Analytics Cloud. You perform this AI analysis by invoking the Vision service from a data flow in Oracle Analytics Cloud. Oracle Analytics Cloud also has integration with OCI Language, which provides sophisticated text analysis for analytics use cases such as customer support feedback.
- Oracle Content Management is Oracle's enterprise digital asset management product. It includes a smart content feature that uses AI to automatically tag documents, images, and video so content creators, authors, and consumers can discover the content when they need it. It includes the following features:
  - Smart search, which analyzes images for color and content and automatically creates tags for them.
  - Smart video transcription, which processes English language videos by using computer vision models and automatically creates transcripts for them.
  - Smart authoring, which recommends images during the authoring process. Users don't need to tag or search for images; images are recommended based on the intent expressed in the current article. Natural language processing (NLP) models derive the main intent from an article.
- Oracle Autonomous Database uses AI to automatically optimize performance, security, resource efficiency, availability, and other aspects of the database. These processes happen without the need for human intervention, so routine tasks traditionally performed by database administrators, like creating, running, and maintaining backup scripts, installing updates, managing security, and optimizing performance are done by the Autonomous Database.

Oracle has AI-driven automation in our analytics, cloud platform, cloud business applications, and database portfolio.

ORACLE
NETSUITE