ORACLE

Generative AI at Oracle

Frequently Asked Questions for Internal Sellers

June 2024, Version 6

ORACLE

Disclaimer

This document in any form, software, or printed matter, contains proprietary information that is the exclusive property of Oracle. Your access to and use of this confidential material is subject to the terms and conditions of your Oracle software license and service agreement, which has been executed and with which you agree to comply. This document and information contained herein may not be disclosed, copied, reproduced or distributed to anyone outside Oracle without prior written consent of Oracle. This document is not part of your license agreement nor can it be incorporated into any contractual agreement with Oracle or its subsidiaries or affiliates.

This document is for informational purposes only and is intended solely to assist you in planning for the implementation and upgrade of the product features described. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described in this document remains at the sole discretion of Oracle. Due to the nature of the product architecture, it may not be possible to safely include all features described in this document without risking significant destabilization of the code.

ORACLE

Table of Contents

ORACLE

Introduction

AI is fundamentally changing the way we interact with the world. Generative AI is a new technology that consists of large language models (LLMs) that are trained on enormous amounts of data and can generate new content, such as text, images, code, video, and audio. Today, generative AI is creating impact across industries and will lead to a new era of business innovation. But choosing the right approach can be a challenge.

Oracle has always focused on enterprise success. This pursuit has driven all of Oracle's technology decision making, product innovation, and acquisitions—which is why today, Oracle can offer an enterprise-focused, end-to-end platform for generative AI built upon our high-performance AI infrastructure and comprehensive portfolio of cloud applications.

The OCI Generative AI service enables customers to accelerate their AI initiatives and drive greater business value. The OCI Generative AI service is a fully managed service that integrates LLMs from Cohere and Meta to address a wide range of business use cases. It will form the basis for generative AI capabilities embedded across Oracle's suite of SaaS applications, including Fusion ERP, HCM, and CX, as well as industry applications, Oracle Health, and NetSuite.

In contrast to other approaches, the OCI Generative AI service provides dedicated AI clusters, enabling customers to run their workloads on dedicated infrastructure—allowing them to control their specific cost and throughput requirements. By virtue of its low-cost, high-speed AI infrastructure, Oracle is the most cost-effective way to consume Cohere and Meta models at scale. OCI also provides fine-tuning support for foundational models to meet organizations' specific business requirements.

Oracle AI

The OCI Generative AI service is part of a larger Oracle AI suite of services:

- AI for applications: Work more efficiently and effectively through integrated, fully-functional AI capabilities in Oracle's business applications.
- AI services: Gain access to pre-trained models, including generative AI models that can be fine-tuned with an organization's own data.
- ML services: Help data scientists collaboratively build, manage, and deploy machine learning models with their favorite open-source frameworks or benefit from the speed of in-database machine learning.
- AI infrastructure: With support for up to tens of thousands of GPUs, OCI Compute virtual machines and bare metal instances can power high-performance applications for computer vision, natural language processing (NLP), recommendation systems, and more.

ORACLE

General Information

1. What's the news?

We are pleased to announce the **General Availability** of multiple new features on the OCI Generative AI service:

1. We now offer **Cohere Command R** and **Command R+** for on demand and dedicated hosting. Command R is an instruction-following conversational model that can perform language tasks at a higher quality, more reliably, and with a larger context window than previous Cohere Command models. Customers can use it for enterprise-grade AI workflows and apply retrieval augmented generation (RAG), Tool Use, and custom agents to exand its capabilities. OCI's hosted version of Command R is smaller than any other version available on any cloud provider, allowing us to offer Command R with the best available price-performance while still supporting a 16K context window. Support for fine-tuning for Command R is planned for an upcoming release. Command R+ is Cohere's most powerful model optimized for longer-context tasks, such as multi-step tool use.
2. We now also support the pre-trained **Meta Llama3-70B** model. This model offers improved capabilities of reasoning, code generation, and instruction as compared with the previous Llama 2 model. With this release, customers can access Llama3-70B via on demand or dedicated hosting. Customers can also fine-tune Llama3-70B model with their own data to help enhance the precision of the model with the Low-Rank Adaptation (LoRA) method.
3. We also introduced a new **Chat API** which supports the new Command R and Llama3 models, accessible via OCI SDK/API or Console Playground. The new Chat API provides an easy integration into the new models by matching the format of the partner model APIs. It also supports the Tool Use for function calling or building complex agents.
4. We are now live in the **Germany Central (FRA) region.**

As the OCI Generative AI service is committed to offering newer and more capable models to our customers, we will be updating our model retirement policy. This policy document will include model retirement dates so that you have the latest information about model availability and can recommend newer models for upgrade. See Retiring the Models for more information.

OCI Generative AI Agents service with a RAG agent, currently in beta, combines the power of LLMs and enterprise search. The beta release is built on OCI OpenSearch to provide contextualized results that are enhanced with enterprise data. This service enables users to converse with diverse enterprise data sources through natural language without the need for specialist skills. The information retrieved is current—even with dynamic data stores—and the results are provided with references to the original source data.

Upcoming releases will support a wider range of data search and aggregation tools and provide access to Oracle Database 23c with AI Vector Search and MySQL HeatWave with Vector Store. This will enable users to build agents that automate their interactions with Oracle and third-party applications, and direct the agents to take actions based on query outcomes. Oracle will also deliver pre-built agent actions across its suite of SaaS portfolio, including Oracle Fusion Cloud Applications Suite, Oracle NetSuite, and industry applications such as Oracle Cerner.

2. Can I show my customer a demo of the service?

Please check Sales Accelerator for the most up-to-date content and demos. Reach out to the AI Product Management team if additional demos are needed: generative_ai_contact_grp@oracle.com or on Slack at #generative-ai-users.

3. How much does it cost to use the OCI Generative AI service?

The OCI Generative AI service is an OCI-native service that provides two experiences, each with its own pricing model:

On-demand model: For inferencing of pre-trained LLMs, customers will be charged for each inference call based on the context length of each request (input + output in terms of characters).

Dedicated AI clusters model: For fine-tuning of LLMs on dedicated AI clusters, as well as hosting of pre-trained and fine-tuned models on dedicated AI clusters. This refers to an equivalent set of GPU shapes for fine-tuning or hosting models for a specific tenant that managed by the OCI Generative AI service.

Please refer to the OCI pricing rate card here for the latest pricing details.

4. What's the minimum level of investment required to leverage dedicated AI clusters?

The minimum level of investment for a dedicated AI cluster can be calculated by multiplying the lowest unit-hr cost by the minimum commitment for hosting hours.

5. When using a fine-tuning dedicated AI cluster, are additional resources needed for inferencing or is it included?

A separate, hosting dedicated AI cluster is required for inferencing. See the documentation for more information on how to provision and use dedicated AI clusters for fine-tuning and hosting customized models.

6. How will this new service compare to other OCI AI services?

The OCI Generative AI service brings LLMs from Cohere and Meta to customers to enable them to create text-based content. This technology is especially suited for use cases such as semantic search, information retrieval for conversational AI, and text generation and summarization. Our other AI services focus on other use cases, such as NLP, chatbots, anomaly detection, and computer vision, as well as tools for data scientists. Learn more here.

7. Will Oracle provide support to help customers refine and tailor these models using their own data?

Oracle will collaborate with customers to support their business objectives with the OCI Generative AI service, including by enabling customers to fine-tune models with their own data. You can reach out to the AI Product Management team at generative_ai_contact_grp@oracle.com or on Slack at #generative-ai-users with specific requirements.

8. What models support fine tuning

Llama3 and the Cohere models (with the exception of Command R at the moment) support fine tuning. It is on the roadmap for Command R. Please reach out to the PM team at generative_ai_contact_grp@oracle.com if your customer has specific needs.

9. Are there plans to offer tools/consulting services to help customers with prompt engineering or templating?

This is on the roadmap. However, we aren't disclosing any specifics at this time, such as the scope of services that may be available or any timelines. If an organization has specific requirements, please reach out to the AI Product Management team at generative_ai_contact_grp@oracle.com or on Slack at #generative-ai-users.

10. To whom can I reach out if I have more questions for a potential customer opportunity?

If you have an immediate business need, contact AI Product Management at generative_ai_contact_grp@oracle.com or on Slack at #generative-ai-users.

ORACLE

Service Capability Questions

1. What capabilities and models will the OCI Generative AI service support?

We provide LLM models for text generation, summarization, and embedding. Text generation is supported on Cohere and Llama3 models. Fine-tuning is supported on Llama3 and Cohere models with the exception of Command R (targeting July '24) and Command R+. Other models are under consideration.

1. What if customers have concerns about hallucinations?

While hallucinations will likely always arise due to the nature of today's LLMs, there are ways to potentially mitigate their risks, including better prompting techniques (e.g., with more specific prompts), retrieval-augmented generation (RAG), and fine-tuning. Please reach out to the AI Product Management team at generative_ai_contact_grp@oracle.com or on Slack at #generative-ai-users with questions or to seek a resource to engage with you or your account.

2. Can customers use the OCI Generative AI service to generate answers about their collection of ever-changing documents?

Yes. The OCI Generative AI service provides customers with the LLMs to transform their current search experience into a Retrieval Augmented Generation (RAG) solution. The OCI Generative AI Agents service with RAG is currently in beta, running OCI OpenSearch.

3. What is an Agent/RAG Agent?

An agent is an autonomous program that operates on behalf of a user or system without continuous human intervention. It is capable of acting automatically, collaborating with other agents or systems, and learning from historical and changing data. Essentially, the agent orchestrates interactions between the LLM, other tools, data, and the user. OCI Generative AI Agents combines the power of LLMs and the potential of RAG with a customer's data to let users query diverse enterprise knowledge bases using natural language.

4. Does enterprise data need to be vectorized via an embedding model before being used in a RAG solution?

Each customer and use case will have different requirements. Work with your cloud engineering team to determine the best solution for your customer. Reach out to the PM team via generative_ai_contact_grp@oracle.com if additional help is needed.

5. Does OCI provide throughput SLAs for the Generative AI service?

Please refer to Oracle PaaS and IaaS Public Cloud Services Pillar Document for SLA details.

6. Are there any performance benchmarks available OCI Generative AI?

There are some published benchmarks, but it is important to keep in mind that there are many different factors that can impact performance and they will vary from deployment to deployment.

7. How often are the models updated to a new version?

New models will be integrated in the OCI Generative AI service after performing internal evaluations. Model availability and timing will vary depending on the nature of the version update and the need for a new or updated SKU.

8. Will our customers get access to the new Cohere and Meta models as they are released?

Yes. New Cohere and Meta models will be integrated in the OCI Generative AI service following our internal evaluations. We will release our retention policy as part of the service documentation.

9. Will Generative AI work with Oracle Database?

Oracle Database can make use of generative AI through REST endpoints. Users will be able to make a call from the database to the model endpoint in a similar way to how they can make a call to a deployed model in OCI Data Science, Language, or Vision services today. In addition, we will inject these LLM capabilities into our database portfolio, by enabling Oracle Database users to issue natural language statements or generating optimized SQL and PL/SQL code. Oracle Database 23c can also be used as a vector database to store and search LLM vector embeddings.

10. How can organizations customize the Generative AI models with their own data?

Customers can fine-tune most models with their own data via the OCI console and APIs.

11. Can these models be applied to the healthcare, manufacturing, financial services, and other industries?

Absolutely. We recommend that customers first check if the pre-built models meet their needs. In our experience, the pre-built models, with a bit of prompt tuning, have been able to address certain specific scenarios across a variety of industries. That said, if an organization's needs are more demanding, customers can fine-tune with their own data.

12. Does Oracle provide industry- or vertical- specific foundation models?

Not as part of the initial GA release of the Generative AI service. However, it's part of our future plans. If you have any comments, please email generative_ai_contact_grp@oracle.com.

13. How is data secured in the Generative AI service?

Refer to https://docs.oracle.com/en-us/iaas/Content/generative-ai/data-handling.htm for details.

14. How well does Cohere perform in languages other than English?

The Cohere Command generative AI models were originally trained to respond primarily in English. While they can understand other languages to a degree, performance depends on the model and specific circumstances, and we encourage customers to test their scenarios. Command R and R+ have been updated to perform well in the following languages: English, French, Spanish, Italian, German, Brazilian Portuguese, Japanese, Korean, Simplified Chinese, and Arabic If your customer is interested in a specific language and would like to try it out, please contact us at generative_ai_contact_grp@oracle.com.

Over 100 languages are supported in the Cohere Embed Multilingual model. Embed models can be used for a variety of scenarios including search, text classification, content aggregation, and recommendations. You can see objective benchmarks for the Cohere Embed models at https://txt.cohere.com/multilingual/. The list of supported languages can be found at https://docs.cohere.com/docs/supported-languages.

15. What data are pre-built Cohere models trained on?

The data is proprietary to Cohere, and we cannot offer any specifics.

16. Is my customer's data sent to Cohere or Meta?

No. Data sent to the OCI Generative AI service is not sent to Cohere or Meta. In addition, thanks to our dedicated AI clusters, customers can choose to run their work streams on hardware dedicated only to them.

17. I heard that the new Generative AI services will leverage Supercluster technology. What is Supercluster technology?

OCI's Supercluster includes OCI Compute Bare Metal, an ultra-low latency RoCE (RDMA over Converged Ethernet) cluster based on NVIDIA networking, and a choice of HPC storage. It has been deployed and validated by NVIDIA to support thousands of OCI Compute Bare Metal instances that can efficiently process massively parallel applications. OCI Supercluster

networking can now scale up to 4,096 OCI Compute Bare Metal instances with 32,768 NVIDIA A100 GPUs. For more information, visit https://www.oracle.com/ai-infrastructure/#rc30p0.

18. Will customers be able to use their own OCI GPUs with the Cohere models?

No. Customers will be able to reserve "dedicated AI clusters" so that they can run their workloads in hardware dedicated only to them, allowing them to control the cost / throughput tradeoff and optimize it for their specific business requirements.

19. Can customers use the Generative AI service with Oracle Digital Assistant (ODA) to create a more intelligent chatbot?

Absolutely. Through ODA's new LLM block capability, customers can integrate OCI Generative AI with their current Oracle Digital Assistant.

20. When will the Generative AI service support Cohere's Classify and Rerank models?

This is on our roadmap, however, we aren't providing specific dates at this time. Until then, classification can be performed with the Command model—with a bit of fine-tuning. Similarly, since the Embed model is exposed, re-ranking is trivial (e.g., by doing cosine similarity, which measures the similarity between two vectors of an inner product space).

21. Will the OCI Generative AI service offer open-source models and support Meta's Llama models like Microsoft does?

The OCI Generative AI service supports Meta's Llama 3 in addition to Cohere's models. Also, OCI Data Science has the feature: AI Quick Actions—a "no-code" feature that enables access to a wide range of open source LLMs, including Mistral AI. This feature provides wider access to the best of what the open-source community offers.

We are also evaluating additional patterns for OCI Generative AI, including a "bring your own model" capability in the future. You can provide input on this topic at generative_ai_contact_grp@oracle.com.

22. What are AI Quick Actions?

AI Quick Actions provide customers with a streamlined, code-free, and efficient environment for working with LLMs, offering a seamless experience from fine-tuning to deployment.

23. Where can I see and use the AI Quick Actions?

The AI Quick Actions catalog is a collection of LLM use cases that can be invoked with a click of a button, all in an easy-to-use user interface inside the Data Science notebook experience.

24. What kind of LLMs are available on AI Quick Actions?

AI Quick actions supports models such as falcon-7b, phi-2, codellama, mistral, in addition to most other LLM models with model artifacts saved in Object Storage by customers (Bring Your Own Model from Object Storage). The current list of Service Curated Models and Verified Models (vpn required).

25. Are machine learning models other than LLMs available via AI Quick Actions?

The initial availability only includes LLMs, but we will be evaluating additional models in the future. However, customers are still able to work with any LLM or classical ML models using Python code on the OCI Data Science platform.

26. What is the difference between Gen AI service and AI Quick Actions?

The OCI Generative AI service provides models-as-a-service, where the customer interacts with the service via web requests/responses only (with no infrastructure required). AI Quick Actions are built on top of the OCI Data Science platform, meaning the models will be fine-tuned and deployed as resources in a customer's tenancy, and then the customer can interact

with the models via APIs. There is still some aspect of resource management from the customer side. Billing is also different because the Generative AI service is billed per character, while Data Science platform is billed by the actual compute consumption.

27. How are OCI Data Science AI Quick Actions billed?

AI Quick Actions utilize OCI Data Science resources; therefore, the billing is the same as with any other Data Science resource – per compute consumption.

28. What is Oracle's strategy for multiple modality (text as well as images, video and audio) models?

Multiple modality generative AI models are under consideration; we will provide additional details at a later date.

29. Microsoft is offering Copilot for code generation. Do you plan to offer a similar capability?

Cohere's models and Meta's Llama3 are capable of generating code in multiple programming languages. Contact Generative AI Product Management for additional details at generative_ai_contact_grp@oracle.com.

30. Can customers embed Generative AI in their APEX applications?

Yes. Generative AI models will be consumable as REST APIs from APEX applications.

Roadmap

1. Which Cloud Regions will support the OCI Generative AI service? When will it be supported in X region?

The OCI Generative AI service is available in the following regions: Chicago, Illinois (ORD); and Frankfurt, Germany (FRA). UK South (London, LHR) is targeted for end of June '24. Other regions will follow. Please see the internal facing roadmap.

2. Does Oracle plan to develop its own LLMs in the future?

We have no updates to share at this time.

3. Does Oracle plan to offer more LLMs in the future?

Additional LLM models are under consideration. We will provide an update at a later date.

Oracle Generative AI service vs competitors

1. What distinguishes OCI Generative AI service from other generative AI offerings?

These statements from industry researchers best encapsulate Oracle's advantages:

IDC:

"With today's news, Oracle is bringing generative AI to customer workloads and their data—not asking customers to move their data to a separate vector database," said Ritu Jyoti, Group Vice President, Worldwide Artificial Intelligence and Automation Research Practice and Global AI Research Lead, IDC. "With a common architecture for generative AI that is being integrated across the Oracle ecosystem from its Autonomous Database to Fusion SaaS applications, Oracle is bringing generative AI to where exabytes of customer data already reside, both in

cloud data centers and on-premises environments. This greatly simplifies the process for organizations to deploy generative AI with their existing business operations."

"Organizations everywhere have struggled with how to deliver generative AI successfully—and Oracle just provided the answer," said David Schubmehl, Research Vice President, Conversational Artificial Intelligence and Intelligent Knowledge Discovery, IDC "With its new Generative AI service, which supports fine tuning so businesses can customize LLMs to their own internal operations, and a comprehensive AI strategy that spans all layers of its tech stack, Oracle has demonstrated that it's focused on solving real world business problems."

Wikibon:

"Oracle is taking a full stack approach to enterprise generative AI," said Dave Vellante, Chief Research Officer, Wikibon. "Oracle's value starts at the top of the stack, not in silicon. By offering integrated generative AI across its Fusion SaaS applications, Oracle directly connects to customer business value. These apps are supported by autonomous databases with vector embeddings and run on high-performance infrastructure across OCI or on-prem with Dedicated Region. Together these offerings comprise a highly differentiated enterprise AI strategy, covering everything from out-of-the-box RAG to a broad range of fine-tuned models and AI infused throughout an integrated stack. Our research shows that 2023 was the year of AI experimentation. With capabilities such as this, our expectation is that 2024 will be the year of showing ROI in AI."

The new OCI Generative AI service from Oracle is seamlessly integrated up and down the entire stack from the hardware through the applications—making the massive "Rube Goldberg"-like effort of integrating generative AI into mission-critical workloads a thing of the past," said Marc Staimer, Senior Analyst, Wikibon. "Suppose you want to bake a cake; with other cloud providers you go to a grocery store, buy all the ingredients and they provide the mixing bowl, pan, and oven to bake it in. But they don't even provide the recipe. In contrast, OCI provides you the entire gen AI cake—already baked."

Additional Oracle Commentary:

The OCI Generative AI service is OCI's service for training and inferencing generative AI models. We have partnered with Cohere, one of the pioneers in generative AI, and offer Meta Llama 3 to bring to OCI customers state-of-the-art models that they can integrate into applications.

Given Oracle's strength in SaaS apps for industries at scale, we shaped our Generative AI service to work well for large-scale enterprise customers by providing:

A broad range of models for enterprise: State-of-the-art models designed for enterprises that customers can integrate into their applications, pre-built or fine-tuned.

Predictable performance and pricing: In contrast with Azure Open AI, we provide *dedicated AI clusters,* so that customers can run their workloads in hardware dedicated only to them, allowing them to control the cost / throughput tradeoff. Azure does not offer this and AWS Bedrock is much more expensive.

Security and privacy: We understand that security and privacy are critical for enterprise customers, and we have decades of experience earning customer trust. Customer data is not shared with Cohere or Meta and is not seen by other customers. In addition, custom models trained on customer data can only be used by that customer.

2. How does Oracle's partnership with Cohere compare to Google's investment in Runway?

Runway's technology is targeting image and video generation with a particular focus on artistic creation. Oracle's strategy is to focus on real-world enterprise use cases. Text-to-text use cases are affecting every single enterprise and, to date, have been identified by organizations as their top priorities. Although images and video generation could have a big impact for certain

industries or verticals (e.g., gaming, entertainment, and retail), most enterprise customers are not leading with those use cases as a high priority or don't have the relevant datasets to fine-tune for their objectives.

3. How is the OCI Generative AI service different from Azure OpenAI?

The OCI Generative AI service is OCI's service for training and inferencing generative AI models.

In contrast with Azure Open AI, we provide dedicated AI clusters, so that customers can run their workloads in hardware dedicated only to them, allowing them to control the cost/throughput tradeoff and optimize it for their specific business requirements. We also provide fine-tuning support. Azure no longer supports fine-tuning for new customers.

In addition, because Generative AI model training and inferencing will run on OCI's high-performance, low-latency Supercluster RDMA technology, we can provide a very competitive offering.

4. Since Cohere is available on AWS and Azure, what makes our partnership with Cohere different?

At AWS, Cohere is an option in a buffet of vendors and partners. In contrast, Oracle is focused on a strategic partnership with Cohere where we can collaborate with the company on their roadmap to support our global customers' most pressing AI business needs, and to work together to develop new models for Oracle's customers. In addition, thanks to our low-cost, high-speed AI infrastructure, Oracle is the most cost-effective way for customers to consume Cohere models at scale.

While there is no exclusivity with Cohere's current models, any models we co-develop with Cohere will not be available on other cloud platforms.

5. Where can I find more information on the competitive differentiators for the OCI Generative AI service?

Oracle's overall AI strategy is our main competitive advantage – focus on enterprise requirements, embedded AI capabilities across the full stack, and emphasis on data security and management. For additional competitive information on the service, visit OCI Generative AI Competitive Info and OCI Generative AI Objection Handling

Selling the OCI Generative AI service

1. When can I sell the OCI Generative AI service?

OCI Generative AI services have been available since the initial release January 25, 2024. If you have an immediate sales opportunity and need support, please contact Generative AI Product Management.

2. Is there any expectation that GenAI embedded in SaaS solutions will lead to price increases? Will these features be offered as a separate SKU?

We have nothing to announce related to SaaS application pricing or SKUs at this time. Please contact SaaS Product Management for questions pertaining to Oracle Fusion Cloud SaaS applications.

Other Questions

1. Does Oracle own the models? If Cohere's business changes tomorrow and we're in production with OCI Generative AI service, what will happen?

We do not own the models developed by Cohere. However, Oracle has a right to host the models in OCI subject to a written agreement. In the unlikely event of a significant change to Cohere that limits its model availability, Oracle would have the right to continue using the models for a reasonable period of time until an alternative can be found.

Glossary of common terms

Agent: In the context of generative AI, an agent is an AI system that can take actions to achieve specific goals or assist users with tasks, often through natural language interactions.

API (Application Programming Interface): A set of protocols and tools that allows developers to integrate AI capabilities into their software applications without building the models from scratch.

Embeddings: Dense vector representations of data (e.g., words, images) in a high-dimensional space, capturing semantic relationships and enabling ML models to understand and reason about the data.

Few-Shot Learning: A type of ML where models can learn to perform new tasks from a very small number of examples, often leveraging knowledge transfer from pre-training on large datasets.

Fine-tuning: The process of further training a pre-trained AI model on a smaller, task-specific dataset to help adapt it for a particular application or domain.

Foundation Model: A large, pre-trained AI model that serves as a base for creating specialized models through fine-tuning, enabling faster and more efficient development of AI applications across various domains.

Generative AI: AI systems that can create new content, such as text, images, or audio, as well as summarize and analyze existing content, based on learned patterns and rules from training data.

GPU (Graphics Processing Unit): A specialized hardware component designed for parallel processing, which is particularly well-suited for training and running deep learning models, including generative AI.

Inference: The process of using a trained AI model to make predictions or generate outputs based on new, unseen input data.

LangChain: An open-source library that provides a standard interface for combining LLMs with other components to build more capable and flexible language models and applications.

ML (Machine Learning): A subset of AI that involves training models to learn patterns and relationships from data, enabling them to make predictions or decisions without being explicitly programmed for each specific task.

MLOps (Machine Learning Operations): The discipline of deploying, monitoring, and maintaining ML models in production environments efficiently and reliably, often using automation and DevOps practices.

Multi-turn: The ability of conversational AI systems to maintain coherent, contextually relevant dialog across multiple user inputs and system responses.

Prompt Engineering: The practice of designing and optimizing input prompts to guide AI models towards generating desired outputs more effectively and consistently.

Prompt: The input text or query provided to a generative AI model, which guides the model to generate a relevant output based on the context and instructions given.

Reranker: A secondary ML model that refines and reorders the outputs of a primary model, such as a search engine or recommendation system, to improve relevance to the user's query or preferences.

Retrieval-Augmented Generation (RAG): A technique to help enhances LLM output by incorporating targeted, up-to-date information from external sources, enabling more contextually relevant and accurate responses.

Scalability: The ability of an AI system to maintain or improve performance as the volume of data and number of users grow, often through distributed computing and efficient resource management.

Transformers: A neural network architecture that uses attention mechanisms to process sequential data, powering state-of-the-art language models like BERT and GPT.

Vector Databases: Databases designed to efficiently store and query high-dimensional vector representations of data, enabling fast similarity searches and powering applications like recommendation systems and semantic search.