# Model comparison

­­

Zolder AI Agent

Contents

[Model comparison 1](#_Toc174033835)

[Introduction 2](#_Toc1933400117)

[Cloud-Based Models: 2](#_Toc220304084)

[Price: 4](#_Toc906607369)

[Locally Hosted Models: 4](#_Toc1290011475)

[Price: 5](#_Toc545764765)

[Conclusion 6](#_Toc2056259086)

# Introduction

This document is intended to compare cloud-based models and locally hosted models. When comparing between these models, several factors should be taken into consideration such as: cost, context, security, and ease of use. Below is a table for in-depth analysis for the mentioned aspects:

## Cloud-Based Models:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Amazon Web Services (AWS)** | **Microsoft Azure**  | **Google Cloud Platform (GCP)** |
| **Context** | Widely used by businesses of all sizes for hosting applications, databases, and storage in the cloud. | Used by businesses and governments for hybrid cloud solutions, integrating cloud services with existing local infrastructure. | Popular among businesses focused on data analytics and machine learning, as well as large-scale web applications. |
| **Services** | Offers Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) solutions. Examples include EC2 (virtual servers), S3 (storage), and RDS (managed databases). | Includes a wide range of services such as Azure Virtual Machines (VMs), Azure App Service, and Azure SQL Database. | GCP provides Google Compute Engine (VMs), Kubernetes Engine (container management), and BigQuery (data analytics). |
| **Use Case** | Startups like Airbnb and large enterprises like Netflix rely on AWS to dynamically scale their infrastructure and handle large amounts of data without investing in local hardware. | Healthcare organizations use Azure for secure data storage and compliance with regulations like HIPAA, with built-in services for data protection. | Spotify uses GCP for music streaming services and to manage millions of real-time data streams and user requests. |
| **Security**  | AWS provides various security features, including encryption, compliance certifications, and identity and access management (IAM). | Offers multi-layered security features and compliance support, including advanced threat protection and security monitoring. | Google Cloud employs end-to-end encryption, threat detection, and compliance support for various industries. |

### Price:

Cloud-based platforms are generally charged on a pay-as-you-go basis, offering flexibility but potentially higher long-term costs, here are the prices on every model mentioned above:

**Google Cloud AI - $0.0113 per vCPU/hour (n1-standard-1)**

**AWS SageMaker - $0.0464 per hour (ml.t3.medium instance)**

**Microsoft Azure AI - $0.008/hour for B1s virtual machine**

**OpenAI API - GPT-4 8k: $0.03/1k tokens (prompt), $0.06/1k tokens (completion). GPT-4 32k: $0.06/1k tokens (prompt), $0.12/1k tokens.**

(note all the prices are excluded from storage + data/training/data transfer costs)

## Locally Hosted Models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **VMware vSphere**  | **Hyper-V (Microsoft)**  | **OpenStack** | **Dell EMC PowerEdge Servers** |
| **Context** | Used for building private clouds in on-premise data centers. | Used for virtualization of data centers and creating private cloud environments. | An open-source platform for creating private and public clouds. | Popular for enterprises requiring local data centers or hybrid environments. |
| **Services** | Provides server virtualization, enabling organizations to host multiple virtual machines on a single physical server. | Allows users to run multiple operating systems as virtual machines (VMs) on Windows servers. | Provides tools for managing compute, storage, and networking resources in a locally hosted environment, supporting on-premise cloud deployment. | High-performance, scalable servers for handling workloads like data analytics, virtualization, and critical applications. |
| **Use Case** | Banks and government agencies use VMware vSphere to virtualize and manage their IT infrastructure in-house, ensuring full control over data and security. | Enterprise environments, including large corporations and research institutions, use Hyper-V for cost-efficient, in-house server management. | CERN (European Organization for Nuclear Research) uses OpenStack to process vast amounts of research data on their private cloud infrastructure. | Retailers and manufacturing companies use Dell PowerEdge servers to store sensitive customer or production data locally while maintaining operational control. |
| **Security**  | Offers strong virtualization security with role-based access, network segmentation, and advanced encryption options. | Built-in security tools for data protection, such as Secure Boot, Shielded VMs, and regular Windows security updates. | Offers security features like encryption, firewall management, and identity management through Keystone. | Includes advanced security features like data encryption, secure boot, and built-in threat detection. |

### Price:

Hosting models locally requires upfront investment in hardware but may be more economical long-term for heavy usage, below are the prices for some local-hosting options:

**NVIDIA RTX 4090 - ~$1,599 for GPU (workstation-grade PC ~$2,500–$4,000)**

**NVIDIA Jetson AGX Orin - ~$1,599**

**Hugging Face Models - Costs associated with local compute and storage; models are free**

# Use case scenario

In this chapter we will give a real-world use case where a company needs to host a machine learning model:

**A text generation model (similar to GPT-4) with a 30-second average inference time per request. it processes 50,000 requests per month and needs 2 vCPUs and 8 GB RAM. each request takes 30 seconds (25, 000 minutes or 416 hours of compute time in a month)**

## Cloud-based model hosting (e.g. AWS, Azure)

### Amazon SageMaker

* **Instance Type**: ml.m5.large (2 vCPUs, 8 GB RAM)
* **Cost per Hour**: $0.12 per hour for **ml.m5.large**
* **Monthly Compute Time**: 416.67 hours per month

**Monthly Pricing:**

* **Compute Cost**:

416.67 hours×$0.12/hour=$50

* **Additional Costs** (Storage, Data Transfer, etc.):
	+ **Model storage (10 GB)**: $0.10 per GB per month = **$1/month**
	+ **Data transfer**: Assuming 100 MB per request, and 50,000 requests per month = **5 TB**. First 1 GB free, then **$90 per month** for data transfer over 1 GB (AWS rates).

**Total cost: $50(compute) + $1(storage) + $90(data transfer) = $141/month**

### Google cloud AI (Vertex AI)

* **Instance Type: n1-standard-2 (2 vCPUs, 7.5 GB RAM)**
* **Cost per Hour: $0.095 per hour for n1-standard-2**
* **Monthly Compute Time: 416.67 hours per month**

**Monthly Pricing:**

* **Compute Cost:**

**416.67 hours×$0.095/hour=$39.58**

* **Additional Costs (Storage, Data Transfer, etc.):**
	+ **Model storage (10 GB): $0.20 per GB per month = $2/month**
	+ **Data transfer: 5 TB at $0.12 per GB beyond free tier = $614.4/month.**

**Total cost: $39.58(computer) + $2(storage) + $614.4(data transfer) = $655.98/month**

## Locally hosted model (self-hosted on-premise)

1. **Hardware Setup:**
	* **CPU: 2-core Intel Xeon or similar (equivalent to Google Cloud's n1-standard-2).**
	* **Memory: 8 GB of RAM.**
	* **GPU: NVIDIA T4 (or a more affordable alternative with similar performance, such as NVIDIA RTX 3060 or RTX 3070).**
	* **Storage: 1 TB SSD.**
	* **Power Usage: For simplicity, assume the power consumption of the system is 400W (a typical server setup with a GPU).**
2. **Hardware Costs:**
	* **Server: Custom-built workstation or server with 2-core CPU and 8 GB RAM = $1,200.**
	* **GPU: NVIDIA RTX 3070 = $600 (close performance to T4).**
	* **Total Hardware Cost: $1,800.**
3. **Ongoing Costs:**
	* **Energy Consumption: Assume the server uses 400W of power and runs for 416.67 hours per month.**
		+ **Power Usage: 400W × 416.67 hours = 166.67 kWh.**
		+ **Electricity Rate: $0.13 per kWh (average in the U.S.).**
		+ **Total Power Cost: 166.67 kWh × $0.13 = $21.67 per month.**
4. **Maintenance Costs:**
	* **Estimated around $20 per month for hardware upkeep, cooling, and minor repairs.**

**Total Monthly Cost (Locally Hosted):**

* **Ongoing Power Cost: $21.67 per month.**
* **Maintenance: $20 per month.**

**One-Time Hardware Purchase (Amortized):**

**If we amortize the $1,800 hardware cost over 3 years (36 months), that comes to:**

* **Amortized Hardware Cost: $1,800 ÷ 36 = $50 per month.**

**Total Monthly Cost (Locally Hosted):**

**21.67(power)+20(maintenance)+50(hardware)=91.67 USD/month**

# Conclusion

The Cloud-Based models are flexible, pay-as-you-go but can become expensive where is the other side, Locally Hosted are high upfront hardware cost, but economical long-term for heavy use. Cloud-based are typically larger but locally hosted depends on model, smaller context windows unless customized.

Cloud-based models are provider-managed security but has a risk of third-party breaches while locally hosted models have full control over security but requires expertise to maintain. Cloud-based models are easy to setup and highly scalable. Locally hosted models are more challenging to set up, but it’s customizable to the environment.

Cloud-based are good for Startups, rapid scaling, variable workloads, SaaS products. Locally hosted are good for privacy-sensitive industries (e.g., healthcare, finance), high control over compute environments, heavy resource use (e.g., research labs).

Cloud-based models are often the go-to for scalability and ease of deployment, while locally hosted models offer better control and potentially lower costs for continuous, heavy usage.

Cloud-based is the best for small-to-medium use cases or where scalability and flexibility are priorities. Local models is the best for heavy use, data privacy requirements, or long-term cost efficiency where upfront investment in hardware is feasible.