HIPAA-Eligible AI Capabilities by Major Cloud Provider

MaddLogic LLC
Document date: 2026-01-19

Purpose: Provider-neutral reference for HIPAA-eligible AI inference (LLMs) and speech-to-text options across major cloud providers.
Disclaimer: Informational only; confirm BAA scope, regions, and service configuration with official provider documentation and your compliance counsel.
Scope: Google Cloud, Amazon Web Services, Microsoft Azure. Focused on **AI inference (LLMs)** and **Voice to Text** in regulated environments.

Assumption: Covered Entity or Business Associate operating under an active BAA with the cloud provider.

Note: This document is intended as a **provider-neutral reference**. Inclusion, ordering, and examples are not recommendations. HIPAA eligibility is service-, feature-, region-, and configuration-dependent—verify against the provider's HIPAA/BAA documentation for your exact deployment.

Glossary of Technical Requirements

This section translates the technical terms used later in the document into plain business language. More detail is provided below in the pricing and self-hosting sections.

Legal / compliance basics

- **BAA in place** = the provider signs a HIPAA Business Associate Agreement.
- **No training on your data** = prompts, audio, and outputs are not used to improve the vendor's models.

Security / network basics

- **Private endpoints / VPC / VNet** = the AI service is only reachable from your private network, not the public internet.
- **Encryption at rest and in transit** = PHI is protected both while stored and while moving between systems.
- **Customer-managed keys (KMS)** = you control the encryption keys used to secure PHI.

Infrastructure basics

- **GPU VM** = a special server with a graphics processor optimized for AI.
- **vGPU** = a virtual slice of a physical GPU shared among multiple VMs to reduce cost.
- **Model runtime (e.g., vLLM)** = the software that actually runs the AI model.

Token usage basics

- **Token** = a small chunk of text (often 3 to 4 characters in English).
- **Billed tokens** = everything you send plus everything the model returns.
- **Common token drivers** = system instructions, user prompt text, conversation history, retrieved documents, and model output.

Examples (approximate token counts; actual counts vary by model):

Example 1: Short FAQ response

- Input: "Summarize this: Annual wellness visit is covered once per year." ~20 tokens
- Output: 2 sentence summary ~60 tokens
- Total billed: ~80 tokens

Example 2: Long prompt with policy excerpt

- Input: 1,200 word policy excerpt (~1,600 tokens) + 100 word question (~140 tokens) + system prompt (~120 tokens)
- Output: 5 paragraph answer (~500 tokens)
- Total billed: ~2,360 tokens

Executive takeaway: long context and verbose answers drive cost. Shorter prompts, strict output limits, and careful use of history reduce spend.

Operational basics

- **Logging and retention** = audit trails and clear rules about how long prompts and transcripts are stored.
- **Patching and uptime** = who owns security updates and 24/7 availability.

Executive shortcut: managed services handle most of the infrastructure and patching; self-hosting gives maximum control but requires a dedicated operations plan.

Microsoft Azure

AI Inference Options

Azure OpenAI (Managed)

- Hosted GPT models inside Azure.
- Covered under Microsoft HIPAA BAA when configured correctly.
- Prompts and outputs not used for model training.
- Supports private endpoints, VNets, managed identity.
- Lower operational burden than self-hosted inference.

Private Inference on Azure VMs

- GPU backed VMs (AI-optimized servers). NC, ND series.
- Run open models. LLaMA, Mistral, Qwen.
- Full control of model, runtime, logging, retention.
- No third party model access.
- Maximum control over security and compliance controls (at the cost of more operational ownership).

Voice to Text

Azure AI Speech. Speech to Text

- Real time and batch transcription.
- Medical and domain vocabularies available.
- Covered under HIPAA BAA.
- Supports private networking.
- Audio and transcripts not used for training.

Summary

- Strong integration with Azure identity and networking patterns (e.g., Managed Identity, Private Link / private endpoints).
- Clear managed vs. self-hosted paths: Azure OpenAI for managed inference; GPU VMs for private inference.
- Common considerations: service access approvals/quotas and per-model regional availability.

Amazon Web Services (AWS)

AI Inference Options

Amazon Bedrock (Managed)

- Access to multiple foundation models.
- HIPAA eligible models must be explicitly scoped into BAA.
- Prompts not used for training by default.
- Requires per-model, per-region, and per-feature eligibility review and configuration validation.

Private Inference on EC2 or EKS

- GPU instances (AI-optimized servers). g4dn, g5, p3, p4.
- Full control of inference stack.
- Mature VPC and IAM model.
- Higher operational ownership (scaling, patching, monitoring, incident response).

Voice to Text

Amazon Transcribe

- Real time and batch transcription.
- Medical transcription supported.
- HIPAA eligible under AWS BAA.
- Integrates with S3 and KMS for secure storage.

Amazon Transcribe — PII-Aware Speech-to-Text (HIPAA Context)

- **Real-time streaming with inline PII redaction**: Detects and masks names, phone numbers, emails, SSNs, and other identifiers before returning text; redacted entities appear as placeholders (e.g., `[NAME]`, `[PHONE_NUMBER]`).
- **Post-call processing + analytics**: For recorded audio, generates a redacted transcript plus optional sentiment, call categorization, key phrase extraction, and structured metadata for QA and reporting.
- **Compliance posture (BAA)**: When used under the AWS BAA and in approved regions, early redaction reduces downstream PHI exposure and simplifies audits and access controls.

Summary

- Strong fit for AWS-centric organizations using IAM/VPC/KMS as standard building blocks.
- Bedrock enables multi-model access; HIPAA eligibility typically needs to be confirmed per model, region, and feature set.
- AI workloads often span multiple services (e.g., Bedrock/Transcribe/S3/KMS/CloudWatch), so governance should cover each component.

Google Cloud Platform (GCP)

AI Inference Options

Vertex AI (Managed)

- Gemini family models.
- HIPAA BAA available.
- Feature availability and regional coverage can differ by model and SKU; confirm for your intended usage.

Private Inference on Compute Engine or GKE

- GPU backed instances (AI-optimized servers) available.

- Full stack control similar to Azure and AWS.

Voice to Text

Cloud Speech to Text

- Streaming and batch transcription.
- Medical speech models available.
- HIPAA eligible when configured correctly.

Summary

- Viable option for organizations already standardized on GCP services and tooling.
- Common considerations: confirming that the specific Vertex AI model/features and Speech-to-Text configuration are in-scope for the BAA and chosen regions.

Comparative Snapshot

Capability	Azure	AWS	Google
Managed LLM under BAA	Yes	Yes	Yes
Private LLM inference	Yes	Yes	Yes
GPU VM availability	Yes	Yes	Yes
HIPAA Speech to Text	Yes	Yes	Yes

What Self Hosting Actually Means

Self-hosted inference means you operate the entire AI execution stack inside your own trust boundary. This is infrastructure, software, and process ownership.

Core Components

Compute

- Dedicated GPU VMs or on premise servers.
- No shared inference fabric.
- Capacity planned for peak concurrency.

Model Runtime

- vLLM as the inference server.
- OpenAI compatible API surface.
- Continuous batching and KV cache reuse for throughput.

Models

- Open weight models selected per tier.
- Version pinned and change controlled.
- No automatic upgrades.

Networking

- Private VNet or VPC only.
- No public ingress unless explicitly required.
- Service to service authentication.

Storage

- Local or managed disks for model weights.
- HIPAA scoped databases for embeddings if used.
- Explicit retention policies.

Operational Responsibilities

You own.

- Scaling decisions and GPU saturation management.
- Model updates and regression testing.
- Security patching and OS hardening.
- Prompt logging policy and redaction.
- Availability and incident response.

vLLM reduces but does not eliminate these responsibilities.

Why vLLM

- High throughput under concurrent load.
- Efficient GPU memory usage.
- OpenAI compatible endpoints.
- Actively maintained and production proven.

When Self Hosting Is the Right Choice

- High volume inference where token costs dominate.
- Tenants with strict data locality requirements.
- Clients uncomfortable with third party model endpoints.
- Long lived prompts with large context windows.

When It Is Not

- Low volume or experimental features.
- Teams without GPU or ML ops experience.
- Workloads with unpredictable spikes and low baseline usage.

Framing for Auditors and Clients

Self-hosted inference is treated as an internal compute workload. There are no downstream subprocessors. All PHI remains inside controlled infrastructure.

Pricing Considerations (High Level)

Pricing varies by region, model size, and usage patterns. The goal here is relative comparison, not exact quotes.

Managed LLM Inference

Azure OpenAI

- Priced per 1K tokens (input and output).
- No infrastructure management cost.
- Predictable for steady workloads.
- Can become expensive for long prompts or high volume tenants.

AWS Bedrock

- Per token or per request depending on model.
- Pricing varies significantly by model provider.
- Slightly higher cognitive overhead to forecast spend.

Google Vertex AI

- Per token pricing.
- Often competitive on paper.
- Healthcare-specific examples and reference architectures vary by provider, region, and partner ecosystem.

HIPAA-Eligible Managed AI Services Pricing (USD, US East)

Scope

- Managed LLM inference and real-time speech-to-text
- Providers: Azure, AWS, Google Cloud
- Region focus: US East
- Pricing in USD
- Only official vendor sources
- Last accessed: **January 2026**

Pricing & Compliance Table

Provider	Service	Pricing Model	Key Rates (US East)	Pricing Link	HIPAA / BAA Documentation	Notes
Azure	Azure OpenAI Service (LLM inference)	Token-based (per 1K tokens)	**GPT-5 Data Zone (Azure OpenAI – East US)** Pricing per **1M tokens**. Input **$1.38**, Cached Input **$0.14**, Output **$11.00**. Batch API: Input **$0.69**, Cached Input **$0.07**, Output **$5.50**. *Region-bound Data Zone model. Required for HIPAA*	Pricing	HIPAA/BAA	Requires Azure OpenAI resource approval. Covered under Azure BAA. East US supported.

Provider	Service	Pricing Model	Key Rates (US East)	Pricing Link	HIPAA / BAA Documentation	Notes
			workloads. **GPT-4 (8K):** $0.03 prompt / $0.06 completion per 1K tokens **GPT-4 (32K):** $0.06 / $0.12 per 1K tokens **GPT-3.5 Turbo:** ~$0.0015 prompt / $0.0020 output per 1K token			
Azure	Azure AI Speech (Speech-to-Text)	Per audio minute (billed per second)	**Standard real-time:** ~$1.00/hour (~$0.0167/min) **Custom models:** ~$2.88/hour (~$0.048/min)	Pricing	HIPAA/BAA	No separate “medical” SKU. HIPAA use allowed under BAA with standard or custom models.
AWS	Amazon Bedrock (LLM inference)	Token-based (per 1K tokens)	(per 1k tokens) **Claude Opus 4.5:** ~$0.005 input / $0.025 output. **Llama 4 Maverick 17B:** $0.00024 in / $0.00097 out (Batch: $0.00012 / $0.000485). **Llama 4 Scout 17B:**	Pricing	HIPAA/BAA	On-demand pricing shown. Batch inference discounted US-East-1 supported.

Provider	Service	Pricing Model	Key Rates (US East)	Pricing Link	HIPAA / BAA Documentation	Notes
			$0.00017 in / $0.00066 out (Batch: $0.000085 / $0.00033). *Model availability and pricing vary by region and tier.*			
AWS	Amazon Transcribe (Speech-to-Text)	Per second of audio	**Streaming with PII Redaction:** ~$0.00240 per minute (≈ **$0.144/hour**) for first tier; volume-discounted at higher usage. **Post-Call Analytics / Medical Transcription:** ~$0.00056/sec (≈ **$2.00/hour**). *Streaming redacts PII inline; post-call processing applies full redaction, sentiment, and analytics. HIPAA-eligible under AWS BAA.*	Pricing	HIPAA/BAA	"Amazon Transcribe Medical" explicitly HIPAA-eligible. Real-time and batch same base rate.
Google Cloud	Vertex AI (Gemini models)	Token-based (per	**Google Gemini (latest): Gemini 3 Pro**	Pricing	HIPAA/BAA	Covered under Google Cloud BAA

Provider	Service	Pricing Model	Key Rates (US East)	Pricing Link	HIPAA / BAA Documentation	Notes
		1M tokens)	**Preview:** $2 input / $12 output per 1M tokens (≤200K context). **Gemini 3 Flash Preview:** $0.50 input / $3.00 output per 1M tokens. *Grounding billed separately; pricing varies by context length.*			when enabled. us-east1 supported.
Google Cloud	Cloud Speech-to-Text	Per audio minute (billed per second)	**Standard:** $0.016/min (0–500k min/month) **Medical models:** $0.078/min	Pricing	HIPAA/BAA	Medical models require BAA. Streaming and batch priced identically.

Private Inference (GPU VMs)

Cost components.

- VM hourly rate.
- GPU type and count.
- Storage and networking.
- Ops and maintenance time.

Typical ranges.

- Entry inference GPU (T4 class). Lower cost, good for small and medium models.

- Mid tier inference GPU (A10). Higher throughput, better latency.
- High end GPU (A100). Reserved for large models or high concurrency.

Tradeoff.

- Higher fixed cost.
- Lower marginal cost per request at scale.
- Full cost predictability.

Voice to Text Pricing

Azure Speech

- Charged per audio hour.
- Separate rates for real time and batch.
- Medical vocabularies may carry premium rates.

AWS Transcribe

- Charged per audio minute.
- Medical transcription priced separately.

Google Speech to Text

- Charged per audio minute.
- Medical models priced separately.

Cost Strategy (Common Patterns)

- Low volume or exploratory. Managed services.
- Steady production workloads. Managed LLM plus strict prompt controls.
- High volume or sensitive tenants. Private inference on GPU VMs.

Private Hosting (GPU VMs + vLLM)

HIPAA-Eligible AI Capabilities — Executive Summary

Scope

- Providers: Azure, AWS, Google Cloud
- Region: US East
- Workload: vLLM-based LLM inference on GPU VMs

- Pricing: On-demand, USD (approximate; varies by region and time)
- Deployment: Managed VM images (DLVM / Deep Learning AMIs)

GPU VM Tier Comparison

Provider	Tier	Example VM (GPU)	Approx $/hr (1 / 2+ GPU)	Notes
Azure	Entry (T4)	NCas T4 v3	~$0.53 / $4.3520 (4 GPU)	Good for small models and low concurrency. 4×T4 available but inefficient per GPU.
AWS	Entry (T4)	g4dn.xlarge	~$0.797 / $4.147(4GPU)	Cost-effective entry GPU. Up to 4×T4 on larger SKUs.
Google Cloud	Entry (T4)	N1 + T4	~0.98/ $1.37(2GPU)/ 2.14(gpu)	Often competitive T4 list pricing; add base VM cost (~$0.15/hr).
Azure	Mid (A10)	NVads A10 v5	~$3.20 / ~$6.50 / N/A	Strong mid-tier inference. Max 2 GPUs per VM.
AWS	Mid (A10G)	g5.xlarge	~$0.79223 / ~$4.4667(4 GPU)	Common price/performance choice. Up to 4 GPUs per VM. Resource
Google Cloud	Mid (L4*)	G2 + L4	~$1.62	*GCP uses L4 instead of A10. Optimized for inference.*
Azure	High (A100)	ND A100 v4	$9,946.76 (per 8 gpu)	High throughput. InfiniBand. Suited for large models. Requires min of 8gpus in increments of 8.
AWS	High (A100)	p4d.24xlarge	$8,496 (per 8 gpu)	Dense 8-GPU nodes. Competitive $/A100 vs. list pricing; confirm current rates. Requires min of 8gpus in increments of 8.
Google Cloud	High (A100)	a2-highgpu-1g	$2,849.19 / $5,572.45(2) / $21,725.30(8) / $41,332.77(16)	Flexible scaling options; per-GPU list price can vary by shape/region.

Total Cost Components

- **GPU VM hourly**: Base VM + GPU accelerator
- **GPU type & count**: Primary cost driver
- **Storage**: OS disk, model weights, checkpoints
- **Networking**: Egress, inter-AZ traffic
- **Ops & maintenance**: Patching, monitoring, scaling, on-call

Cost Model Tradeoffs

Private GPU hosting has **fixed hourly costs** regardless of request volume. This provides predictable performance and full data control but can result in idle spend. Managed or serverless inference shifts costs to **per-request pricing**, which is efficient for bursty workloads but introduces latency, scaling limits, and vendor coupling. Private hosting is favored when compliance, data residency, or sustained throughput matters.

> **Savings note:** Reserved instances and savings plans can reduce GPU VM costs by ~30–60% but reduce flexibility.

Common Selection Criteria (Non-Prescriptive)

When choosing a provider and deployment model, teams typically evaluate:

- **BAA scope**: which exact services/features/models are covered and in which regions.
- **Network isolation**: private endpoints and egress controls for PHI boundaries.
- **Identity and access**: IAM/Entra roles, service identities, least-privilege patterns, and auditability.
- **Logging and retention**: where prompts/transcripts/logs land, retention controls, and redaction strategy.
- **Model availability**: which LLMs are available under HIPAA constraints, plus quotas and throughput limits.
- **Operational model**: managed APIs vs. private inference (GPU VMs/Kubernetes) and your on-call/patching capacity.
- **Cost drivers**: token volume, context size, output limits, caching, and fixed GPU utilization vs. burst usage.

References (Official Sources)

Azure

- Azure OpenAI pricing
- Azure AI Speech pricing
- Azure HIPAA / BAA
- Azure VM pricing
- Azure VM series docs (NCas / NVads / ND)

AWS

- Amazon Bedrock pricing
- Amazon Transcribe pricing
- AWS HIPAA-eligible services list
- EC2 on-demand pricing
- EC2 instance types (G4 / G5 / P4)

Google Cloud

- Vertex AI pricing (Gemini)
- Cloud Speech-to-Text pricing
- Google Cloud HIPAA / BAA
- GPU pricing
- GPU machine types (A2 / G2)