# Latest LLM, Embedding Models & Vector DB Comparison Guide

By Rajesh Srivastava



Vector databases, interactive · aimlcompanion.ai

# Comprehensive LLM API Pricing Table (Nov 2025)

## OpenAI Models

| Model | Input ($/1M tokens) | Output ($/1M tokens) | Cached Input | Context Window | Notes |
|---|---|---|---|---|---|
| **GPT-5** | $1.25 | $10.00 | $0.125 | 400K | Flagship model, released Aug 2025 |
| **GPT-5 Mini** | $0.25 | $2.00 | $0.025 | 200K | Cost-efficient GPT-5 variant |
| **GPT-5 Nano** | $0.05 | $0.40 | $0.005 | 32K | Fastest, cheapest GPT-5 |
| **GPT-4.1** | $2.00 | $8.00 | $0.50 | 1M | Improved coding, long-context |
| **GPT-4.1 Mini** | $0.40 | $1.60 | $0.10 | 1M | Same as o3 pricing |
| **GPT-4.1 Nano** | - | - | - | - | Entry-level |
| **o3** | $0.40 | $1.60 | - | 200K | Reasoning model, 80% price cut |
| **o3 (standard rate)** | $2.00 | $8.00 | - | 200K | Full reasoning mode |
| **o4-mini** | $1.10 | $4.40 | - | 200K | Efficient reasoning model |
| **o4-mini-high** | $1.10 | $4.40 | - | 200K | Enhanced reasoning |
| **GPT-4o** | $5.00 | $20.00 | - | 128K | Being deprecated Feb 2026 |
| **GPT-4o Mini** | $0.15 | $0.60 | - | 128K | Legacy mini model |

---

## Anthropic Claude Models

| Model | Input ($/1M tokens) | Output ($/1M tokens) | Context Window | Notes |
|---|---|---|---|---|
| **Claude Opus 4.1** | $15 | $75 | 200K | Frontier reasoning model |
| **Claude Sonnet 4.5** | $3 | $15 | 200K (1M beta) | Best for coding, current flagship |
| **Claude Sonnet 4** | $3 | $15 | 200K | Balanced mid-tier |

| Model | Input ($/1M tokens) | Output ($/1M tokens) | Context Window | Notes |
|---|---|---|---|---|
| **Claude Haiku 4.5** | $1 | $5 | 200K | 4-5x faster than Sonnet 4.5 |
| **Claude Haiku 3.5** | $0.80 | $4 | 200K | Previous fast model |
| **Claude Haiku 3** | $0.25 | $1.25 | 200K | Cheapest Claude |

**Claude Discounts:**

- Batch API: 50% discount on input and output
- Prompt caching: Cache writes 1.25x base price, cache hits only 0.1x base price

---

## Google Gemini Models

| Model | Input ($/1M tokens) | Output ($/1M tokens) | Context Window | Notes |
|---|---|---|---|---|
| **Gemini 3 Pro Preview** | $2.00 (≤200K), $4.00 (>200K) | $12.00 (≤200K), $18.00 (>200K) | 2M | NEW - Best multimodal |
| **Gemini 2.5 Pro** | $1.25 (≤200K), $2.50 (>200K) | $10.00 (≤200K), $15.00 (>200K) | 2M | Coding/reasoning flagship |
| **Gemini 2.5 Flash** | $0.30 (text/image/video), $1.00 (audio) | $2.50 | 1M | Hybrid reasoning model |
| **Gemini 2.5 Flash-Lite** | $0.10 (text/image/video), $0.30 (audio) | $0.40 | 1M | Most cost-efficient |
| **Gemini 2.0 Flash** | $0.10 (text/image/video), $0.70 (audio) | $0.40 | 1M | Balanced multimodal |
| **Gemini 2.0 Flash-Lite** | $0.075 | $0.30 | 1M | Budget option |
| **Gemini Embedding** | $0.15 | - | - | New embedding model |

**Google Batch Pricing:** 50% discount on batch processing

---

## DeepSeek Models

| Model | Input Cache Hit | Input Cache Miss | Output ($/1M tokens) | Context | Notes |
|---|---|---|---|---|---|
| **deepseek-chat (V3.2-Exp)** | $0.07 | $0.27 | $1.10 | 64K | Non-thinking mode |
| **deepseek-reasoner (V3.2-Exp)** | $0.14 | $0.55 | $2.19 | 64K | Thinking mode |
| **DeepSeek V3.2-Exp (latest)** | $0.028 | $0.28 | $0.42 | 128K | 50% cheaper than V3.1 |

---

## xAI Grok Models

| Model | Input ($/1M tokens) | Output ($/1M tokens) | Cached Input | Context Window | Notes |
|---|---|---|---|---|---|
| **Grok 4** | $3.00 | $15.00 | $0.75 | 256K | Released July 2025 |
| **Grok 4 Fast** | $0.20 | $0.50 | - | 256K | High-throughput variant |
| **Grok 3** | $3.00 | $15.00 | - | 131K | Standard tier |
| **Grok 3 (fast)** | $5.00 | $25.00 | - | 131K | Fast mode |
| **Grok 3 Mini** | $0.30 | $0.50 | - | 131K | Lightweight |
| **Grok 3 Mini (fast)** | $0.60 | $4.00 | - | 131K | Fast mini |

**Grok Live Search:** $25 per 1,000 sources ($0.025 per source)

---

# Embedding Models

| Model | Dimensions | Use Case | Cost |
|---|---|---|---|
| CodeBERT | 768 | General code (Recommended) | Free |
| StarEncoder | 768 | Code search optimized | Free |
| OpenAI text-embedding-3-small | 1536 | High quality, general purpose | ~$0.02/1M tokens |
| OpenAI text-embedding-3-large | 3072 | Highest quality | ~$0.65/1M tokens |
| OpenAI text-embedding-ada-002 | 1536 | Legacy model | ~$0.10/1M tokens |
| Cohere embed-v3 | 1024 | Code + text hybrid, multilingual | ~$0.10/1M tokens |
| Google text-embedding-004 | 768 | General purpose, latest Gemini embedding | ~$0.025/1M tokens |
| Google text-multilingual-embedding-002 | 768 | Multilingual support | ~$0.025/1M tokens |
| Voyage Code-2 | 1536 | Code-specialized | ~$0.12/1M tokens |
| Jina AI v2 | 768 | 8K context, bilingual | ~$0.02/1M tokens |
| Nomic embed-text-v1.5 | 768 | Long context (8K), open source | Free |

# Vector DB

## Comparison

| Database | Latency | Filtering | Monthly Cost (10M vectors) | Deployment Options | Free Tier |
|---|---|---|---|---|---|
| **Pinecone** | Sub-50ms at billion-scale | Advanced - namespaces, hybrid search (vector + keyword) | $50/month minimum (Standard), $500/month minimum (Enterprise) | Serverless Cloud only (AWS, GCP, Azure) | **Starter**: Free, 5 indexes, 2GB storage, AWS us-east-1 only |
| **Weaviate** | 23-34ms p95 | Excellent - GraphQL, hybrid search, native AI modules | $25-153/month (serverless) | Serverless Cloud, Dedicated Cloud, BYOC, Self-host | 14-day free sandbox |
| **Qdrant** | Lowest latencies, 4x RPS gains | Excellent - pre-filtering HNSW, payload filters, geo-locations | $27-102/month (with quantization) | Qdrant Cloud, Hybrid Cloud, On-premise, Self-host | Free tier available |
| **ChromaDB** | ~20ms p50, ~170ms p99 | Good - metadata filtering, hybrid search, range searches | $0 (self-host) or usage-based cloud | Self-host, Chroma Cloud (serverless) | $5 free credits on Cloud |
| **Milvus** | Sub-10ms | Excellent | $0 (OSS) / $99+ (Zilliz) | Cloud + Self-host | Billion-scale, GPU support |
| **pgvector** | 10-100ms | Good | $0 (extension) | Self-host / managed PG | SQL integration, existing PG users |
| **AWS S3 Vectors** | Sub-second | Basic | ~$0.04/GB storage | Cloud only (AWS) | Cost-optimized, infrequent queries |
| **LanceDB** | <20ms | Excellent | $0 (OSS) / Cloud pricing | S3 + Self-host | Multimodal, serverless, cost-efficient |

## Performance & Scale

| Database | Best For | Max Scale | Max Dimensions | Notes |
|---|---|---|---|---|
| **Pinecone** | Production apps requiring real-time responses at billion-scale | Billions of vectors | 20,000 (dense) | Serverless architecture handles sharding, replication, load balancing automatically |
| **Weaviate** | Complex data relationships, knowledge graphs, multi-modal | Hundreds of millions | No hard limit | Hybrid filtering with 22% lower monthly bill vs Pinecone at steady traffic |
| **Qdrant** | Fast query processing, complex metadata filtering | Billions of vectors | 65,536 (dense) | Written in Rust for speed under high load |
| **ChromaDB** | Prototypes, development, applications under 10M vectors | Tens of millions (single machine) | No hard limit | 2025 Rust rewrite: 4x faster writes/queries |
| **Milvus** | Billion-scale enterprise, GPU-accelerated indexing | Billions of vectors | 32,768 (configurable) | Milvus is available as a completely free open-source project under the Apache 2.0 license. Supports GPU acceleration for indexing |
| **pgvector** | SQL integration, teams already using PostgreSQL | Postgres/pgvector realistically maxes out at 10–100 million vectors before it slows unacceptably. | 2,000 (standard), 4,000 (halfvec), 64,000 (binary) | pgvector 0.8.0 demonstrates 3-5× query throughput improvements over previous versions in production workloads. |
| **AWS S3 Vectors** | Cost-optimized storage, infrequent queries, AI agent memory | Billions of vectors | 4,096 | The service currently supports only floating-point (float32) vector data.Sub-second latency, not for high QPS |
| **LanceDB** | Multimodal AI, serverless on S3, | We are currently working with customers who | No hard limit | Storing 10M vectors with 50% overwrites |

| Database | Best For | Max Scale | Max Dimensions | Notes |
|---|---|---|---|---|
| | cost-efficient at scale | regularly perform operations on 200M+ vectors, and we're fast approaching billion scale and beyond. | | and 500K queries can cost $500+ with a typical vector database. This can easily drop to $50/month with LanceDB. |

---

## Pricing Details

| Database | Pricing Model | Minimum Commitment | Cost for 1M vectors (768-1536 dims) |
|---|---|---|---|
| **Pinecone** | Usage-based (read units, write units, storage) | $50/month (Standard), $500/month (Enterprise) | ~$50+ (minimum applies) |
| **Weaviate** | Vector dimensions + storage + backups | $25/month (serverless) | Our test of 1536 dimensions with 1 million reads and writes works out to $153 but if you choose the less performant compression version, it's only $25. |
| **Qdrant** | CPU, memory, disk storage usage | Pay-as-you-go (free tier available) | A pricing calculator is available, and based on our standard test the price is $102 on AWS us-east without quantization, which can reduce memory usage. With disk caching and quantization turned on, this can be reduced to $27. |
| **ChromaDB** | GB written/stored/queried (cloud) | None ($5 free credits) | Chroma's cloud offering has simple usage based pricing with a nice pricing calculator which works out to $81 for 1536 dimension vector with 1 million writes and 1 million queries. |
| **Milvus/Zilliz** | vCU consumption + storage (managed) | $99 per month for dedicated tiers. Free (self-hosted) | $0 (OSS) / $99+ (Zilliz Cloud dedicated) |
| **pgvector** | PostgreSQL hosting costs | $0 (extension is free) | Varies by PG hosting (~$20-100 for managed PG) |

| Database | Pricing Model | Minimum Commitment | Cost for 1M vectors (768-1536 dims) |
|---|---|---|---|
| **AWS S3 Vectors** | Storage + queries | None (pay-as-you-go) | Storage costs from $0.30 to $0.04 per GB per month. Very low for infrequent queries |
| **LanceDB** | S3 storage + compute (serverless) | None (OSS is free) | ~$50/month or less with S3 backend |

---

## Features Comparison

| Feature | Pinecone | Weaviate | Qdrant | ChromaDB | Milvus | pgvector | S3 Vectors | LanceDB |
|---|---|---|---|---|---|---|---|---|
| **Hybrid Search** | ✅ Yes | ✅ Dense + sparse BM25 | ✅ Yes | ✅ Basic | ✅ Sparse + dense | ✅ With extensions | ❌ Vector only | ✅ Yes |
| **Built-in Embeddings** | ❌ No | ✅ OpenAI, Cohere, HuggingFace | ✅ Cloud Inference (July 2025) | ✅ Sentence Transformers | ✅ Multiple integrations | ❌ No | ✅ Bedrock integration | ✅ Multiple integrations |
| **Multi-tenancy** | ✅ Namespaces | ✅ Native | ✅ Flexible sharding | ✅ Collections | ✅ Partitions | ✅ Schemas/tables | ✅ Vector indexes | ✅ Tables/namespaces |
| **Quantization** | ✅ PQ | ✅ PQ, BQ | ✅ 24x compression | ❌ Limited | ✅ PQ, SQ, BQ | ✅ halfvec, binary | ❌ float32 only | ✅ PQ |
| **Self-host Option** | ❌ No | ✅ Yes | ✅ Yes | ✅ Yes | ✅ Yes | ✅ Yes (PG extension) | ❌ No (AWS only) | ✅ Yes |
| **SOC 2 / HIPAA** | ✅ Both | ✅ SOC 2 | ✅ Available | ❌ No | ✅ SOC 2 Type II, ISO 27001, GDPR (Zilliz) | Depends on PG host | ✅ AWS compliance | ❌ No |
| **GPU Acceleration** | ❌ No | ❌ No | ✅ GPU indexing | ❌ No | ✅ Yes | ❌ No | ❌ No | ❌ No |
| **Full SQL Support** | ❌ No | ❌ GraphQL | ❌ No | ❌ No | ❌ No | ✅ Full PostgreSQL | ❌ No | ✅ SQL queries |

---

## When to Choose Each

| Database | Best Use Case | Avoid If... |
|---|---|---|
| **Pinecone** | Need turnkey scale, guaranteed SLAs, minimal ops overhead | Budget <$50/month or need self-hosting |
| **Weaviate** | Need vector search + complex data relationships, GraphQL interface | Need absolute lowest latency |
| **Qdrant** | Need complex filters, OSS flexibility, best price-performance | Prefer simpler learning curve |
| **ChromaDB** | Prototyping, development, <10M vectors, Python-native | Need enterprise scale or built-in HA |
| **Milvus** | Billion-scale workloads, need GPU acceleration, full OSS control | Organizations must meet strict regulatory or compliance standards often choose this path for maximum control. Want zero ops (use Zilliz Cloud instead) |
| **pgvector** | Already using PostgreSQL, need SQL + vector in one DB, moderate scale | Billion-Scale Vector Requirements: Datasets exceeding 500 million vectors often benefit from specialized infrastructure. Need >100M vectors or sub-10ms latency |
| **AWS S3 Vectors** | Use S3 Vectors for large, long-term vector data that doesn't require the high-throughput performance of in-memory vector databases. AI agent memory, cost-optimized archival | Need high QPS, real-time responses, or non-AWS infrastructure |
| **LanceDB** | LanceDB makes sense when you need something that is efficient but not at the level of the best tools, good enough and cheaper. Multimodal data, serverless on S3, cost-sensitive | Need managed enterprise support or highest possible QPS |

---

## Deployment Options Summary

| Database | Cloud Managed | Self-Hosted | BYOC | Serverless |
|---|---|---|---|---|
| **Pinecone** | ✅ Only option | ❌ No | ❌ No | ✅ Yes |
| **Weaviate** | ✅ Weaviate Cloud | ✅ Docker/K8s | ✅ Yes | ✅ Yes |
| **Qdrant** | ✅ Qdrant Cloud | ✅ Docker/K8s | ✅ Hybrid Cloud | ✅ Yes |
| **ChromaDB** | ✅ Chroma Cloud | ✅ Docker/pip | ❌ No | ✅ Yes |
| **Milvus** | ✅ Zilliz Cloud | ✅ Docker/K8s/Helm | ✅ Yes | ✅ Yes |
| **pgvector** | ✅ RDS/Aurora/Cloud SQL | ✅ Any PostgreSQL | N/A | Via managed PG |

| Database | Cloud Managed | Self-Hosted | BYOC | Serverless |
|---|---|---|---|---|
| **AWS S3 Vectors** | ✅ Only option | ❌ No | ❌ No | ✅ Yes |
| **LanceDB** | ✅ LanceDB Cloud | ✅ Embedded/S3 | ✅ Yes | ✅ Yes |

---

## Quick Decision Matrix

| Your Priority | Recommended Database |
|---|---|
| **Zero ops, enterprise SLAs** | Pinecone or Zilliz Cloud |
| **Best price-performance** | Qdrant or LanceDB |
| **Already using PostgreSQL** | pgvector |
| **Billion-scale + GPU** | Milvus |
| **Complex data relationships** | Weaviate |
| **Rapid prototyping** | ChromaDB or LanceDB |
| **AWS-native, cost-optimized archival** | AWS S3 Vectors |
| **Multimodal + serverless** | LanceDB |
| **Hybrid search critical** | Weaviate or Qdrant |
| **Compliance-heavy (HIPAA, SOC 2)** | Pinecone, Zilliz Cloud, or Weaviate Enterprise |

---