

SERVERLESS 2025: FROM CONCEPT TO CLOUD – IMPLEMENTING GENAI APPLICATIONS IN SERVERLESS ARCHITECTURES

by Guillermo Galvan Soltero

BEYOND SERVERS: KEY INSIGHTS DRIVING THE GENAI ERA

The serverless architecture market is experiencing rapid growth. Valued at approximately \$10.21 billion in 2023, it's projected to reach \$78.12 billion by 2032, with a compound annual growth rate (CAGR) of 25.42% from 2024 to 2032.

<https://www.thebusinessresearchcompany.com/report/generative-artificial-intelligence-ai-in-architecture-global-market-report>

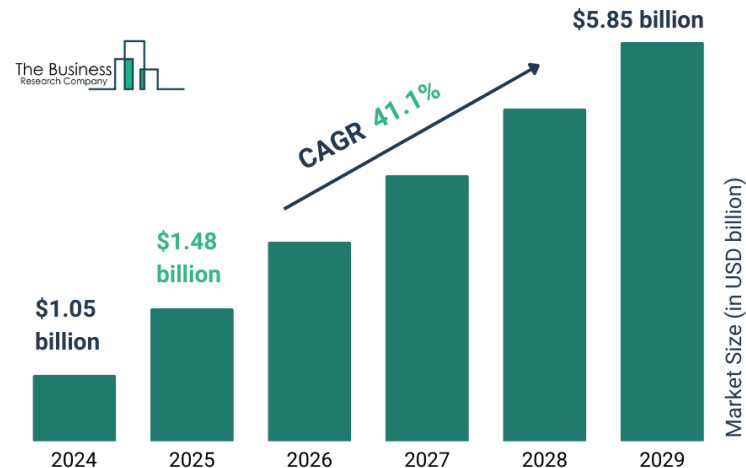
Serverless architectures are accelerating GenAI adoption by offering flexibility and scalability. They allow developers to focus on innovation without managing underlying infrastructure, making them ideal for GenAI applications.

<https://www.antstack.com/blog/how-serverless-is-accelerating-gen-ai-adoption/>

Financial institutions are leveraging serverless architectures to enhance their AI capabilities. For instance, JPMorgan Chase has integrated AWS's AI tools for massive data processing, improving both security and scalability.

<https://www.businessinsider.com/aws-wall-street-jpmorgan-bridgewater-mufg-rocket-mortgage-2025-2>

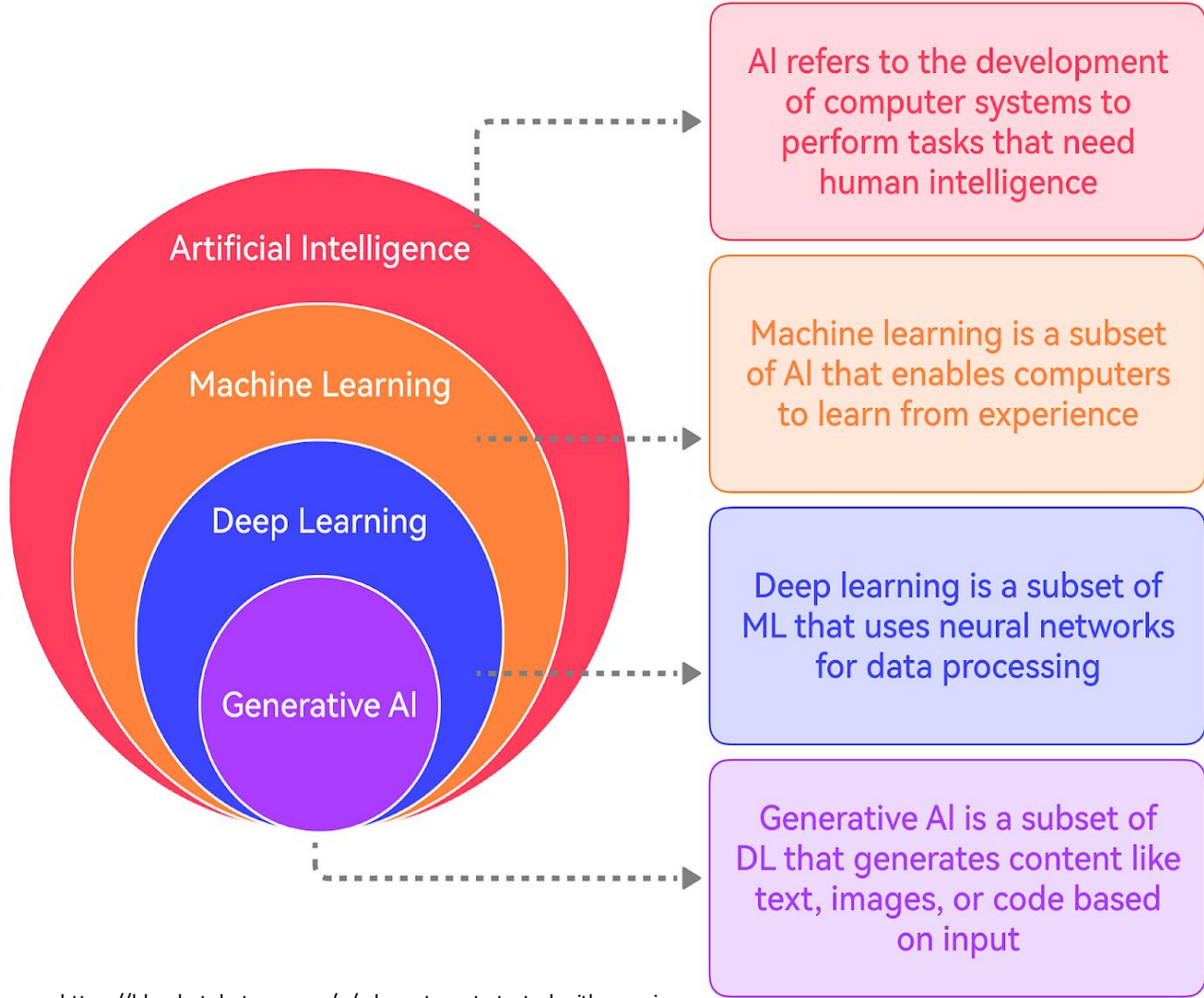
Generative Artificial Intelligence (AI) In Architecture Global Market Report 2025



Generative AI (GenAI) in architecture is also on the rise. The market is expected to grow from \$1.05 billion in 2024 to \$1.48 billion in 2025, reflecting a CAGR of 41.3%

<https://www.thebusinessresearchcompany.com/report/generative-artificial-intelligence-ai-in-architecture-global-market-report>

GENAI AND THEIR MODELS



<https://blog.bytebytego.com/p/where-to-get-started-with-genai>

Model Type	Description	Examples
Text Generation Models	Generate human-like text, enabling tasks such as content creation, summarization, and translation.	GPT-4, LaMDA, LLaMA, BLOOM
Text Generation Models	Produce or suggest programming code, assisting in software development and debugging.	OpenAI Codex, GitHub Copilot
Image Generation Models	Create images from textual descriptions or other images, useful in design and creative industries.	DALL-E, Midjourney, Stable Diffusion
Audio Generation Models	Generate music, speech, or other audio forms, facilitating tasks like music composition and voice synthesis.	Jukebox, MusicLM
Video Generation Models	Produce video content based on text prompts or existing footage, aiding in media production.	Sora by OpenAI, Runway's Gen-1 and Gen-2, Meta's Make-A-Video
Multimodal Models	Handle and integrate multiple data types, such as text, images, and audio, for comprehensive understanding and generation.	GPT-4 (multimodal version), CLIP, Gemini
3D Modeling Models	Generate 3D models from text, images, or videos, applicable in gaming, virtual reality, and animation.	3D-GPT
Molecular Generation Models	Design new molecular structures, aiding in drug discovery and material science.	AlphaFold

SERVERLESS 101








What is Serverless

- No server management (focus on code)
- Automatic scaling (capacity on demand)
- Pay-as-you-go (billed only on execution)

Why serverless

- Cost efficiency (no idle servers)
- Scalability (handles unpredictable workloads)
- Faster deployment (focus on features, not setup)

SERVERLESS GENAI APPS ON AWS

		Purpose	Key features	Use in Gen AI Apps
	Amazon Lambda	Serverless compute	Automatic scaling, pay-per-use, event-driven	Handle user requests, process AI calls.
	AWS Step Functions	Serverless workflow orchestration	Visual workflow creation, integration with various services, scalable	Orchestrate complex GenAI workflows, manage model training and deployment processes
	Amazon Bedrock	Inference for GenAI models	Access to multiple pre-trained models, RAG support, guardrails	Generate content, enhance responses with RAG
	Amazon SageMaker Inference	Serverless ML model inference	Managed endpoints, supports various ML models, scalable	Deploy custom or pre-trained ML models for real-time inference in GenAI apps
	Amazon RDS with Aurora PostgreSQL (pgvector)	High-performance vector database for GenAI apps	High performance, scalable, supports vector search, supports Aurora Serverless v2	Used for large-scale RAG implementations in GenAI apps
	Amazon OpenSearch	Vector database	Vector search, scalability, integration with Lambda	Store and query embeddings for RAG

APPLICATION USE CASE DEMO

DEBT MANAGER APP

A personal finance app that helps users simulate and manage debt repayment plans using real-time data, offering tailored insights for smarter financial decisions.



BUILDING OUR GENAI APP STEP BY STEP

softserve

AWS
User Groups

THANK YOU!