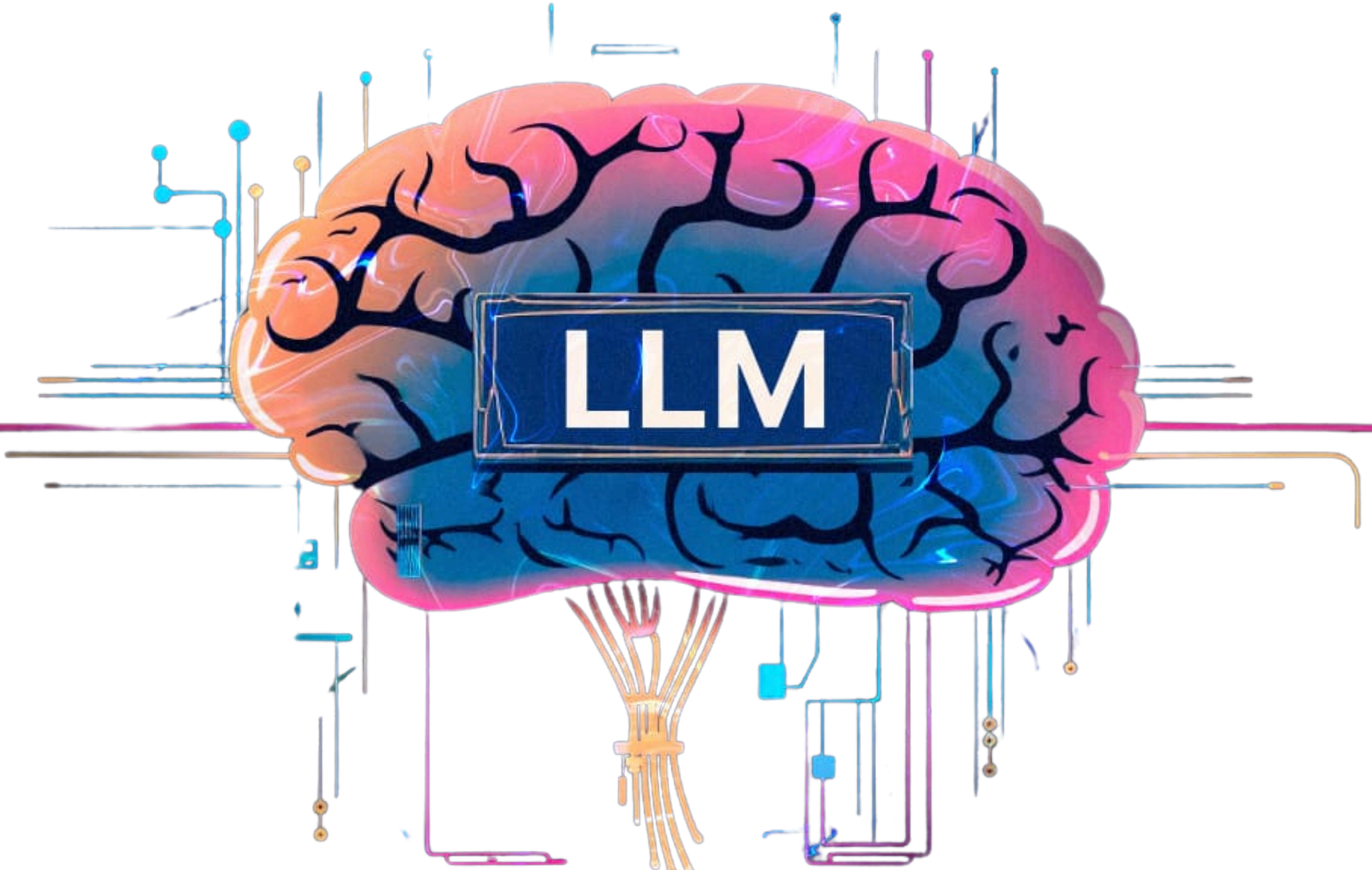
Greg Coquillo
Product Leader at AWS

TOP 7 POPULAR LLMS EXPLAINED




Greg Coquillo
Product Leader at AWS

What Are LLMs?

Large Language Models (LLMs) are AI systems trained on billions of texts to understand, generate, and reason using natural language..

They power tools like ChatGPT, Google Bard, Claude, and more—yet each LLM is unique in how it processes data and performs.

Greg Coquillo
Product Leader at AWS

BERT

Google's Language Understanding Breakthrough

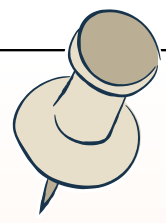
Bidirectional Encoder Representations from Transformer

- Reads input from both directions at once for full context
- Uses Masked Language Modeling (predict missing words) and Next Sentence Prediction
- Great for: Sentiment analysis, Q&A, sentence pair tasks
  📏 Sizes: BERT Base (110M), BERT Large (340M)
  ⚙️ Architecture: Encoder-only Transformer

Greg Coquillo
Product Leader at AWS

GPT

The Pioneer of Generative AI



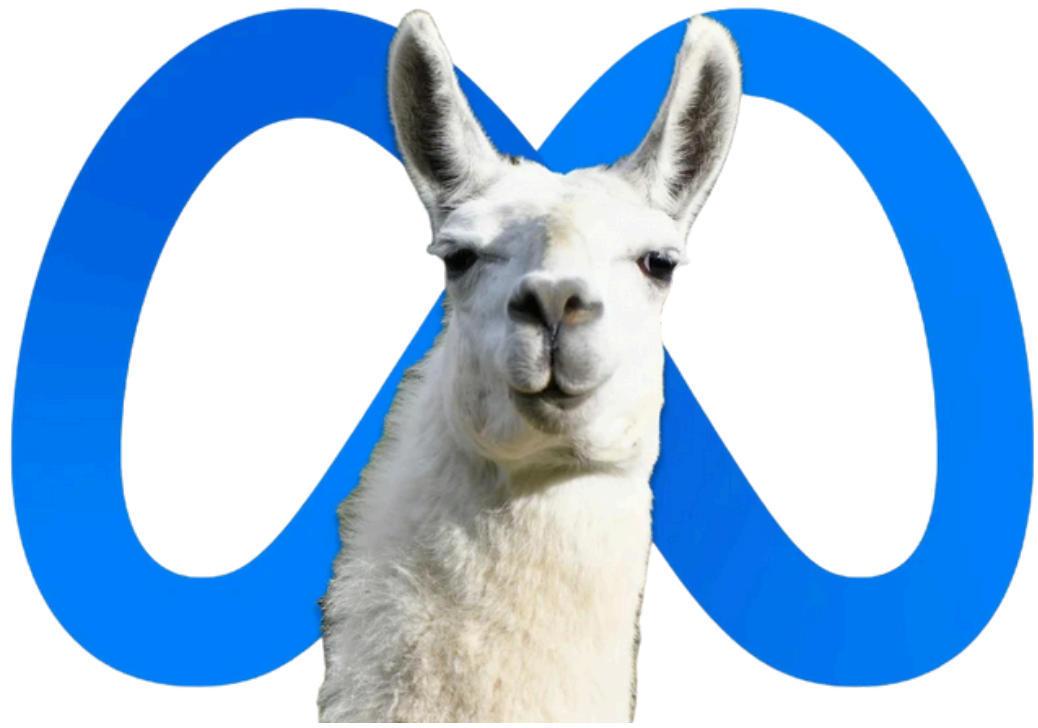
Generative Pre-trained Transformer by OpenAI

- Uses a decoder-only architecture for next-word prediction
- Pretrained on massive datasets, then fine-tuned or prompted for tasks
- GPT-4o (2024) added multimodal support (text + images)
- Great for: Creative writing, coding, chatting, and multi-turn conversations
  Note: Latest versions are proprietary with limited transparency

Greg Coquillo
Product Leader at AWS

LLAMA

Meta's Open-Source Powerhouse



Large Language Model Meta AI

- Open-source, decoder-only models ranging from 7B to 65B+ parameters
- Includes innovations like SwiGLU activations, Rotary Positional Embeddings (RoPE), and RMSNorm
- LLaMA 3 and 4 models show performance comparable to GPT-3.5 and GPT-4, at smaller sizes
- Great for: Research, on-device deployment, community fine-tuning

Greg Coquillo
Product Leader at AWS

PALM

Google's Scaling Supermodel

Pathways Language Model (2022–2023)

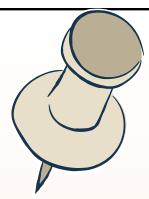
- Trained on 780B tokens using TPUs for extreme efficiency
- PaLM 2 enhanced multilingual, logic, and coding skills
- Powers tools like Google Bard and Workspace AI features
- Uses multi-query attention for memory efficiency
- Great for: Few-shot learning, translation, complex reasoning tasks

Greg Coquillo
Product Leader at AWS

GEMINI

DeepMind's Multimodal Future

Gemini



Google's next-gen LLM with native multimodal support

- Handles text, images, audio, video, and code in one model
- Uses Mixture-of-Experts (MoE) to activate only parts of the model per task—saving compute
- Gemini 2.5 (2025) supports 1M-token context, tool integrations, and enterprise applications
- Comes in sizes: Ultra, Pro, Nano
- Great for: Long-context tasks, reasoning, tool-augmented interactions

Greg Coquillo
Product Leader at AWS

Mistral

Open-Source Meets Efficiency

MISTRAL AI_

Developed by Mistral AI (France), launched 2023

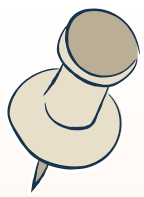
- Mistral 7B: Compact, fast, and outperforms larger models
- Mixtral 8×7B: A sparse Mixture-of-Experts (MoE) model combining scale with efficiency
- Mistral Medium 3 (2025): Proprietary model matching Claude Sonnet at lower cost
- Great for: Coding, logic, enterprise-grade reasoning
- Note: Early models were open; recent ones are enterprise-only

Greg Coquillo
Product Leader at AWS

DeepSeek

China's Logic-Oriented LLM

deepseek



Built by DeepSeek (spin-off from High-Flyer AI, 2023)

- Uses sparse MoE Transformer architecture with 670B total parameters but activates only 37B per response
- Highly efficient: balances power and cost
- Known for reasoning, multilingual generation, and FP8 training optimizations
- Great for: High-performance reasoning tasks, real-world deployment at scale

Greg Coquillo
Product Leader at AWS

RECAP – 7 LLMS AT A GLANCE

Model	Type	Speciality	Open/Closed
BERT	Encoder-only	Sentence understanding	Open
GPT	Decoder-only	Fluent generation, few-shot tasks	Closed
LLaMA	Decoder-only	Open & efficient large models	Semi-Open
PaLM	Decoder-only	Coding, multilingual, reasoning	Closed
Gemini	MoE + Multi	Multimodal, long context	Closed
Mistral	MoE + Sparse	High performance, open/closed mix	Mixed
DeepSeek	Sparse MoE	Reasoning, logic, low compute use	Open

Greg Coquillo

Product Leader at AWS