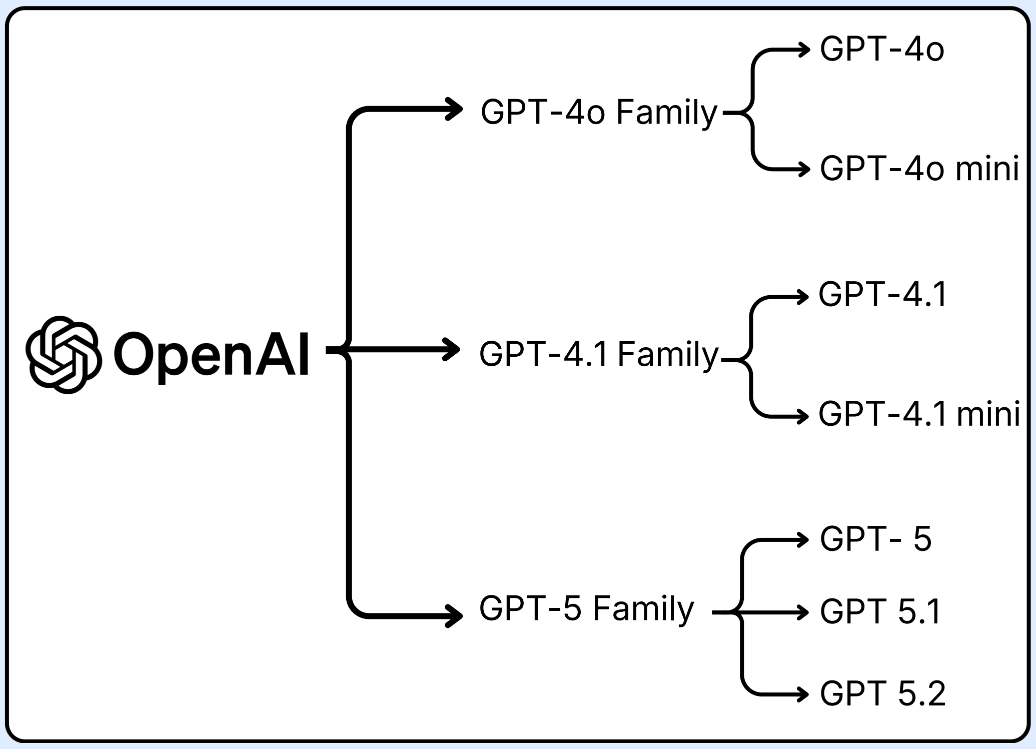
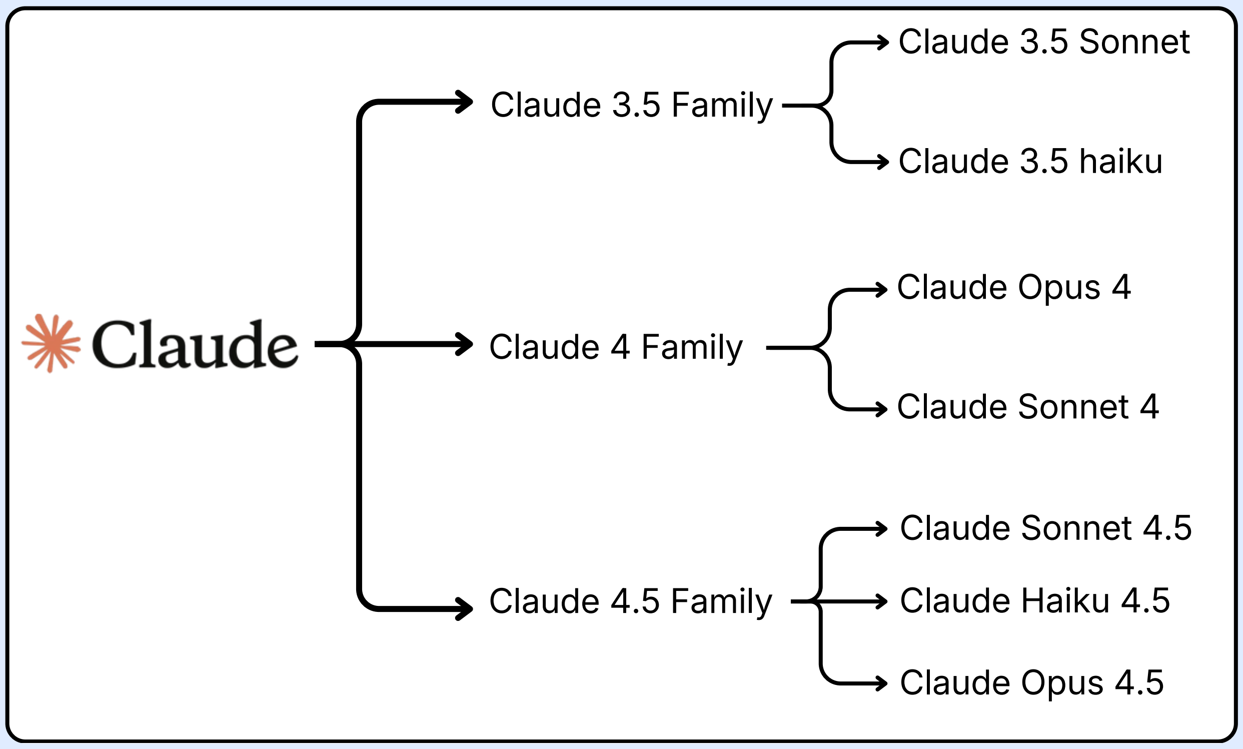
Analytics Vidhya

AI Accelerator Program

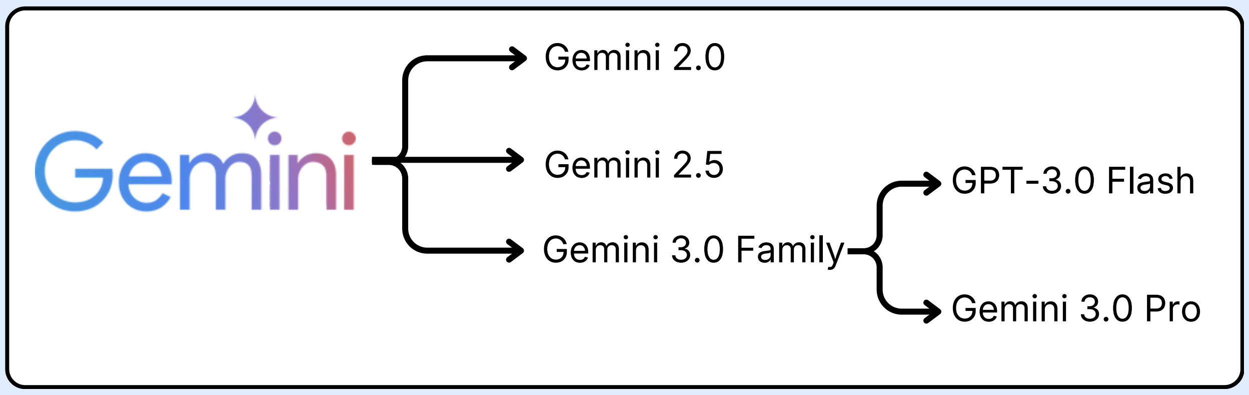
Current LLM Landscape



OpenAI
- GPT-4o Family
  - GPT-4o
  - GPT-4o mini
- GPT-4.1 Family
  - GPT-4.1
  - GPT-4.1 mini
- GPT-5 Family
  - GPT- 5
  - GPT 5.1
  - GPT 5.2

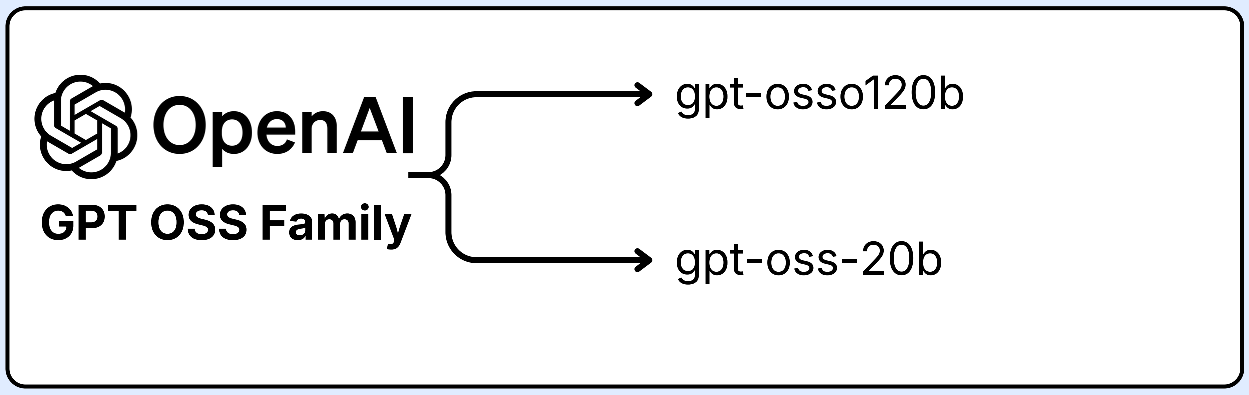


Claude
- Claude 3.5 Family
  - Claude 3.5 Sonnet
  - Claude 3.5 haiku
- Claude 4 Family
  - Claude Opus 4
  - Claude Sonnet 4
- Claude 4.5 Family
  - Claude Sonnet 4.5
  - Claude Haiku 4.5
  - Claude Opus 4.5

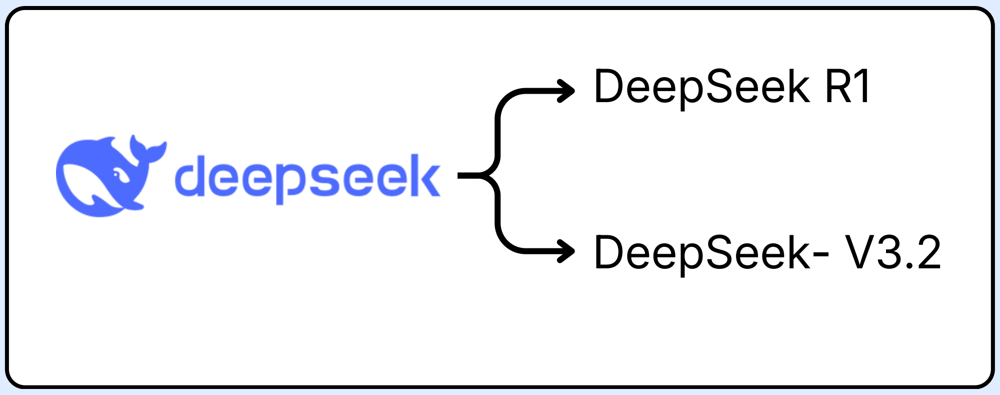


Gemini
- Gemini 2.0
- Gemini 2.5
- Gemini 3.0 Family
  - GPT-3.0 Flash
  - Gemini 3.0 Pro

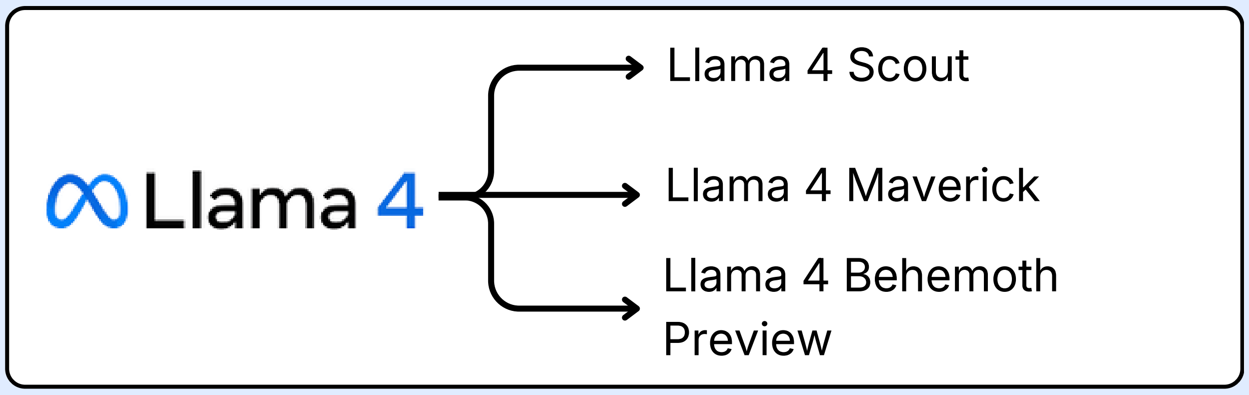
These all are **commercial LLMs** that means they are proprietary language models developed and offered by companies as paid services for business and practical applications.



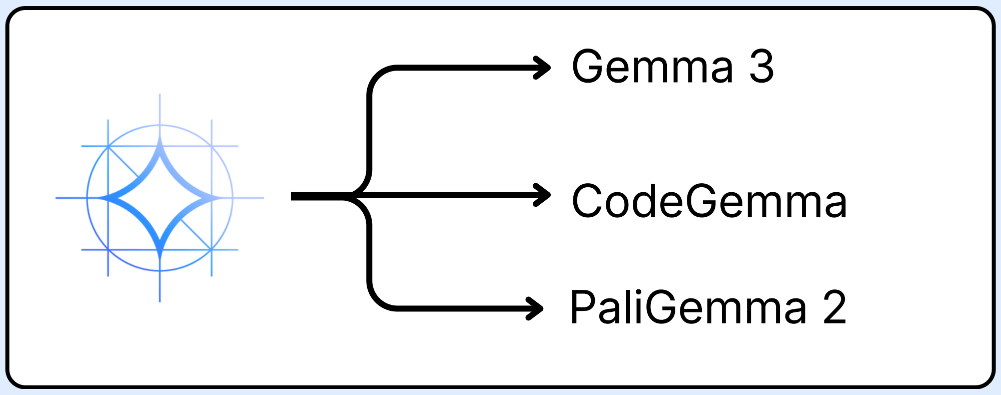
OpenAI
GPT OSS Family
- gpt-osso120b
- gpt-oss-20b



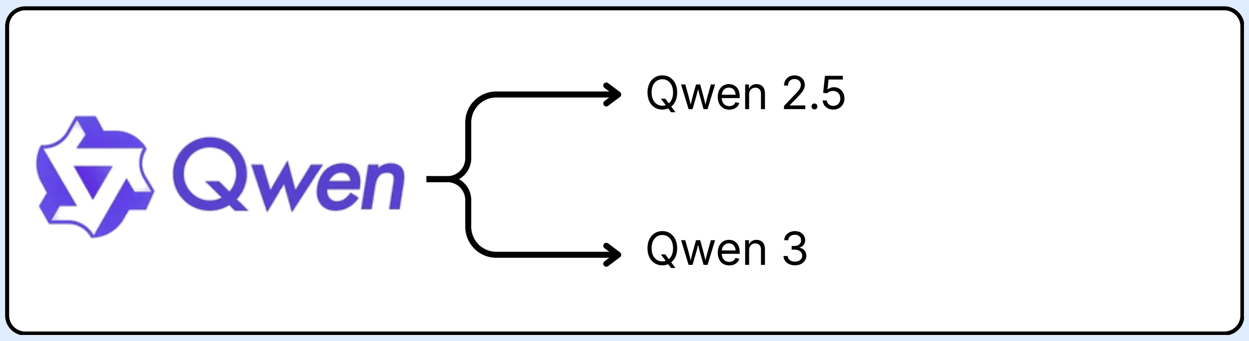
deepseek
- DeepSeek R1
- DeepSeek- V3.2



Llama 4
- Llama 4 Scout
- Llama 4 Maverick
- Llama 4 Behemoth Preview



- Gemma 3
- CodeGemma
- PaliGemma 2



Qwen
- Qwen 2.5
- Qwen 3

These all are **open-source LLMs**, which means their model code and often their weights are publicly available for anyone to use, modify, and build on.

What are Commerical LLMs?

Commercial LLMs are large language models that are built, owned, and offered by companies as **paid products** or services.

They're trained on massive datasets and run on expensive infrastructure, so companies package them into APIs, apps, or platforms and charge for access.

Some well-known commercial LLMs include:

- OpenAI models like GPT-4 and GPT-5 (used in ChatGPT and APIs)
- Google's Gemini models
- Anthropic's Claude models
- Microsoft Copilot models
- Meta Platforms's enterprise AI offerings

These are offered through subscriptions, usage-based pricing, or enterprise licenses.

What are Open-Source LLMs?

Open-source LLMs are large language models whose code and usually their trained weights are publicly available, so anyone can download, inspect, modify, and run them.

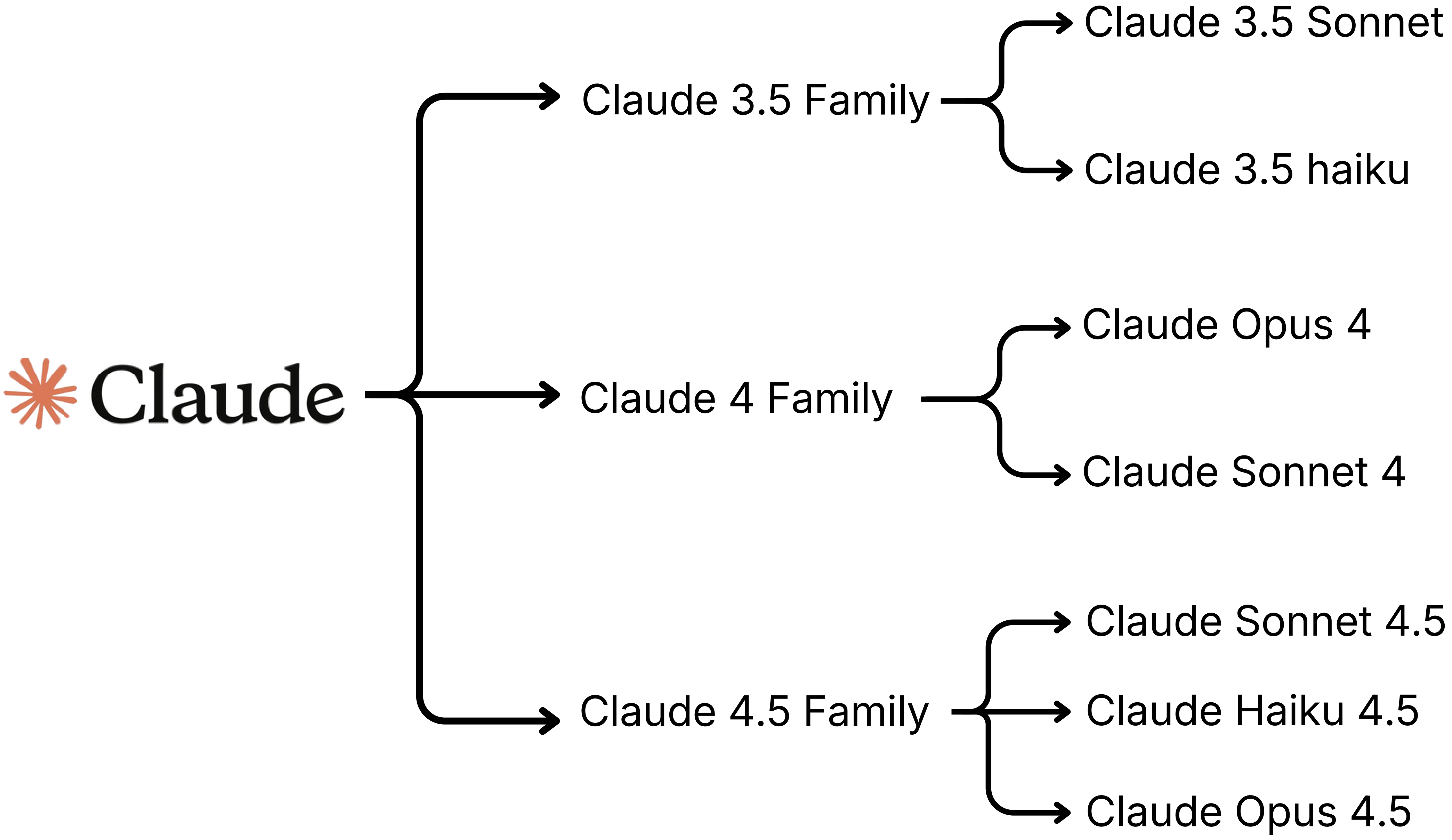
Instead of paying a company to access a model through an API, you can host an open-source LLM yourself or build on top of it.

Some well-known open-source or open-weight LLMs include:

- LLaMA by Meta Platforms
- Mistral by Mistral AI
- Falcon by Technology Innovation Institute
- Models shared through Hugging Face

These models are often released under licenses that allow research and, in many cases, commercial use.

Claude: A Commercial LLM



Claude

Claude 3.5 Family

Claude 3.5 Sonnet

Claude 3.5 haiku

Claude 4 Family

Claude Opus 4

Claude Sonnet 4

Claude 4.5 Family

Claude Sonnet 4.5

Claude Haiku 4.5

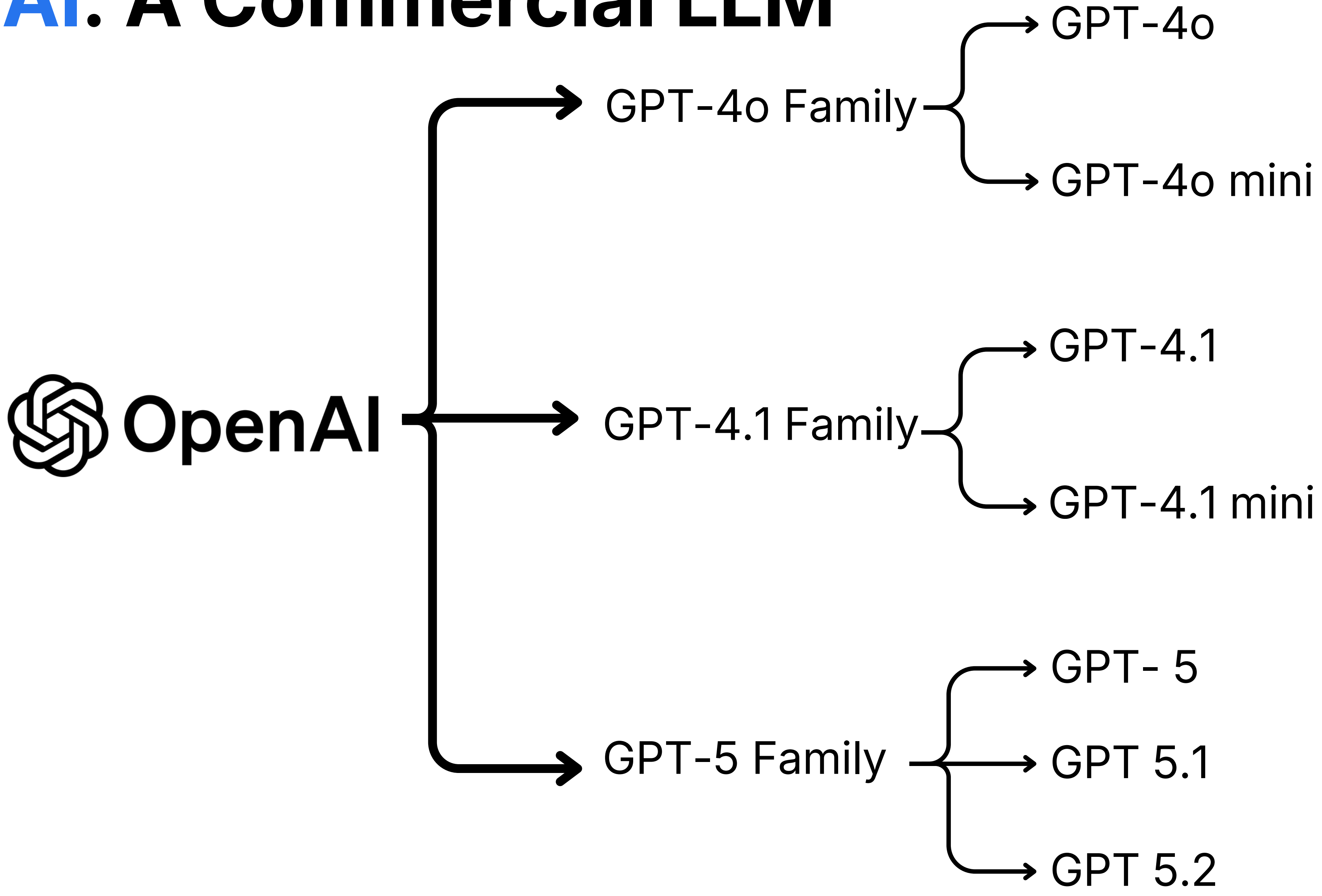
Claude Opus 4.5

Claude 3.5 Family: Claude 3.5 Sonnet set new benchmarks, surpassing the previous Claude 3 Opus in speed and intelligence, and **Claude 3.5 Haiku** is a **fast**, efficient model for quick, natural interactions.

Claude 4 Family: Claude Opus 4 is a flagship-level model focused on advanced reasoning, coding, and agentic tasks, while Claude Sonnet 4 is a slightly lighter but powerful variant within the Claude 4 family, optimized for broader use cases.

Claude 4.5 Family: Claude Sonnet 4.5 is an upgraded Claude 4 model with stronger agentic and coding capabilities, Claude Haiku 4.5 is a faster, cost-efficient model in the 4.5 series, and Claude Opus 4.5 is the most advanced Claude model, with top-tier reasoning and coding capabilities.

OpenAI: A Commercial LLM



OpenAI

GPT-4o Family

GPT-4o

GPT-4o mini

GPT-4.1 Family

GPT-4.1

GPT-4.1 mini

GPT-5 Family

GPT- 5

GPT 5.1

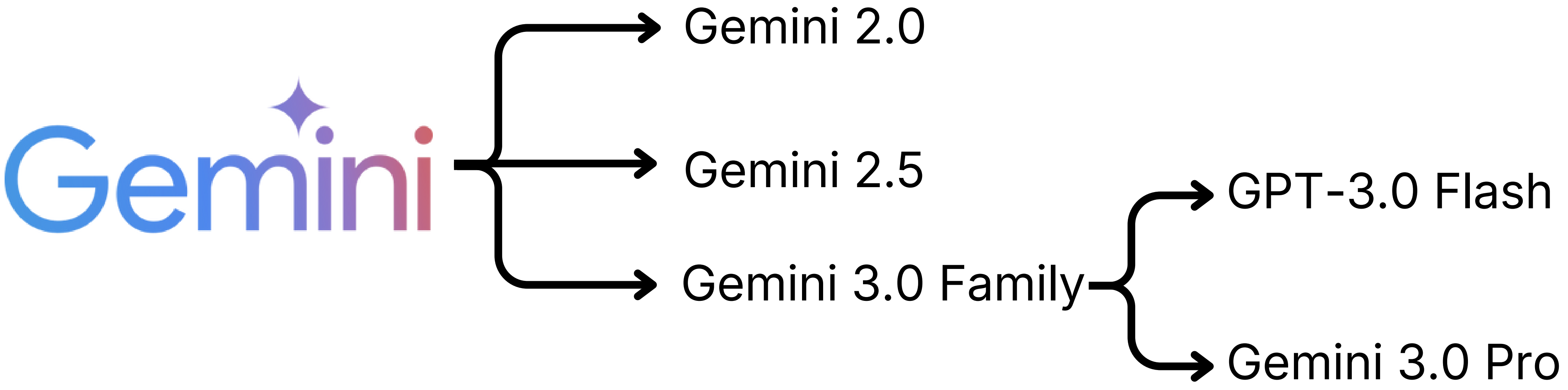
GPT 5.2

GPT-4o Family: GPT-4o is a multimodal model for text, vision, and audio, while GPT-4o mini is a lighter, cheaper version optimized for fast, everyday use.

GPT-4.1 Family: GPT-4.1 improves reasoning, instruction following, and coding, and GPT-4.1 mini is a smaller, faster option that balances cost and performance.

GPT-5 Family: GPT-5 is a flagship model for advanced reasoning and long-context tasks, GPT-5.1 refines quality and consistency, and GPT-5.2 is the most powerful version with the strongest reasoning performance.

Gemini: A Commercial LLM



Gemini 2.0

Gemini 2.5

Gemini 3.0 Family

GPT-3.0 Flash

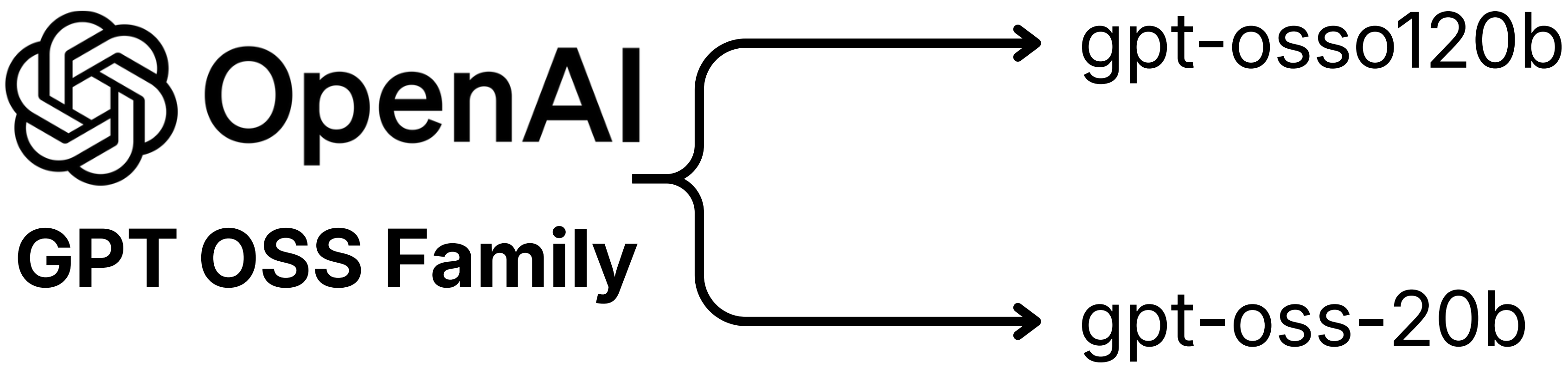
Gemini 3.0 Pro

Gemini 2.0 marks Google's shift to a multimodal-first Gemini generation, bringing strong improvements over Gemini 1.x with better reasoning, instruction following, and cross-modal understanding.

Gemini 2.5 enhances reasoning depth and analytical performance, improves results in coding and complex problem solving, and is designed for reliability, long-context tasks, and structured reasoning, making it suitable for demanding production workloads.

The Gemini 3.0 family includes Gemini 3.0 Flash, a fast, low-latency model built for real-time scale and responsiveness rather than deep reasoning, and Gemini 3.0 Pro, the flagship Gemini 3.0 model designed for deep reasoning and enterprise-level tasks.

The GPT-OSS family: Open Weight Models



OpenAI

GPT OSS Family

gpt-osso120b

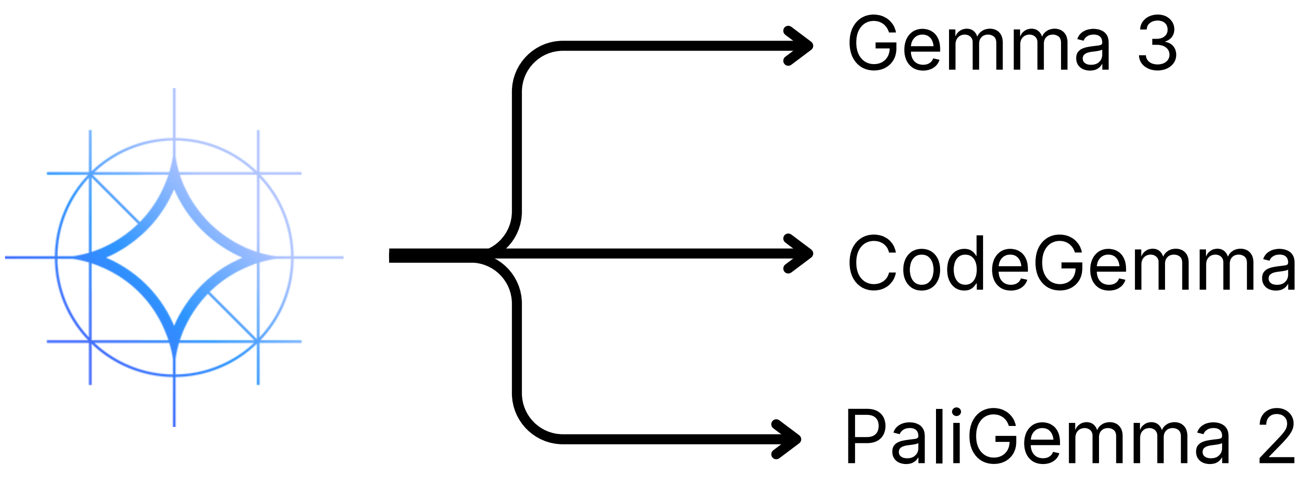
gpt-oss-20b

The GPT-OSS family is OpenAI's open-weight reasoning model lineup released under the Apache 2.0 license, enabling developers to run, fine-tune, and deploy GPT-class models on their own infrastructure.

The gpt-oss-120b is the flagship model, offering near-parity reasoning performance with OpenAI's o4-mini while being optimized to run on a single 80 GB GPU.

The gpt-oss-20b is a smaller, deployment-friendly variant designed to run on systems with around 16 GB of memory, targeting local inference, edge use cases, and rapid experimentation, and delivering reasoning performance comparable to o3-mini at a fraction of the infrastructure cost.

Google Gemma: Open Source LLMs



Gemma 3

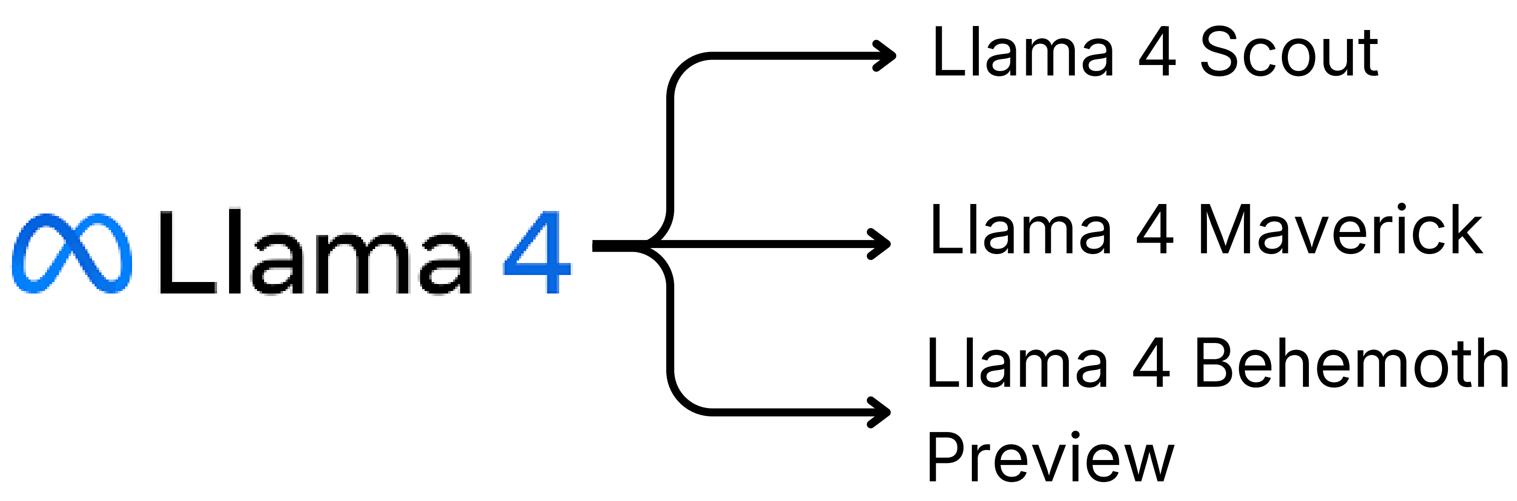
CodeGemma

PaliGemma 2

The Gemma family is a series of lightweight, open-weight language models developed by Google DeepMind, built using the same foundational research and technology that powers Google’s proprietary Gemini models, and designed to be easy to run, efficient to deploy, and broadly accessible for developers, researchers, and enterprises.

Gemma 3 can solve a wide variety of generative AI tasks with text and image input, supports over 140 languages, and offers a long 128K context window. CodeGemma completes programming tasks with a lightweight, coding-focused generative model. PaliGemma 2 builds visual data processing AI solutions with a model designed to be fine-tuned for image data processing applications and is available in multiple resolutions.

Meta Llama 4: Open Source LLMs



Llama 4

Llama 4 Scout

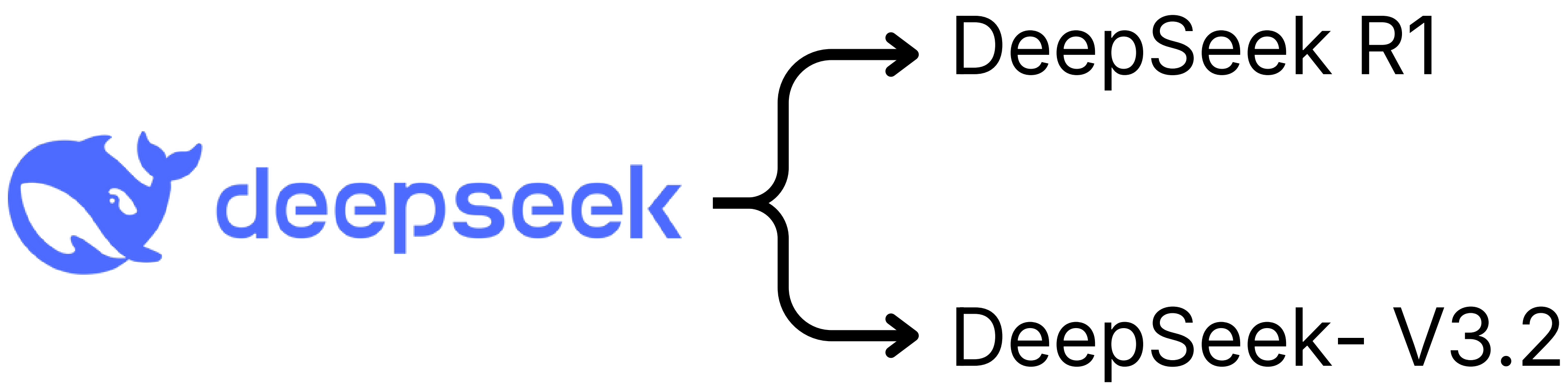
Llama 4 Maverick

Llama 4 Behemoth Preview

Llama 4 models are designed with native multimodality, leveraging early fusion that allows the model to be pre-trained with large amounts of unlabeled text and vision tokens. Llama 4 Scout is a model that offers superior text and visual intelligence, single H100 GPU efficiency, and a 10M context window for seamless long document analysis.

Llama 4 Maverick is a multimodal model for image and text understanding with groundbreaking intelligence and fast responses at a low cost. Llama 4 Behemoth Preview is an early preview, still in training, of the Llama 4 teacher model used to distill Llama 4 Scout and Llama 4 Maverick.

DeepSeek: Open Source LLMs



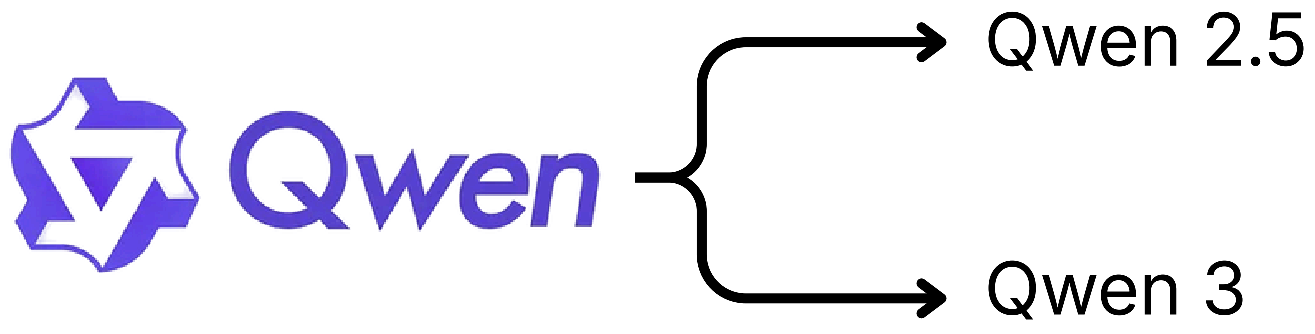
DeepSeek R1

DeepSeek- V3.2

The DeepSeek family is a set of open-source large language models developed by the Chinese AI company DeepSeek, designed to advance reasoning, coding, and general LLM capabilities while remaining fully accessible for modification, deployment, and commercial use under permissive licensing.

DeepSeek R1 represents the family's reasoning-first generation, developed using reinforcement learning techniques that enable structured problem solving and improved reasoning depth. DeepSeek-V3.2 is a later family iteration that integrates agent-style thinking directly into tool use, supporting complex instruction execution across environments and offering smoother API experiences.

Qwen: Open Source LLMs



The Qwen family is a broad, open-source series of large language models developed by Alibaba Cloud, designed to support text, multimodal, and reasoning applications under a permissive Apache 2.0 license.

Qwen 2.5 represents a mid-generation advancement that improves multilingual understanding, reasoning, and multimodal processing over earlier Qwen releases.

Qwen3 is the latest and most advanced open-source generation, offering significant improvements in reasoning, multilingual coverage, and context handling, and the Qwen3 family includes a range of dense models from 0.6B to 32B parameters.