LLMs

1. Tokenization layer
2. Embedding Layer
3. Transformers Block

* Rotary positional Embedding (RoPE)
* KV cache (During Inference)
* LayerScale, DeepNorm, RMSNorm
* LORA / Adapters / PEFT modules
* Mixture of Expert (MoE

Tokenization:

My name is Bappy. I am a good boy.

1. word Level Tokenization
2. sentence Level Tokenization

word_token = ["My", "name", "is", "Bappy"]
["I", "am", "a", "good", "boy"]

sentence_token = ["My name is Bappy",
"I am a good boy"]

Open Source LLMs

Definition:

Models that are publicly released with access to **weights, code, and training details** — enabling users to **inspect, modify, fine-tune**, and deploy locally or on their infrastructure.

Popular Open Source Models:

Model	Creator	Context Length
LLaMA 2 / LLaMA 3	Meta	4K–32K
Mistral / Mixtral	Mistral AI	8K–32K
Falcon	TII (UAE)	2K–4K
BLOOM	BigScience	2K
Phi-2 / Phi-3	Microsoft	2K–4K
Qwen	Alibaba	8K+

Pros:

- **Transparent** architecture and weights
- **Customizable** (fine-tuning, quantization, PEFT, etc.)
- **Runs locally** (better for privacy & control)
- **No vendor lock-in**
- **Great for research and innovation**

Cons:

- May require **infrastructure & engineering expertise**
- Often **smaller-scale** or **less optimized** than proprietary LLMs
- Security & performance tuning is **your responsibility**

✅ **Pros:**

- **State-of-the-art performance**
- Scalable, reliable APIs
- Fully **managed infrastructure**
- Access to **massive context windows** (100K+ tokens)
- Ideal for **enterprise use cases** and speed to deploy

❌ **Cons:**

- ❌ No access to weights or training data
- ❌ Expensive (pay-per-token)
- ❌ Risk of **vendor lock-in**
- ❌ Limited customization/fine-tuning options
- ❌ Potential **privacy/compliance concerns** (data leaves your infra)

Feature / Model	LLaMA (Meta)	GPT (OpenAI)	Claude (Anthropic)	Mistral	Gemini (Google DeepMind)
🧠 Latest Version	LLaMA 3 (2024)	GPT-4 Turbo (2024)	Claude 3 (Opus, Sonnet, Haiku)	Mixtral 8x22B / Mistral 7B	Gemini 1.5 Pro / Flash (2024)
🧱 Model Type	Decoder-only Transformer	Decoder-only Transformer	Constitutional AI + Transformer	MoE (Mixture of Experts) + Dense	Multimodal Transformer (text/image/audio/code)
🌐 Open Source	✅ Yes (LLaMA 2/3)	❌ No	❌ No	✅ Yes	❌ No
📦 Sizes Available	8B, 70B (LLaMA 2/3)	GPT-3.5, GPT-4 (unknown size)	Haiku (small) → Opus (largest)	7B, 12.9B, 45B (sparse)	Not disclosed
🧠 Context Window	8K–32K (LLaMA 3)	4K–128K (GPT-4 Turbo)	Up to 200K tokens	8K–32K	Up to **1M tokens** (Gemini 1.5 Pro)
🔧 Fine-tuning	✅ Supported (via PEFT, LoRA)	✅ Proprietary fine-tuning only	❌ Not yet	✅ Easy via open tools	❌ No
🛠️ Tooling	HuggingFace, Transformers	OpenAI API	Claude API	HuggingFace, Transformers	Gemini API (via Google AI Studio)
🔒 Local Inference	✅ Yes	❌ No	❌ No	✅ Yes	❌ No
🧠 Multimodal	❌ Text only (as of LLaMA 3)	✅ GPT-4V (vision support)	✅ Claude 3 (image + charts)	❌ Text only	✅ Image, Audio, Code, Text
🔥 Unique Strength	Best OSS for fine-tuning	Industry-leading performance & tools	Best at long reasoning	Blazing fast OSS MoE models	Unmatched context + multimodality
💰 Pricing	Free (OSS)	Paid (OpenAI API)	Paid (Anthropic API)	Free (OSS)	Paid (via Google Cloud & Studio)

Feature	Pretraining	Fine-tuning	Instruction-tuning
✅ Goal	Learn general knowledge	Specialize for a use case	Learn to follow human instructions
Data	Unlabeled, massive	Domain-specific	Instruction/response pairs
Supervision	Self-supervised	Supervised or unsupervised	Supervised
Customizable	❌ Not specific	✅ Task/domain specific	✅ Dialogue/following instructions
Output Behavior	Generic, knowledge-rich	Narrow/specialized	Conversational, task-aware