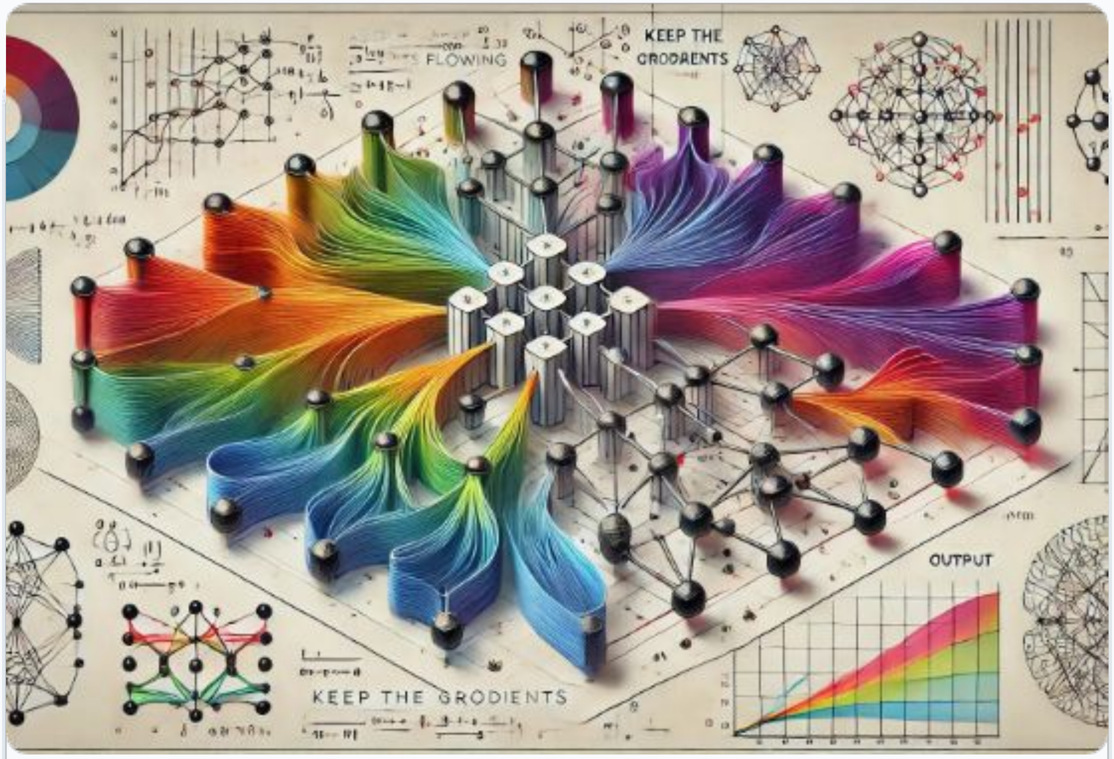
CLOSED MODELS
OPEN WEIGHT MODELS
OPEN SOURCE MODELS

WHAT IS AN AI "MODEL"?

For today’s topic, a model is a **Large Language Model (LLM)**: a neural network trained to predict the next token in a sequence.

We will discuss and compare three characteristics of models when it comes to how “open” they are:

- **The Recipe:** The training code, architecture and datasets used to calculate the weights.
- **Parameters (Weights):** Numerical values defining how data is processed. These are the "learned intelligence."
- **Inference:** Executing the weights on a GPU to generate a result based on user input.



01. CLOSED (PROPRIETARY) MODELS

The Vendor Lock-In Reality

Proprietary models are managed entirely by providers. You are fully dependent on their infra, alignment, and internal updates.

- **Total Dependency:** Every feature relies on the provider's uptime, pricing, and API accessibility.
- **Silent "Nerfing":** Models are updated "behind-the-scenes." A prompt that works today might fail tomorrow due to mid-cycle weight changes.
- **Opaque Guardrails:** Mandatory safety filters can change without warning, breaking specific business or logic use-cases.

RECIPE	SECRET
WEIGHTS	OPAQUE
INFERENCE	MANAGED

02. OPEN WEIGHT MODELS

RECIPE	SECRET
WEIGHTS	SHARED
INFERENCE	LOCAL

Immutable Sovereignty

Models where the final weights are shared. Once you download the file, it is yours to control and run indefinitely.

- **Version Immutability:** The model never changes unless YOU update it. No silent nerfing or mid-project prompt drift.
- **Local Execution:** Deploy air-gapped on private hardware for maximum data sovereignty and privacy.
- **Full Control:** Ability to apply custom fine-tuning (LoRA), quantization (GGUF/EXL2), and in some cases even remove guardrails.

03. OPEN SOURCE MODELS

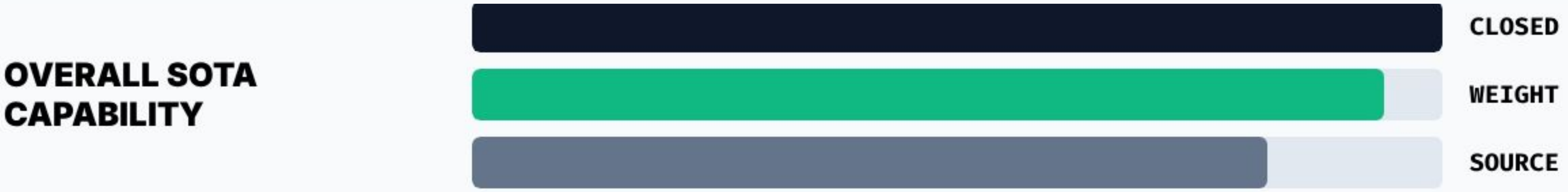
The Glass Box Paradigm

Total transparency. Providers release the weights, the full training datasets, and the code used to train them.

- **Scientific Rigor:** Public datasets allow for auditable research on how models actually learn and generalize.
- **Auditability:** Scrutinize pre-training data mixtures for bias, copyright, and safety training effectiveness.
- **Total Lineage:** Absolute control over the model's history. No proprietary "secret sauce" in the recipe.

RECIPE	PUBLIC
WEIGHTS	SHARED
INFERENCE	LOCAL

Performance comparison



Frontier Closed Models still hold the absolute "bleeding edge" in multimodal integration (native video/audio reasoning) and complex agentic reliability.

Open-Weight Models often reach parity with these closed models in coding and mathematics within one quarter of their release.

Open Source Models are generally about **9–12 months behind**, as the massive compute and data cleaning required for frontier-level training are still largely gated by corporate resources.

// BENCHMARK SOURCES & METHODOLOGY

LMSYS Arena: Human-preference Elo ratings for GPT-5.2 and Claude 4.5 Opus.

MMLU-Pro / GPQA: Graduate-level reasoning and STEM task understanding scores.

SWE-bench Verified: Real-world software engineering issue resolution benchmarks.

OPENAI

GPT-5.2

> Unified Reasoning: Default CoT logic
> 400,000 Token Context Window
> Agentic Orchestration: 50+ sub-tasks

ANTHROPIC

Claude 4.5 Opus

> Effort Parameter: Reflection toggle
> 64,000 Token single-pass output limit
> Vision-based UI interaction ("Zoom")

CLOSED FRONTIER

GOOGLE

Gemini 3 Pro

> Native Temporal Video/Audio Processing
> 2M - 10M Context Window (Ultra)
> Deep Workspace & Disco integration

OPENAI

o1-pro

> Reasoning-time compute scaling
> RL-trained internal thinking cycles
> Hidden CoT for security & logic

OPEN WEIGHT LEADERS

META

Llama 4 Behemoth

- 405B+ Params; 15T token pre-training
- Native Multi-Token Prediction (MTP)
- Scout variant: 10M context capability

OPENAI

gpt-oss-120b

- MXFP4 Quantization: Single 80GB GPU
- 117B MoE (5.1B active per token)
- Optimized for agentic tool-use calls

DEEPSEEK

DeepSeek-R1

- MLA Architecture (671B MoE)
- RL-based logic thinking steps (CoT)
- MIT License (Unrestricted)

MINIMAX

MiniMax M2.1

- Interleaved Thinking Verification
- iOS/Android native UI optimization
- 204,000 Token local context window

MOONSHOT AI

KIMI 2.5

- Multimodal (image and video)
- 15T tokens training
- Performance rivalling closed models

AI2

OLMo 2

- Full Dolma v3 Dataset Disclosure
- 500+ intermediate checkpoints public
- Scientific glass-box architecture

TRULY OPEN SOURCE

DATACOMP-LM

DCLM-7B v2

- Dataset: 240T token curated pool
- Public filtering & quality-scoring scripts
- Maximum parameter efficiency (SOTA)

ZHIPU AI / THUDM

GLM-4.7-9B

- Agentic Terminal: Native self-correction
- 200K long-context code window
- SFT & RLHF strategies disclosed

```
$ cat architecture_summary.csv
```

METRIC	CLOSED	OPEN WEIGHT	OPEN SOURCE
Weights Access	Locked (API)	Public (Local)	Public (Local)
Training Recipe	Secret	Secret	Public & Auditable
Model Stability	Variable (Nerfing)	Fixed (Immutable)	Fixed (Immutable)
Strategic Edge	Frontier Scaling	Sovereignty	Scientific Rigor

QUESTIONS