Keiji AI

TUTORIAL · OPEN-WEIGHT MODEL TRAINING AND EVALUATION

How LLM Training Actually Works

You don’t need the biggest model. You need the right data.

Keiji AI Inc · keiji.ai · jimeng@keiji.ai

Jimeng Sun · August 2026

WHAT THIS COVERS

Three questions to ask of any model release

You will be handed numbers like “2.8 trillion parameters,” “104B activated,” “1M context,” “MoE,” “MXFP4.” Vendors use them to mean whatever they need them to mean.

Can we run it?

Total parameters × bytes, plus KV cache. Ask what precision the quote assumes.

May we use it?

Revenue triggers, attribution at scale, field-of-use and jurisdictional limits.

Is it good enough?

On *your* task. A headline benchmark on software engineering tells you nothing.

PART 0

Two cases where the small model won

Smaller models, trained on data the organization already owned, outperformed every frontier model they were tested against — at roughly an order of magnitude less cost. Neither required a new algorithm.

CASE 1 · FINANCIAL JUDGMENT

Prompting plateaued. Training did not.

AVERAGE ACCURACY ACROSS SIX INFORMATION-FILTERING TASKS

29.8%
fewer mistakes than the best frontier model

13.8×
cheaper per task

80%
was the threshold prompting could not cross

CASE 1 · WHAT ACTUALLY DECIDED IT

Label quality decided the outcome

THE PROBLEM

The first training set came from commercial vendors using **non-expert labelers** . Models trained on it stayed poor — not because the training was wrong, but because the labels were.

THE FIX

Train on the cheap labels, then route **only the examples the model disagrees with** to experts. If a model can’t reproduce its own training example, the case is hard or the label is bad.

Their tasks were ones where experts *have* judgment but cannot state the rule — the shape of adverse-event triage, eligibility screening and endpoint adjudication. A prompt carries only the part of an intuition you can put into words. Training carries the rest.

CASE 2 · CLINICAL TRIAL REASONING

8B beat 70B on all eight tasks

We built **TrialPanorama** first — 1.6M trial records from 15 registries, 152,000 samples across eight systematic-review and trial-design tasks. Only then did we train against it.

SPECIALIZED, TRAINED ON TRIALPANORAMA

8B

~4.5 GB at 4-bit · single workstation GPU · over 100 tokens/second

Wins 8 of 8 tasks

GENERIC, LARGER

70B

~35 GB at 4-bit · multi-GPU · and a 2.8T frontier model is a rack, not a server closet

Loses 8 of 8 tasks

PART 0 · THE PATTERN

What the two cases share

The work was judgment-heavy and unwritten

Experts performed it easily and could not specify it. That gap is where a prompt fails and training succeeds.

The data came first, the model second

Expert labels in one case, a published benchmark in the other. In neither was the model the asset.

The winner was small enough to host

Both students run on hardware you can own, inside your data boundary, with pinnable weights.

Frontier progress did not close the gap

Successive releases improved little on these tasks, and cost more. Waiting was not a strategy.

> The finance team called it *differentiated intelligence*. **The frontier model was never the asset — the labeled judgment was.**

PART 1

The two numbers that decide everything

A modern large model has two different parameter counts. They mean completely different things, and a vendor quoting only one is hoping you don’t ask.

PART 1 · DEFINITIONS

Total parameters and activated parameters

TOTAL PARAMETERS

How much model you have to store

Drives memory and hardware cost.

ACTIVATED PARAMETERS

How much model runs on a token

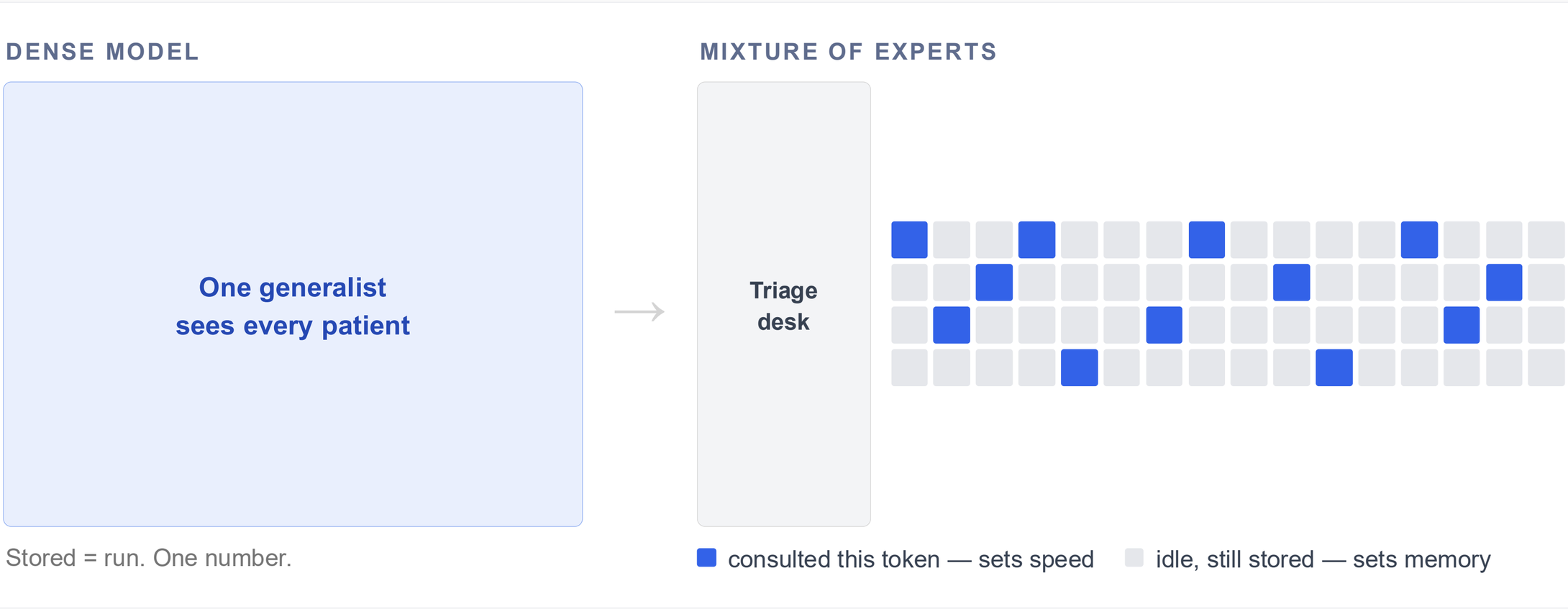
Drives compute and speed.

You must have enough memory for the total, but you only get the speed of the activated.

PART 1 · ARCHITECTURE

Mixture of Experts, in one picture

A hospital with 896 specialists and a triage desk. A question arrives, 16 specialists handle it, 880 sit idle — and you still employ all 896.



PART 1 · THE EVIDENCE

The gap between stored and run

MODEL	TOTAL	ACTIVATED	EXPERTS
Kimi K3	2.78 trillion	104.2 billion	16 of 896 routed
DeepSeek-V4-Pro	1.6 trillion	49 billion	—
GLM-5	744 billion	40 billion	256 total
DeepSeek-V4-Flash	284 billion	13 billion	—

Kimi K3 is genuinely open — and for almost every hospital system and mid-size pharma, unrunnable. **Open weights and “we can run it” have become different statements.**

PART 2

The four stages of training

Understanding what each stage does tells you what you can and cannot change about a model you download.

PART 2 · THE PIPELINE

Four stages, two of them yours

STAGE 1

Pre-training

Predict the next word over trillions of tokens. All knowledge arrives here.

10^8–10^9 · not you

STAGE 2

Mid-training

Stretch the context window in stages: 8K → 64K → 256K → 1M.

lab only

STAGE 3

Supervised fine-tuning

Instruction → good response. Turns a completion engine into an assistant.

10^4–10^5 · you can do this

STAGE 4

Reinforcement learning

Reward verifiable outcomes. Learning to be right, not just plausible.

needs a verifier

Knowledge is installed in stage 1 — and essentially nowhere else.

Behaviour, format and discipline are shaped in stages 3–4. This is your window.

PART 2 · STAGE 1

Knowledge arrives in pre-training

GLM-5 read 28.5 trillion tokens and out of pure next-word prediction acquired grammar, facts, reasoning, translation and code. Nobody fully understands why this works as well as it does — it is an empirical result, not a designed one.

THE FINDING

Fine-tuning examples containing knowledge the model does not already hold are learned slowly — and once learned they raise the hallucination rate roughly linearly.

Gekhman et al., arXiv 2405.05904 (EMNLP 2024)

THE CONSEQUENCE

Fine-tuning transfers *form* reliably and *facts* badly. If a dosing fact wasn't in pre-training, later stages will not reliably install it.

Tune for form. Retrieve for facts.

PART 2 · STAGE 4

Reinforcement learning needs a verifier

DEEPSEEK-R1 · AIME 2024

15.6% → 77.9%

During RL alone, rewarding only verifiably-correct final answers. The model invented its own strategies, including self-checking.

Nature 645 · doi:10.1038/s41586-025-09422-z

RL works where answers are verifiable. Math has a right answer. Code either runs or it doesn’t. That is why every current model is strong at math and code, and comparatively weaker at open-ended judgment — those are the domains where the reward signal is clean.

THE READ

“Strong at agentic coding” does not transfer to “reliable at adverse-event extraction.” The reason is structural: one has a verifier and the other doesn’t.

PART 3

Distillation — how a small model gets good

The teacher does not have to be a model. It can be a stronger model, a group of human experts, or both.

PART 3 · THE BASELINE

The behaviour-cloning recipe

1

Collect tasks

Representative examples from the workflow you want to automate.

2

Create demonstrations

A stronger model or human experts produce high-quality answers.

3

Review the outputs

Correct errors, resolve disagreements, remove unsafe examples.

4

Train the student

Fine-tune a smaller model to reproduce the demonstrated behaviour.

5

Evaluate

Test on a held-out, independently adjudicated evaluation set.

Steps 2 and 3 are the expensive ones — not step 4. A stronger model supplies scale; experts define and control the standard, spending their attention on the cases where judgment matters.

PART 3 · THE REFINEMENT

On-policy distillation

OFF-POLICY · THE OLD WAY

Teacher writes 100,000 worked solutions → Student studies them → Student's first own mistake — territory the textbook never covered

Exposure bias. Like learning surgery entirely from recordings of a surgeon who never slips.

ON-POLICY · THE NEW WAY

Student attempts the problem itself →  teacher grades every token the student produced → Corrected on the states it actually reaches, not the ones the teacher visited

Dense feedback on your own mistakes rather than sparse feedback on someone else's successes — an attending physician standing behind you.

Now standard post-training at Alibaba, DeepSeek, Z.ai, Xiaomi and NVIDIA. It requires compatible thinking patterns and a teacher with something genuinely new to offer — otherwise start with an off-policy cold start.

PART 3 · THE LIMIT

Distillation raises floors, not ceilings

THE EXPERIMENT

Students from 1.5B to 13B trained on up to 150M tokens of ChatGPT output. Human raters judged them roughly competitive with the teacher.

THE RESULT

Targeted benchmarks showed they had closed almost none of the real gap. Imitation transfers style more than capability.

> If your evaluation of a distilled model is **“our reviewers liked the outputs,”** you have reproduced this experiment and will get its result.

Gudibande et al., The False Promise of Imitating Proprietary LLMs — arXiv 2305.15717

PART 4

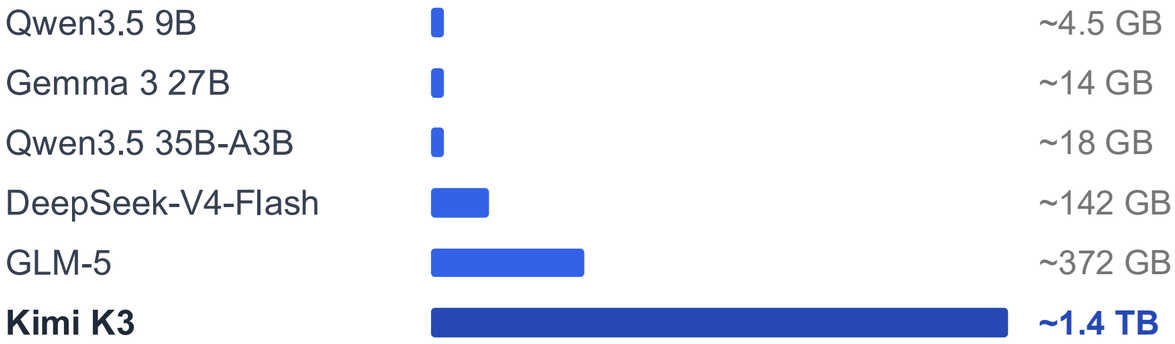
What it takes to run one

The models in the news are the trillion-parameter ones. The models you would actually deploy are mostly not.

PART 4 · ARITHMETIC

Two independent dials

MEMORY FOLLOWS TOTAL PARAMETERS



At 4-bit, ≈0.5 GB per billion parameters. Bars are proportional to Kimi K3; the three smallest are floored so they stay visible. Then add the KV cache — it scales with context and concurrency, and at long context routinely exceeds the weights.

SPEED FOLLOWS ACTIVATED PARAMETERS

```
tokens/second ≈ memory bandwidth ÷
(activated params × bytes per param)
```

Generating a token is memory-bandwidth-bound, not compute-bound: the GPU must read the active weights out of memory. Which is why MoE is fast — only the activated slice gets read.

Kimi K3 needs roughly **eighteen H100s just to hold it** at 4-bit, before any KV cache.

PART 4 · DEPLOYMENT

Five deployment tiers

TIER	REPRESENTATIVE MODELS	TOTAL	ACTIVATED
Edge	Gemma 3 1B / 4B · Qwen3.5 0.8B–4B · Ministral 3B	1–4B	same
Workstation	Qwen3.5 9B · Gemma 3 12B / 27B · Mistral Small 4	7–32B	same
Single-node MoE	Qwen3.5 35B-A3B · Mistral Medium 3.5	30–130B	**3–13B**
Multi-GPU	DeepSeek-V4-Flash · GLM-5 / GLM-5.2	284–744B	13–40B
Cluster only	DeepSeek-V4-Pro · Kimi K3 · Qwen3.8-2.4T	1.6–2.8T	49–104B

Qwen3.5 35B-A3B holds 35 billion parameters and activates 3 billion — workstation memory at the speed of something five times smaller. **MoE is not only a frontier-lab technique.**

PART 4 · SEQUENCING

Don’t self-host first

STEP 1 · EVALUATE

The labs’ own APIs

DeepSeek, Moonshot and Z.ai serve their models directly. Cheapest honest evaluation.

STEP 2 · PRODUCTION

Serving providers

Together, Fireworks, Baseten, Groq, Cerebras. Dedicated instances restore version pinning.

STEP 3 · ONLY IF FORCED

Self-host

When volume, residency or validation demands it — not before.

A router is not one endpoint. Different providers at different quantizations on different days — disqualifying for anything you must reproduce unless you pin both.

The data boundary comes back. Open weights served by someone else is a licensing win, not a data residency win.

Ask what quantization is served. A model that fails your evaluation may be a 4-bit copy of one that would have passed.

PART 5 · LICENSING

Three licence patterns

CLEAN

MIT / Apache 2.0

DeepSeek V4 · GLM-5 · open Qwen · Gemma 4 · Muse Glimmer. No thresholds, no attribution. Apache adds a patent grant.

BESPOKE “OPEN”

Vendor-drafted

Kimi K3 · Nemotron 3. Broad rights, own terms: K3’s $20M service-revenue threshold becomes a negotiation, plus UI branding at scale.

READ CAREFULLY

Community licences

Older Llama · Gemma 3. Restrictions are field-of-use and jurisdictional — Llama 4’s multimodal carve-out excludes EU-headquartered companies.

Terms follow the **checkpoint** , not the brand — Gemma 4 and Gemma 3 differ, and “Qwen” spans Apache 2.0 and fully proprietary. And **a fine-tune inherits its base model’s licence.**

PART 6 · WHERE TO SPEND

What depreciates, what appreciates

↓ DEPRECIATES

Fine-tuned weights — reproducible from a public base and a public recipe

Generic training corpus — literature is public, de-identified records increasingly commodity

↑ APPRECIATES

Adjudicated gold standards — clinician time, disagreement resolved, years to accumulate

Evaluation harness — encodes your definition of correct; gates the treadmill

Verification layer — deterministic, hardest to copy because it’s unglamorous

Validation package — per-configuration, auditable, slow to produce

A competitor with the same open base and a good engineer reproduces your fine-tune in weeks. Nothing about the weights is defensible. **What surrounds them is.**

CLOSING

Six conclusions I would defend

01 **“Open source caught up” is two true clauses joined by a false therefore.** Open models closed much of the capability gap. Open models you can host have not.

02 **Licence is a gating criterion, not a footnote.** Check it before the technical evaluation, and read the LICENSE file rather than the announcement.

03 **Capability is task-shaped.** Near-parity on extraction, structuring and retrieval-grounded work; a real gap on long-horizon agentic planning.

04 **Distillation is the right tool, with a known failure mode.** It won’t install knowledge the base lacks. Evaluate against benchmarks, never reviewer preference.

05 **Frozen weights are a scientific requirement before a security one.** A trial runs five to ten years; frontier models are deprecated in twelve to eighteen months.

06 **Specialized beats generic on narrow work, and it beats it cheaply.** The question is not which frontier model to license. It is which of your judgments is worth labeling.

Keiji AI

You don’t need the biggest model. You need the right data.

The most valuable AI asset is already inside your walls. The organizations that win will be the ones that start compounding it now.

Jimeng Sun · jimeng@keiji.ai

keiji.ai/research/benchmarks