Inkling 975B vs Your 3090: The Real Memory Math (2026)

July 16, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Open weights are winning. The winners no longer fit on your hardware. Inkling landed at 975B parameters and Apache 2.0, genuinely free to download, and its smallest 1-bit quant is 270GB, which means a maxed-out consumer desktop with four 64GB sticks comes up short. Not short on VRAM. Short on DIMM slots. The GPU was never the thing standing in your way. But the panic version of this story is wrong: the frontier tier left, the tier you actually run didn't move an inch. Qwen 3.6 still fits your 3090 with room for a quarter-million tokens of context. This is the memory math on every open model that made headlines this month, where each one's real entry point sits, and why the headline releases stopped being aimed at you.

📚 **Related:** How to Run GLM 5.2 Locally · Best Way to Run Qwen 3.6 35B MoE Locally · China Made Open Source a Strategy · MoE Models Explained

Thinking Machines dropped Inkling on July 15. Apache 2.0, full weights on Hugging Face, no gate, no waitlist, no acceptable-use rider. By every definition the local AI crowd has been asking for since 2023, this is the good outcome. Open won.

Then you look at the repo. The BF16 checkpoint is **1.9 terabytes**.

That’s the shape of 2026. The open frontier keeps delivering exactly what we asked for, and the models keep arriving in sizes that have quietly stopped being about us. Inkling at 975B. GLM 5.2 at 753B. Kimi K3, which I’ll get to, because what Moonshot did this week is the most interesting part of the story and almost nobody has framed it right.

None of them run on a 3090. Not one.

So let’s do the math nobody else is doing. Every site on the internet is publishing “Inkling is here!” today. Here’s what 975B actually costs you in memory, and where the real entry point sits for each of these models. Then the part that keeps this honest: why your position tonight is exactly as good as it was last week.

Image: Memory footprint of Inkling, GLM 5.2, and Qwen 3.6 at their smallest usable quants, against a 24GB RTX 3090 and a 256GB consumer desktop ceiling

What actually shipped

Inkling is a Mixture-of-Experts transformer: **975B total parameters, 41B active per token**, 66 layers, 6 of 256 routed experts plus 2 shared on each token. 1M context. Pretrained on 45 trillion tokens of text, images, audio and video. It takes text, images and audio in, and emits text only. Video was a pretraining diet, not an input you can use. All of that is straight from the model card.

One wrinkle worth knowing before you quote the headline number: Hugging Face's own metadata for the repo counts **952B** parameters, not 975B. TML says 975B. The gap is probably an embedding or expert-counting convention, but the two primary sources disagree, so don't cite 975B as though HF corroborates it.

The benchmark story is where it gets interesting, and where the press releases stop being useful. Artificial Analysis put Inkling at **41** on its Intelligence Index and called it the leading open-weights release from a U.S. lab. True. That "U.S." is carrying a freight train.

Model	AA Intelligence Index	Origin
GLM 5.2	**51**	Z.ai (CN)
Kimi K2.6	**44**	Moonshot (CN)
Inkling	**41**	Thinking Machines (US)
DeepSeek V4 Flash	**40**	DeepSeek (CN)
Nemotron 3 Ultra	38	NVIDIA (US) — prior US leader

Inkling is the best open model America has produced. It is fourth on this list. On Terminal Bench 2.1 it posts 63.8% against GLM 5.2's 82.7% in TML's own comparison table, a gap of 18.9 points against a model that's been out for a month. (Artificial Analysis independently scores GLM 5.2 at 78 on that benchmark rather than 82.7, which would make the gap about 14. Either way Inkling loses, and I'd rather show you both numbers than pick the flattering one.) Kimi K2.6 also beats it there, 71.3% to 63.8%, so "beats Kimi on agentic tasks" is true on some rows and false on others, depending which table you're selling.

AA's Omniscience eval clocks Inkling's **hallucination rate at 63%**, with 40% accuracy. It costs **$1.87 in / $4.68 out** per million tokens against GLM 5.2's $1.40/$4.40 and DeepSeek V4 Flash's $0.14/$0.28, and that Inkling price is currently running at a limited-time 50% discount. It is the most expensive model on the list, and the third-smartest.

To Thinking Machines' considerable credit, they say this out loud, in their own announcement:

> "Inkling is not the strongest overall model available today, open or closed."

I checked that quote against the raw page source rather than trusting a summary of it. They really wrote that. A lab shipping a 975B model and telling you in the launch post that it isn't the best is doing something rarer than the model.

The size reality

At BF16, Inkling's Hugging Face repo is **1,904.8 GB**. There's an NVFP4 checkpoint at 592 GB. Nobody outside a serving company touches either.

So you quantize, which is the whole reason a model this size is discussable at all. TML worked with Unsloth for llama.cpp support, and the GGUF repo went up day one. These are live file sizes, pulled fresh:

Quant	Size	What you're getting
UD-IQ1_S	**270.2 GB**	1-bit. The floor. Visibly dumber.
UD-IQ1_M	285.0 GB	1-bit.
UD-Q2_K_XL	317.3 GB	2-bit. Usable in a pinch.
UD-Q3_K_XL	432.8 GB	3-bit. Still feels like Inkling.
UD-Q4_K_XL	587.0 GB	4-bit. Near-indistinguishable.
Q8_0	856.8 GB	Effectively lossless, effectively pointless.
BF16	1,894.3 GB	The thing nobody runs.

Now the number that matters. Not the 1.9TB. That one's easy to dismiss as a datacenter problem and move on. The one that matters is **270.2 GB**, the smallest quant Unsloth ships. That's the floor — Inkling wearing every compromise available to it.

Where the real entry point sits

Here's the part I haven't seen anyone write.

A modern consumer desktop (AM5, four DIMM slots) tops out at 64GB per stick. Four sticks, 256GB. That is already an exotic and expensive configuration that almost nobody reading this owns.

256GB is 14 gigabytes short of the smallest Inkling quant that exists.

Read that again, because it reframes the whole thing. There is no GPU you can buy that fixes this. Your 3090 isn't the problem. A 5090 isn't the problem. Four 5090s aren't the problem either. The expert-offload trick that makes huge MoE models reachable, pinning the routed experts to system RAM with `-ot "exps=CPU"` while attention and the KV cache stay on the card, only turns "needs 270GB of VRAM" into "needs 270GB of something." And a consumer platform cannot physically hold 270GB of anything. You run out of DIMM slots before you run out of money.

Inkling doesn't need a better graphics card. It needs a different class of motherboard. Threadripper, Epyc, Xeon: eight or twelve DIMM slots, registered ECC memory. That's the entry ticket, and it's a platform decision, not a GPU upgrade.

What that costs right now, in a memory market that has roughly tripled since the shortage began: DDR5 is running $12–14/GB, and registered ECC is worse. 256GB of DDR5 RDIMM is about **$9,200** on NeweggBusiness today. To clear Inkling's 1-bit floor with headroom you want 384GB, call it **$14,000 in RAM alone**, and what you have bought for that money is the privilege of running a 1-bit quant slowly.

The GPU path is worse:

Target quant	Size	RTX Pro 6000 96GB cards	GPU cost
UD-Q2_K_XL	317.3 GB	4	~$53,000
UD-Q3_K_XL	432.8 GB	5	~$66,250
UD-Q4_K_XL	587.0 GB	7	~$92,750

At $13,250 per card on NVIDIA's own listing, up 55% from launch MSRP because GDDR7 is scarce, the quant that "still feels like Inkling" is a five-card, $66,000 machine before you buy a chassis.

And the escape hatch everyone reaches for is gone. The Mac Studio was the honest answer to this class of problem for two years: one quiet box, unified memory the GPU can address, 512GB if you paid for it. Apple pulled the 512GB config in March 2026 and the 256GB by May. The 128GB went at some point after that; Apple never announced when. **Today the biggest Mac Studio you can order new is 96GB, at $5,299.** An M3 Ultra that costs $1,300 more than it did in the spring, for a fraction of the memory. And this wasn't only config trimming — on March 26 Apple discontinued the Mac Pro outright, deleting the 192GB machine that sat above the Studio, with no replacement planned. The only Mac that goes higher than 96GB now is a laptop: the M5 Max MacBook Pro still configures to 128GB, which makes a notebook Apple's highest-memory

Mac. It's no workstation substitute, running 614 GB/s of memory bandwidth against the M3 Ultra's 819, and 128GB doesn't reach Inkling's 270GB floor anyway. The "just buy a big Mac" path to giant MoE models didn't get expensive. It stopped existing.

Here's the whole article in one table:

Model	Smallest shipping quant	Size	What actually holds it
Inkling 975B	UD-IQ1_S	**270.2 GB**	Threadripper/Epyc, 384GB RDIMM (~$14K of RAM)
Inkling 975B (decent)	UD-Q3_K_XL	432.8 GB	512GB workstation, or 5× Pro 6000 (~$66K)
GLM 5.2 753B	UD-IQ2_M	239 GB (~245 GB RAM)	256GB workstation (~$9.2K of RAM)
Inkling-Small 276B	not shipped yet	~123 GB projected at 3-bit	128GB workstation — if it lands as expected
Qwen 3.6-35B-A3B	UD-Q3_K_XL	**16.8 GB**	**Your RTX 3090. Tonight.**
Qwen 3.6-27B	Q4_K_M	**16.8 GB**	**Your RTX 3090. Tonight.**

16.8 GB against 270.2 GB. The open frontier is **16× past the card most of us run**, at its most compromised setting.

The one row worth watching is Inkling-Small: **276B total, 12B active**. TML says, again verbatim, "We are currently finishing the testing of Inkling-Small and will release its full weights once that work is complete." The weights aren't out (the HF repo 404s), so the ~123 GB above is my projection from Inkling's own bytes-per-parameter ratio, not a measurement. Treat it as an estimate and hold me to it when the real files land. But if it comes in near that, Inkling-Small at 3-bit is a 128GB workstation model. Still not a 3090. Considerably less absurd than $66,000.

The MoE trap: fast if you can hold it

There's a tempting misread of MoE models that I want to kill before it costs someone money.

41B active on 975B total sounds like salvation. Only 41B of parameters fire per token, so the arithmetic per token is 41B-class, genuinely quick. People see that and conclude Inkling is somehow a "41B model with a big closet."

It isn't. **Active parameters set your compute. Total parameters set your memory bill.** The router can pick any 6 of 256 experts for any token, so all 975B have to be resident and reachable. There is no partial load. You pay for the whole model to sit there and use 4% of it per token.

Worse, the moment you solve the memory problem with system RAM instead of VRAM, the speed advantage you were promised evaporates. Those expert lookups now cross the DDR5 bus, and RAM bandwidth becomes the bottleneck rather than the GPU. On the GLM 5.2 path, same architecture story at 40B active on 753B total, RAM-offloaded inference lands in low single-digit tokens per second. Fine for batch jobs you kick off before bed. Miserable for anything you sit and watch.

So the MoE promise is real but conditional: **it runs fast if you can hold it in fast memory.** Holding it in fast memory is the entire problem, and it's the problem MoE doesn't solve.

Kimi K3: the frontier that didn't ship weights at all

This is the part everyone's getting wrong, and it's the most important signal of the week.

The story circulating for months was that Kimi K3 would be a ~2.5-trillion-parameter open model landing in Q3. I went looking for the source of that 2.5T figure. It traces to a Sina Finance piece from April 28 that uses the verb 被曝, "was exposed," which is Chinese tech-press shorthand for somebody told us and we can't say who. No Moonshot source. A Tencent follow-up two days later upgraded it to "according to information already disclosed," attributing it to nothing at all. Every content farm on the internet has been republishing that leak as spec ever since.

Then K3 actually shipped, around July 15, and the leak turned out to be wrong twice.

First, the parameter count. Moonshot's own quickstart doc says K3 is "Kimi's most capable model to date, with 2.8 trillion parameters." Not 2.5T. If you read 2.5T anywhere today (and you will, because most of the July coverage says it), you're reading an April rumor that the primary source now contradicts.

Second, and this is the one that matters: **there are no weights.** K3 is live on Moonshot's API, priced at $3.00 in / $15.00 out per million tokens, 1M context. Hugging Face has nothing. GitHub has nothing. There's no model card, no license, no benchmark table. Moonshot didn't even publish a launch post. Sina noted on July 16 that no official channel had run one.

Consider what that breaks. Moonshot open-weighted K2. And K2.5. And K2.6. And K2.7-Code. All of them 1T total, 32B active, all under the same Modified MIT license whose only real condition is that you display "Kimi K2.6" in your UI if you clear 100 million monthly users or $20

million monthly revenue. Four straight releases. The most reliable open-weights shipper in the business.

Then the flagship arrives — and it's an API product.

I want to be careful here, because the honest position is narrower than the fun one. Moonshot has not said K3 will stay closed. Weights may follow in a month; K2 shipped that way before. But the argument "K3 will be open because Moonshot is always open" is not a prediction anymore. It's a pattern that just got interrupted, and pointing at K2's license doesn't fix that. For anyone reading this to decide what to run locally, K3 is currently a model you can rent and cannot have.

So the trend line has two segments. The open models that shipped weights are too big for your hardware. The newest frontier model didn't ship weights at all.

Meanwhile, on your 3090: absolutely nothing happened

Now the part that keeps this honest, because the doom read of everything above is wrong.

You didn't lose anything this week. Go check. **Qwen 3.6-27B**, Apache 2.0, 262K native context, dense, 16.8 GB at Q4_K_M from Unsloth (18.0 GB from bartowski; the repos genuinely differ by about a gigabyte on embedding handling, so use bartowski's number if you're calculating what fits). It runs on a 3090 with roughly 6GB left for KV cache, which at its 64 KB/token works out to about 90K of context. It claims 77.2 on SWE-bench Verified, level with Sonnet 4.5.

Or **Qwen 3.6-35B-A3B**, the MoE, 35B total and 3B active. We benched this one: UD-Q4_K_XL is 22.4 GB and does ~101 tok/s on a single 3090. That's a tight fit, under a gigabyte of headroom once the weights are in, so if you want long context, drop to UD-Q3_K_XL at 16.8 GB. Here's the nice part: this model's hybrid attention means only 10 of its layers are full attention, so its KV cache costs about **20 KB per token**. Twenty kilobytes. That 6GB of headroom buys you north of 256K of context on a six-year-old card.

Both are vision-capable, incidentally, which most coverage of them forgets to mention.

That's the same hardware you had in June, running the same models at the same speed, doing the same work. The tier that serves you is healthy, Apache-licensed, and fits with room to spare.

What changed isn't your position. It's who the headlines are for.

For about three years, a frontier open-model release was aimed at you. Llama 2 70B, Mixtral, the early Qwens. Those landed and the immediate question was "what quant fits my card," and there was always an answer. That era is over. When Inkling lands at 975B, the honest answer to "what

quant fits my card" is none of them, and also that was never the question this model was built to answer. Inkling is aimed at labs and at the people renting H200 nodes. Its openness is real and it matters: for researchers, for auditors, for the outfits that will fine-tune it, for the principle that frontier weights can exist in public at all. It's just not aimed at your desk.

That's a loss of relevance, not a loss of capability. Those feel identical when you're reading launch coverage. They are not remotely the same thing, and conflating them is how you end up believing you got locked out of something you actually still have.

What to do about it

Running local tonight? Nothing changes. Qwen 3.6-27B for coding, 35B-A3B if you want speed and long context. 16.8 GB either way. Stop reading launch posts for hardware advice.

Want to use Inkling? Rent it. $1.87/$4.68 per million, currently half off, from Together, Fireworks, Modal, Databricks or Baseten. The weights are Apache 2.0 and genuinely yours, which is worth something real. Just not $66,000 of Blackwell.

Actually building a box for this class of model? Buy DIMM slots, not GPUs. The platform is the constraint. A Threadripper or Epyc board with 384GB of RDIMM (~$14K of memory at today's prices) runs Inkling's 1-bit quant slowly; five RTX Pro 6000s runs its 3-bit quant fast for $66K. Both are real. Neither is a consumer purchase, and don't let anyone tell you a Mac bridges this now. The Studio caps at 96GB, and the 128GB MacBook Pro above it is still 142GB short of Inkling's floor.

Waiting on something better? Watch Inkling-Small. 276B/12B, weights promised after testing, projecting to around 123 GB at 3-bit. If it lands near that, it's a 128GB workstation model, the first thing from this generation a serious enthusiast could plausibly own. That's the release to actually care about, and almost nobody is covering it because 276B doesn't make a headline the way 975B does.

The open frontier is winning and pulling away from you at the same time. Both are true. The weights are free and the memory to hold them costs more than a car, and the model you'd actually use on a Tuesday night was never in that conversation to begin with. Download Inkling if you like. You've got every legal right to it. You just can't open it.

> **A note on this guide:** This is not a benchmark. I have no firsthand numbers on Inkling, GLM 5.2, or Kimi K3 and I'm not going to pretend otherwise. Miu, the 3090 box everything here gets tested on, has 64GB of RAM and physically cannot hold the smallest Inkling quant four times

> over. What you're reading is memory math: real file sizes from Unsloth's GGUF repo and the Inkling model card, checked against real memory capacities and live prices, plus capability data from Artificial Analysis. Every tok/s figure here is either from our own Qwen bench or labeled as somebody else's. The Inkling-Small sizes are my projection from Inkling's bytes-per-parameter ratio and are marked as such. Where primary sources disagree (975B vs HF's 952B, TML's Terminal Bench numbers vs AA's), I've shown you both instead of picking. That's the whole point: this is guidance math, and the arithmetic is checkable even when the hardware isn't in my hands.
