# Distilled vs Frontier Models for Local AI — What You're Actually Getting

February 25, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Distilled models score within 5-10% of frontier on benchmarks but break in predictable ways during extended agentic work: tool use rigidity, error recovery loops, context coherence failures. For short tasks (chat, summarization, code completion), distilled local models deliver 90% of the quality at 15% of the cost. For multi-hour agent sessions, the gap is real and growing. Test any model yourself: give it a complex task, then change one constraint and watch whether it adapts or regenerates from scratch.

On February 23, 2026, Anthropic disclosed that three Chinese labs ran **16 million automated conversations** across **24,000 fake accounts** to systematically extract Claude's capabilities. MiniMax alone pulled over 13 million exchanges. Moonshot targeted agentic reasoning and tool use with 3.4 million. DeepSeek ran 150,000 focused on step-by-step logic. When Anthropic released a new model mid-campaign, MiniMax pivoted within 24 hours, redirecting half their traffic to capture the fresh capabilities.

That's not research. That's an industrial extraction pipeline. And the models built from it are in your Ollama library right now.

The espionage angle is a sideshow, though. The real question is what distillation does to a model's actual capabilities, because the answer explains failures you've probably already noticed.

## What Distillation Actually Is

Distillation is not copying a model. It's compression.

The technical version: a smaller "student" model trains on the outputs of a larger "teacher" model, using temperature-scaled probability distributions to capture the teacher's decision-making patterns. The student learns to mimic the teacher's chain-of-thought reasoning, its confidence levels, its output style.

The practical version: it's like making an MP3 from a vinyl record. The hits come through clearly. The depth doesn't.

A frontier model like Claude or GPT-4 occupies a wide capability space. It can handle a broad range of tasks because it was trained on diverse, carefully curated data with extensive reinforcement learning. A distilled model occupies a narrower slice of that space, optimized for whichever tasks the distiller chose to target.

Think of it like watching football highlights versus watching full games. The highlights show you every touchdown. They skip the play development, the read progressions, the defensive adjustments that created those touchdowns. You know what happened. You don't know why it worked.

Moonshot's 3.4 million conversations targeted agentic reasoning, tool use, and coding. DeepSeek's 150,000 focused on foundational logic and step-by-step reasoning. They were extracting the touchdowns, but the play development stayed locked in Anthropic's weights.

## Why Distilled Models Look Good on Benchmarks

This is where things get misleading.

Standard benchmarks (MMLU, HumanEval, MT-Bench, SWE-Bench) test exactly the types of tasks that distillers optimize for. Well-defined, narrow problems with clear right answers. Distilled models routinely score within 5-10% of frontier models on these evals.

Enterprise buyers run standard evaluations, see comparable scores, conclude the models are equivalent. They are not. The benchmarks measure the narrow center of a model's capability space, the area where distillation works best.

Dario Amodei, Anthropic's CEO, put it bluntly: "There was a test recently where some of these models scored very highly on the usual SWE benchmarks…But then when someone made a held-back benchmark — one that had not been publicly measured — the models did a lot worse on that."

The Kimi K2 Thinking model is a case study. Moonshot claimed 44.9% on Humanity's Last Exam. Independent verification found **23.9%**. That's not a rounding error — the real score was barely half the claimed number.

We wrote about this pattern in depth: The Benchmarks Lie: Why LLM Scores Don't Predict Real-World Performance. The benchmarks aren't wrong exactly. They measure a real thing. That thing is benchmark performance, not general capability.

## Where Distilled Models Actually Break

For the first hour or two of a task, a distilled model feels comparable to frontier. The cracks show up later.

Tool use gets rigid first. Distilled models use tools in the patterns they were trained on. Give them a standard API call sequence and they'll nail it. Ask them to improvise a novel combination to solve an unexpected problem and they'll refuse, hallucinate a tool that doesn't exist, or force-fit a familiar pattern onto an unfamiliar situation. Kimi K2 literally called tools that weren't declared in the current request, pulling from memorized patterns instead of reading its actual context.

Error recovery loops come next. A frontier model hits an unexpected failure and rethinks its approach. A distilled model retries the same failing strategy, produces a technically valid but strategically wrong fix, or generates increasingly elaborate workarounds that miss the root cause. The agent trust decay problem hits distilled models faster and harder.

Context coherence degrades over long sessions. Distilled models forget constraints established 30 messages ago. They contradict their own earlier reasoning. They solve step 7 in a way that invalidates their solution to step 3. The model memorized what good reasoning looks like without internalizing how to actually reason across a sustained argument.

Generalization failure is the big one. Change one variable in a task that the model previously handled well. A frontier model adapts, identifies which parts of its previous reasoning still apply, revises the rest. A distilled model regenerates from scratch, or worse, force-fits its old solution onto new constraints.

Kimi K2 is the poster child. It hallucinated sources in 3 of 10 tests when writing technical explainers with citations. Its coding abilities were described as "destroyed" after quantization. Users reported it "wasted huge numbers of tokens while producing broken code anyway." The distillation captured output patterns but missed the underlying reasoning structure.

## The Manifold, Explained Without Math

One mental model clears up the whole distillation debate.

Picture a frontier model's competence as a plateau. Wide, flat surface. Step sideways from coding to creative writing to formal logic to tool orchestration, you stay on solid ground.

A distilled model's competence is a mountain peak. It reaches the same height as the plateau on specific benchmarks. Step sideways and you hit a cliff.

| | Frontier Model | Distilled Model |
|---|---|---|
| Short, well-defined tasks | Strong | Strong (within 5-10%) |
| Multi-step reasoning | Strong | Degrades after ~2 hours |
| Novel tool combinations | Adapts | Force-fits familiar patterns |
| Off-benchmark tasks | Solid | Steep quality dropoff |
| Extended agentic work | Maintains coherence | Progressive degradation |

The hard numbers back this up. DeepSeek's own R1 distillation shows the pattern clearly:

| Model | MATH | MMLU Formal Logic | Size |
|---|---|---|---|
| R1 Teacher | 90.5% | 97.6% | 236B MoE |
| R1-70B Student | ~88% | ~90% | 70B |
| R1-1.5B Student | 65.6% | 47.6% | 1.5B |

The 70B student is close on benchmarks, close enough to look equivalent in a standard eval. The 1.5B student drops 50 points on formal logic. But even the 70B student is standing on a mountain peak, not a plateau. The benchmark measures the peak height. It doesn't measure the cliff just out of frame.

## What This Means for Local AI

Most popular local models have some distillation in their lineage. This isn't speculation — it's documented.

DeepSeek R1 was built from 150,000 exchanges extracting Claude's step-by-step reasoning and alignment patterns. Anthropic caught them asking Claude to "articulate internal reasoning step-by-step," generating chain-of-thought training data directly.

Moonshot's Kimi models came from 3.4 million exchanges targeting agentic reasoning, tool use, coding, and computer vision. K2.5 can self-direct up to 100 sub-agents with 1,500 tool calls. On paper, impressive. In practice, the hallucination and coherence issues are direct consequences of the narrowed manifold.

MiniMax ran over 13 million exchanges, more than three-quarters of all detected extraction traffic. They pivoted to capture a new Claude release within 24 hours. That's a production pipeline.

Google confirmed Gemini was targeted by over 100,000 structured prompts in a coordinated extraction attempt.

OpenAI separately accused DeepSeek of running similar operations against their models, describing “obfuscated third-party routers” designed to evade detection.

Worth noting the uncomfortable angle: Anthropic settled a copyright lawsuit for $1.5 billion in September 2025 after a judge ruled they’d pirated over 7 million copies of books from Library Genesis. The “unauthorized extraction of capabilities” framing is rich coming from any lab in this industry. Everyone’s hands are dirty. The question is what the extraction does to the resulting models.

The downstream problem is monoculture. When every popular local model distills from the same two or three frontier models, they all inherit identical reasoning patterns and similar failure modes. A jailbreak that works on one works on all. A blind spot in the teacher propagates to every student. Research on model collapse shows that even a tiny fraction of synthetic training data (1 in 1,000) can degrade model diversity over time.

None of this makes local models bad. For narrow, well-defined tasks, distilled models are excellent value. But if you’re building agents that run for hours or doing complex coding sessions, model provenance matters.

## The Practical Framework: Matching Model to Task

Stop thinking “local vs cloud” and start thinking “narrow vs wide.”

Narrow tasks like email classification, summarization, code completion, simple chat, data extraction: these sit squarely on the mountain peak. Distilled local models handle them at 90% of frontier quality for roughly 15% of the cost. Run them locally. Save your money.

Wide tasks like multi-step research across unfamiliar domains, autonomous coding across multiple repos, agent workflows that run for hours: these require the plateau. The manifold width matters, and frontier models earn their cost.

The skill to develop is model routing, knowing where your task sits on the narrow-to-wide spectrum. Same principle behind a tiered AI model strategy: route simple subtasks to cheap local models, route complex reasoning to frontier. Your intent engineering determines how well this works.

For agent frameworks: let the local model handle individual tool calls, data lookups, straightforward code generation. Escalate to a frontier model when the agent needs to rethink strategy, recover from unexpected failures, or synthesize information across domains.

You can estimate VRAM requirements for any model you're considering with the Planning Tool, and our llama.cpp vs Ollama vs vLLM comparison covers inference backends for different use cases.

## How to Test for Yourself: The Off-Manifold Probe

Forget leaderboards. This 15-minute test tells you more about a model's actual depth than any benchmark.

Pick a complex task in your domain. Not a benchmark problem, something you actually work on. Write a detailed prompt and run it on the model you want to evaluate.

If the model succeeds (and for many tasks, it will), change one constraint. Add a requirement that conflicts with the obvious approach. Remove a resource the model relied on. Shift the target format or audience.

Then watch.

A model with representational depth, one that internalized the reasoning and not just the patterns, will adapt. It identifies which parts of its previous approach still hold, revises the rest, and produces a modified solution that accounts for the new constraint.

A model running on memorized patterns will regenerate from scratch. Or worse, it'll force-fit the old solution onto new constraints, producing output that looks structured but fails under examination.

If both models handle the first constraint change, add another. Increase the pressure. The plateau model keeps walking. The mountain peak model falls off the cliff.

This test works because it probes off-manifold performance, the exact capability that distillation strips away. It's specific to your work, which matters more than any generic benchmark.

Distilled local models are genuinely useful for a huge range of tasks. Run them, save money, keep your data private. But understand what you're getting: a model that learned what the frontier model says without fully learning why it says it. For most local AI work, that's plenty. For the agentic future where AI runs for hours on open-ended problems, it's the gap that defines your ceiling.

Know the difference. Route accordingly.

---