GPT-5.4 Just Dropped. Here's Why I'm Not Switching.

March 5, 2026 · by Mark Bartlett

Download this post as PDF

OpenAI shipped GPT-5.4 today. It's their best model by a wide margin, and I want to be honest about it before I make the case for why it doesn't matter to most of us.

What GPT-5.4 actually is

The headline numbers:

Benchmark	GPT-5.4	GPT-5.2	Notes
OSWorld-Verified	**75.0%**	47.3%	Beats human performance (72.4%)
SWE-Bench Pro	**57.7%**	—	Real GitHub issue resolution
GDPval (professional tasks)	**83.0%**	—	44 professions tested
MMMU-Pro (vision)	**81.2%**	—	Visual understanding

OSWorld is the one that'll get the headlines. It measures whether a model can navigate a real desktop environment through screenshots and mouse/keyboard actions. GPT-5.4 scores 75%, which is above the human baseline of 72.4%. That's a first.

The other big deal: it's the first model that combines GPT-5.3 Codex-level coding with reasoning and computer use in a single model. Previous GPT-5 variants made you pick your lane. This one does all three.

Context window is 1.05 million tokens via the API. That's roughly 3,000 pages of text in a single prompt.

What it costs

Tier	Input	Output
Standard	$2.50/1M tokens	$15.00/1M tokens
Pro	$30.00/1M tokens	$180.00/1M tokens

Tier	Input	Output
Batch	$1.25/1M tokens	$7.50/1M tokens

Prompts over 272K tokens get charged at 2x input and 1.5x output for the full session. So that 1M context window gets expensive fast if you actually use it.

For comparison, Claude Opus 4.6 runs $5/1M input and $25/1M output. GPT-5.4 standard is cheaper on input, comparable on output. The Pro tier is wild though – $180/1M output tokens. That's for the "xhigh" reasoning mode, and it'll drain budgets in a hurry.

ChatGPT Plus subscribers ($20/month) get GPT-5.4 Thinking. That's the consumer play.

Where it beats local models

I'm not going to pretend otherwise. GPT-5.4 is better than anything you can run at home in several ways.

Computer use is the big one. No local model navigates a desktop well yet – clicking buttons, filling forms from screenshots, that kind of thing. The 75% OSWorld score is in a different league from anything open-source.

Context length is another gap. 1M tokens is roughly 200x what you're running on an 8GB GPU. Even a 128GB Mac tops out at 128K-256K in practice. GPT-5.4 can hold entire codebases in context at once.

The GDPval benchmark tested 44 professions and GPT-5.4 hit 83%. Local 7B-14B models can't match that breadth. And the new "tool search" feature reduced token usage by 47% on the MCP Atlas benchmark – it's getting good at picking the right tool from a large set.

None of this is close on consumer hardware right now. Worth being honest about that.

Where local still wins

But GPT-5.4 doesn't change the reasons I run local models. Those reasons were never about matching frontier benchmarks.

Every prompt to GPT-5.4 passes through OpenAI's servers. Your code, your internal docs, your client projects. If you're working on anything proprietary, that's a liability. My Ollama instance processes everything locally. Nothing leaves my network.

I paid for my RTX 3090 once. Every token after that is free. GPT-5.4 at $15/M output tokens adds up fast – a heavy coding session burns through millions of tokens. At local, the meter isn't running.

OpenAI has changed pricing, removed features, and throttled access before. My local setup works the same today as it did yesterday. Nobody can remotely update my model weights or deprecate my toolchain. And it works on planes, bad WiFi, air-gapped environments. GPT-5.4 needs the internet. My Mac Mini doesn't.

Meanwhile, open-source keeps closing the gap. A year ago, the distance between frontier and local was enormous. Today, Qwen 2.5 Coder 32B hits 92.7% on HumanEval and matches GPT-4o-era performance on the Aider benchmark. Qwen3-Coder-Next scores 70.6% on SWE-bench Verified. These models run on a single GPU. The gap shrinks with every release.

The actual question

GPT-5.4 vs local isn't an either/or. The people freaking out on Twitter today are framing it wrong.

The question isn't "should I switch to GPT-5.4?" It's "what's worth paying for vs what's worth owning?"

For me, the split looks like this: I run local models for daily coding, private work, and anything I want to keep on my machine. If I ever needed 1M-token context or desktop automation at superhuman level, I'd pay for a GPT-5.4 session. Different tools for different jobs.

Most developers I know who run local AI already use a mix. Local for the 90% of tasks that a 7B-32B model handles fine. Cloud API for the 10% that genuinely needs frontier capability. GPT-5.4 makes that 10% better. It doesn't make the 90% more expensive.

What I'm watching

Two things to track in the coming weeks:

First, the open-source response. Every major frontier model release accelerates open-source development. When GPT-4 shipped, open-source caught up within 18 months. When GPT-5 shipped, it took 6 months. The cycle keeps compressing. Qwen, DeepSeek, and Meta will be studying these benchmarks.

Second, the pricing ceiling. GPT-5.4 Pro at $180/M output tokens is a signal. The best reasoning comes at a premium, and that premium is going up, not down. The more expensive frontier gets, the more attractive "good enough locally for free" becomes.

Bottom line

GPT-5.4 is a great model. The OSWorld score alone is worth paying attention to. If you need frontier capability and don't mind cloud dependency, it's the best option available today.

But if you're running local models, nothing about today's announcement changes your setup. Your GPU still works. Your models are still free. Your code still stays private. And the open-source ecosystem that feeds your local stack just got another target to aim for.

The cloud keeps getting better. So does local. They're not playing the same game.

Source: https://insiderllm.com/blog/gpt-5-4-what-it-means-for-local-ai/

Free guides for running AI locally