14: Keeping up without drowning

How to follow a field that ships every week, and when to ignore it

Why this matters

- A major model, tool, or benchmark ships almost every week
- You can't track all of it, and you don't need to
- Quality goes up *and* down, usually for reasons outside the model
- Access can vanish overnight for non-technical reasons
- The fundamentals change slowly, so build on those

How fast it moves: three months of 2026

Date (2026)	What shipped
Jul 24	Claude Opus 5
Jul 30, Aug 28	Terminal-Bench 3.0, then 4.0 replaces it
Sep 1, Sep 3	Claude Fable 5.1; OpenAI's GPT-6 Astra
Sep 22	Claude Opus 5.5 and GPT-6 Sol and Luna, same day
October	Google's Gemini 4 models

Codex CLI was at version 0.167 by November 6, 2026.

What stays stable

Durable skill	Why it survives
Giving the model the right context	Every model is limited by what you show it
Verifying output	Models remain confidently wrong some of the time
Security habits	Attacks target the agent pattern, not one model
Breaking work into checkable steps	Reliability falls with task length, for every model
Your own eval set	Tells you what any release means for your work
Cost awareness	Prices change; the habit of measuring doesn't

For any technique, ask: *will this still matter after the next model?*

Ups and downs

Quality changes, and it's usually not the weights

Quality comes from a stack of layers

Layer	What can change without notice
You	Task, prompt, session length, repo size
Harness	CLI or IDE version, system prompt, default effort, MCP servers, compaction
Product	Plan, usage limits, model routing, model picker
Serving	Hardware, load balancing, sampling code, caching
Model	Weights and version

People blame the bottom row. Documented causes sat more often in the middle.

A field guide to recent ups and downs

When	What users saw	Cause, per the company
Apr 2025	Sycophantic GPT-4o, rolled back in days	Overweighted short-term user feedback
Aug–Sep 2025	Worse Claude answers for weeks	Three infrastructure bugs
Sep–Oct 2025	Codex "got worse"	No single cause: hardware, compaction, a sampling bug
Mar–Apr 2026	Six weeks of Claude Code complaints	Three product-layer changes; weights unaffected
Jun–Jul 2026	Claude Fable 5 suspended for all users	US export controls
Jul–Sep 2026	Opus 5: strong benchmarks, mixed reviews	Behavior: verbosity, scope, communication

Two lessons from the record

- Labs' own monitoring often missed what users noticed
- Anthropic, 2025: its evaluations "simply didn't capture the degradation"
- Stated causes were rarely "we made it worse on purpose"
- You can't verify a vendor's promise. You can verify your golden tasks.

Claude and Codex

> **Bryan's take:** Claude and Codex have both had ups and downs.

It feels worse: check in this order

Check	If yes
1. Status page or incident report?	Wait it out, or switch tools for now
2. Did your harness version, plan, defaults, or MCP setup change?	Pin or roll back; rerun a 5-task smoke test
3. Are golden tasks passing less often than usual?	Likely a real regression: collect specifics, report, consider switching
4. Is today's task harder: bigger repo, longer session, vaguer spec?	Fix the context: fresh session, smaller step, clearer spec
None of these	Log it in your vibes log; recheck in a week

When to switch tools

Good reasons	Bad reasons
Golden tasks drop in two runs, a week apart	One bad afternoon
A rival beats yours beyond your error bar, affordably	A launch-day benchmark chart
Pricing, limits, data policy, or security changed	A viral thread
A long bad stretch: no fix, no postmortem, no word	Fear of missing out

Always keep a second tool warm.

Filtering the noise

Launch claims, sources, and a routine you can keep

Weighing launch claims

- Self-reported benchmarks? Which harness and settings?
- What was dropped since the last comparison? Who's missing?
- Is the demo live, edited, or sped up?
- "Available today" may mean a limited set of organizations
- Customer quotes and internal runs are testimonials, not tests
- Wait for independent numbers; they usually come within days

Every lab has a mixed record

Lab	A documented episode
Meta	Apr 2025: promoted the Arena rank of a chat-tuned experimental variant
xAI	Jul 2025: the `@grok` account posted antisemitic content for about 16 hours
OpenAI	Feb 2026: publicly retired SWE-bench Verified, its own benchmark, after an audit
Anthropic	Detailed postmortems in 2025 and 2026, after users noticed first

Judge each on replication, honesty about regressions, demos vs product, stable pricing.

Google's launch claims

> **Bryan's take:** I don't trust Google's launch claims much. In my experience they overhype Gemini, and it underperforms what was promised, especially for coding. I'll use Gemini for some specific tasks, but not as my coding agent.

Google: the record cuts both ways

Against	For
Dec 2023: the "Hands-on with Gemini" video was staged	Jul 2025: IMO gold, certified by IMO coordinators
Jun 2026: Gemini 3.1 trailed GPT-5.5 on Harbor-Index, though no comparison was statistically significant	Nov 2025: Gemini 3 Pro topped Arena at launch
	Nov 2026: Gemini 4 Carbon #1 on Arena's text leaderboard (preliminary)

Your golden tasks will tell you whether it shows up in *your* work.

An information diet: pick two or three sources

Source	Good for	Keep in mind
Simon Willison's Weblog	Fast hands-on notes; security	Enthusiast who often gets early access
Latent Space	Long interviews with builders	Industry insider
Interconnects	Post-training and open models	Open-model advocate
The Pragmatic Engineer	How real teams use AI	Big-tech lens; deep dives paywalled
METR, Epoch AI	Data, trends, studies	Risk or quantitative focus
Changelogs, status pages	What actually shipped	Silent about what got worse

Never cite SEO or auto-generated "AI news" sites.

A sustainable routine

Cadence	Time	What to do
Weekly	30 min	Tool changelogs (10 min), one or two newsletters (15), one try-list item (5)
Monthly	60–90 min	Run golden tasks on one or two releases; record the result; drop one unused tool
Quarterly	Half a day	Cost vs value, refresh golden tasks, review security, update the radar

Rules that prevent churn

- Wait two weeks, unless it solves a problem you have today
- Change one thing at a time: model, harness, or config, not all three
- It has to win on your tasks before you switch
- Ignore small benchmark moves; a few points is noise for you
- Mute sources that mostly make you anxious

A team AI radar

Ring	Meaning
Adopt	Proven on our golden tasks; the recommended default
Trial	Promising; a few people use it on real work
Assess	Worth an afternoon; someone is testing it
Hold	Don't start new work with it, and say why

Rotate a monthly scout. Items move at most one ring per review, with evidence.

Evaluating a new release in an afternoon

Time	Step
0:00–0:20	Read the claims; pick two or three to test
0:20–0:40	Check independent signals
0:40–2:30	Run 10 golden tasks in your real harness, graded blind
2:30–3:15	Try one real task from this week's backlog
3:15–3:45	Measure cost, quota, and speed per task
3:45–4:00	Decide: Adopt, Trial, Assess again in two weeks, or Hold

A close result means "not yet."

The pace itself

> **Bryan's take:** The rapidly changing nature of all this is interesting too.

Try it

On your own (about an hour, then ongoing):

- Subscribe to two sources; bookmark your tools' changelogs and status pages
- Start a `vibes-log.md`: date, model, harness and version, task, result
- Log five entries in two weeks, then run a short afternoon eval

As a class or team, start an AI radar, or present one incident from the field guide.

Discussion

1. Most incidents weren't in the weights. Why do people blame the model first?
2. How could an outside user test "we never degrade quality due to demand"?
3. How much time should engineers spend keeping up weekly? On company time?
4. How do you tell healthy experimentation from tool churn on a team?

Read more

- Module 14: Keeping up without drowning
- A postmortem of three recent issues (Anthropic, September 2025)
- Ghosts in the Codex Machine (OpenAI Codex team, October 2025)
- Thoughtworks Technology Radar