



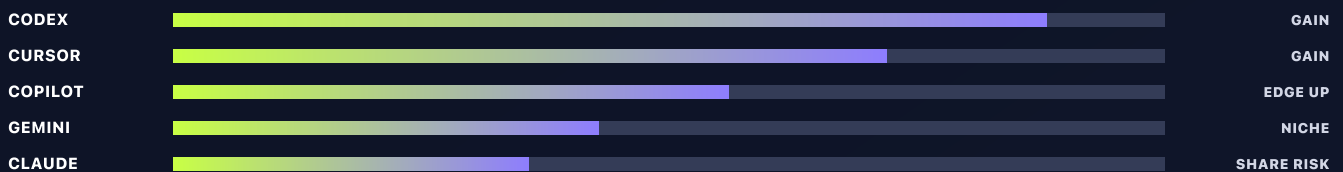
THE TOKEN REPORT

The coding-agent race just changed.

Eight weeks of user evidence across Claude Code, Codex, Cursor, GitHub Copilot and Gemini reveal a new competitive axis: not who writes the best code—but who can be trusted with the longest leash.

What thousands of builders already learned. Inside your agent's reasoning loop. Updated daily.

9-MONTH RELATIVE TRACTION OUTLOOK



THE BIG SHIFT

Capability is abundant. Control is scarce.

The last eight weeks were not a simple model horse race. Every major coding system gained power. The surprise was what users started complaining about once that power arrived.

Cloud agents, nested subagents, automatic routing and million-token models expanded what developers could delegate. But the conversation moved just as quickly toward context loss, runaway usage, invisible quotas, session failures and output that looked right while being wrong.

The durable advantage is shifting from model intelligence to **operational trust**: durable state, observable execution, predictable cost, stable releases and safe rollback.

“Users cannot reliably predict the outcome, cost or collateral behavior of a long-running coding-agent task.”

THE TOKEN REPORT, BASE-CASE THESIS

IN THIS ISSUE

- 03

Claude Code
Power without a governor
- 04

OpenAI Codex
The momentum trade
- 05

Cursor
The harness becomes a platform
- 06

GitHub Copilot
Distribution meets agent mode
- 07

Gemini / Antigravity
The beautiful wildcard
- 08

Nine-month forecast
Winners, losers, watchpoints
- 09

Practitioner playbook
How to allocate the stack
- 10

Use-case radar
Proven, emerging, fringe but real
- 11

Effectiveness map
Best, average, poor—and how to improve

5 SYSTEMS TRACKED	8 WEEKS OF EVIDENCE	70% CODEX GAIN PROBABILITY	60% CLAUDE SHARE-LOSS RISK
----------------------	------------------------	-------------------------------	-------------------------------

01 • CURRENT LEADER, HIGHEST VOLATILITY

Claude Code: power without a governor

9-MONTH CALL

RELATIVE SHARE RISK

60% probability of losing traction

MAY 31
tool-call regressions

JUN 11
agent view lands

JUN 29
quality backlash

JUL 1-10
cost + control waves

WHAT SURPRISED USERS

Orchestration at startling scale. Users observed five-level nested subagents and 105 agents in a single session. The new agent view turned the terminal into a control plane.

Accuracy can pay for itself. One practitioner attributed a 40% cost reduction to improved accuracy:

“And now my costs are actually down 40%, presumably because of the better accuracy.”

CLAUDE CODE DISCORD USER • JUL 6

WHAT DISAPPOINTED THEM

Release velocity outran reliability. Malformed tool calls, retry loops, timeout ceilings and ignored project rules repeatedly made interactive work unusable.

Long sessions lost the plot. Reports described circular reasoning, compaction loss, runaway subagents and unauthorized actions.

UNSOLVED PAIN

Predictable control of autonomous execution

Bound the agents, tokens, permissions and behavioral drift—without neutering the system.

THE BET: Claude remains top-tier, but Anthropic must trade some feature velocity for regression discipline, hard spend controls and reliable rollback.

02 · THE MOMENTUM LEADER

Codex: the strongest gain setup

9-MONTH CALL

LIKELY GAINER

70% probability of gaining traction

MAY 24
usability praise

JUN 10-15
value + one-shot wins

JUN 23
Windows regressions

JUL 7-9
5.6 surge, context cap

WHAT SURPRISED USERS

It finished work that rivals drifted away from. Users reported plans drifting in Claude while Codex completed the implementation in one shot.

The \$20 plan went further than expected. Code-review users praised allowance value; bankable limit resets felt like an effective price cut.

250–
353k

approximate CLI context reported by users—even when 1M was available through the API

WHAT DISAPPOINTED THEM

Windows and desktop state were brittle. Missing sessions, mixed queues, broken model pickers and indefinite “Working...” states made production use risky.

The context ceiling felt artificial.

“That’s basically a deal breaker for my research flow—one question often uses 200–300k tokens.”

DISCORD USER · JUL 9

UNSOLVED PAIN

Durable long-horizon context

Keep tools, task state and history intact across sessions, compaction, platforms and updates.

THE BET: Codex becomes the usage momentum leader if OpenAI exposes full context and fixes session continuity before release churn recreates Claude’s trust problem.

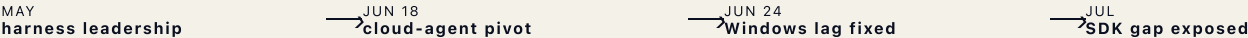
03 · THE HARNESS BECOMES A PLATFORM

Cursor: the integrated-team bet

9-MONTH CALL

LIKELY GAINER

60% probability of gaining traction



WHAT SURPRISED USERS

Cloud agents became operational. Reusable VMs, visible terminals, `/in-cloud`, parallel branches, local handoff and PR babysitting moved beyond remote chat.

The context stack handled real scale.

“Subagents have been great... Mine is over 100k LoC and connected to three other services.”

R/CURSOR USER · JUL 8

WHAT DISAPPOINTED THEM

Composer’s speed sometimes displaced reasoning. Users saw lazy automation on tasks that required deliberate implementation.

Headless did not equal IDE. The same task took about 3m15s in the IDE and 26 minutes through the SDK—with no semantic searches and 122 edit calls.

~8×

reported headless slowdown versus the Cursor IDE

UNSOLVED PAIN

Predictable usage economics

Translate models, cache, fast mode and tokens into productive hours and a believable bill.

THE BET: Cursor wins teams that value an integrated cloud-agent environment—but opaque economics leave an opening for cheaper CLIs and enterprise bundles.

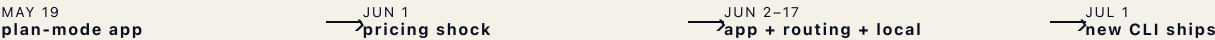
04 · DISTRIBUTION MEETS AGENT MODE

Copilot: the default, not yet the favorite

9-MONTH CALL

EDGE UP / STABLE

55% probability of gaining traction



WHAT SURPRISED USERS

A non-code workflow became an application. Cassidy Williams dragged a Photoshop template into plan mode, described the workflow and generated a functioning desktop tool.

“For this small little problem, I can now just click a button.”

CASSIDY WILLIAMS · MAY 19

Copilot became a multi-model surface. Local models, automatic routing and model swarms pushed it well past completion.

WHAT DISAPPOINTED THEM

June pricing damaged trust. Reports described roughly 3x renewal increases and a shift toward less predictable consumption billing.

The agent workflow remained fragmented. Plan/agent switching was cumbersome; shared CLI tools initially raised review cost while finding fewer issues.

UNSOLVED PAIN

A predictable cost-to-quality relationship

Show users what routing chose, what a task will cost and why agent mode is better than completion.

THE BET: Copilot gains enterprise seats through GitHub distribution, but power-user preference stays elsewhere unless the agent experience becomes coherent.

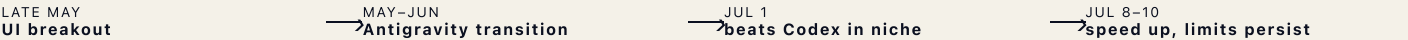
05 · THE BEAUTIFUL WILDCARD

Gemini: specialist upside, trust deficit

9-MONTH CALL

NICHE GAIN

40% probability of generalist gains



WHAT SURPRISED USERS

Frontend aesthetics looked designed, not generated. Practitioners praised its layouts, visual judgment and escape from default “AI UI.”

A real niche win over Codex.

“Gemini is beating Codex hands down by a huge margin.”

DISCORD USER, SMALL GRAPHICAL APPS · JUL 1

WHAT DISAPPOINTED THEM

Beautiful pages contained invented facts. One evaluation praised the visual result while finding nonexistent integrations and fabricated workflow claims.

Quota reality did not match the meter. A single-page project consumed a large share of a weekly allowance and still hit quota despite visible capacity.

“The design... looks fantastic, but the information is hallucinated in a lot of different places.”

COLE MEDIN · MAY 22

UNSOLVED PAIN

Grounded correctness

Preserve the visual imagination; remove fabricated repository facts and architectural assumptions.

THE BET: Gemini earns a place in multi-model stacks for UI and visual work. Generalist leadership waits on grounding, state recovery and sane quotas.

BASE CASE • THROUGH APRIL 2027

Everyone grows. Relative share diverges.

This forecast concerns practitioner preference and serious active usage—not bundled seats or revenue. Product launches can overturn it; current trajectory sets the prior.

1 CODEx 70% GAIN Switcher momentum, value and less drift. Fixable product debt.	2 CURSOR 60% GAIN Harness, cloud execution and workflow data compound.	3 COPiLOT 55% GAIN Enterprise distribution outpaces developer preference.	4 GEMiNI 40% GAIN Specialist UI adoption; generalist trust remains weak.	5 CLAUDE 60% LOSS RISK Still elite; reliability threatens relative share.
---	--	---	--	---

WHAT WOULD BREAK THE FORECAST?

- **Claude:** hard spend/agent controls plus a stability-first release cycle.
- **Gemini:** repository grounding paired with Google Cloud and Android distribution.
- **Copilot:** transparent routing that demonstrably improves quality per dollar.
- **Cursor:** a simpler, credible pricing meter—or a pricing backlash.
- **Codex:** full context and durable sessions—or recurring release regressions.

THE DECIDING VARIABLE

Operational trust

The market will not be won by the highest benchmark. It will be won by the first system to combine intelligence with durable memory, predictable cost, observable execution, safe rollback and stable behavior.

DON'T PICK A RELIGION. DESIGN A PORTFOLIO.

The stack for the next quarter

PRIMARY IMPLEMENTER

Codex or Claude Code

Use Codex where session stability is proven; keep Claude for deep reasoning and orchestration with explicit agent and spend caps.

INTEGRATED TEAM WORK

Cursor

Best fit for codebase context, local-to-cloud handoff, background agents and pull-request flow. Track spend outside the product.

ENTERPRISE DEFAULT

GitHub Copilot

Exploit repository and policy integration. Benchmark routed models and agent mode against a simpler completion baseline.

VISUAL SPECIALIST

Gemini

Delegate UI exploration and graphical apps. Never accept product facts, APIs or architecture without independent verification.

FIVE RULES FOR OPERATORS

- 1. Set a token and wall-clock budget before launch.**
Stop conditions are part of the prompt, not an afterthought.
- 2. Separate planning, implementation and review models.**
Cross-model review catches shared harness blind spots.
- 3. Persist state in the repository.**
Assume the next session remembers nothing.
- 4. Require evidence-bearing completion.**
Tests, diffs, logs and uncertainty—not “done.”
- 5. Measure the system, not the demo.**
Time-to-verified-output, repairs, regressions and total cost.

WHAT BUILDERS ARE DOING NOW

Where coding agents are actually going

The common use case is no mystery. The important signal is what happens when builders connect code execution to a measurable outcome.

PROVEN NOW

214

LINKED OBSERVATIONS

Bounded implementation

Give the agent a concrete change and an objective check it can run.

FEATURE	Build against acceptance tests.
BUG	Fix a reproducible failure.
REFACTOR	Change structure while behavior stays green.

Method: verification loop · 119 observations
Problem defeated: false completion · 41 observations

EMERGING

73

LINKED OBSERVATIONS

Closed-loop computational research

The agent proposes, runs, measures and revises—not merely writes the research code.

BRAIN2QWERTY	An Auto Research workflow found word-error-rate improvements beyond conventional hyperparameter optimization.
SOCIAL SCIENCE	Frontier coding agents reproduced computational findings as reliable workflow executors.

Signal: strongest for Claude Code, with additional Codex research evidence.

FRINGE, BUT REAL

1-3

SIGNALS PER CASE

Coding escapes software

GENEALOGY

Structured family-history and archival auto-research.

POKER

Solver bots used to test reasoning and optimization.

EDGE CAMERAS

An MCP server running directly on an AXIS IP camera.

Pattern: these differ by system. The interface determines the edge.

THE BIGGER SHIFT

Code is becoming the agent’s universal actuator.

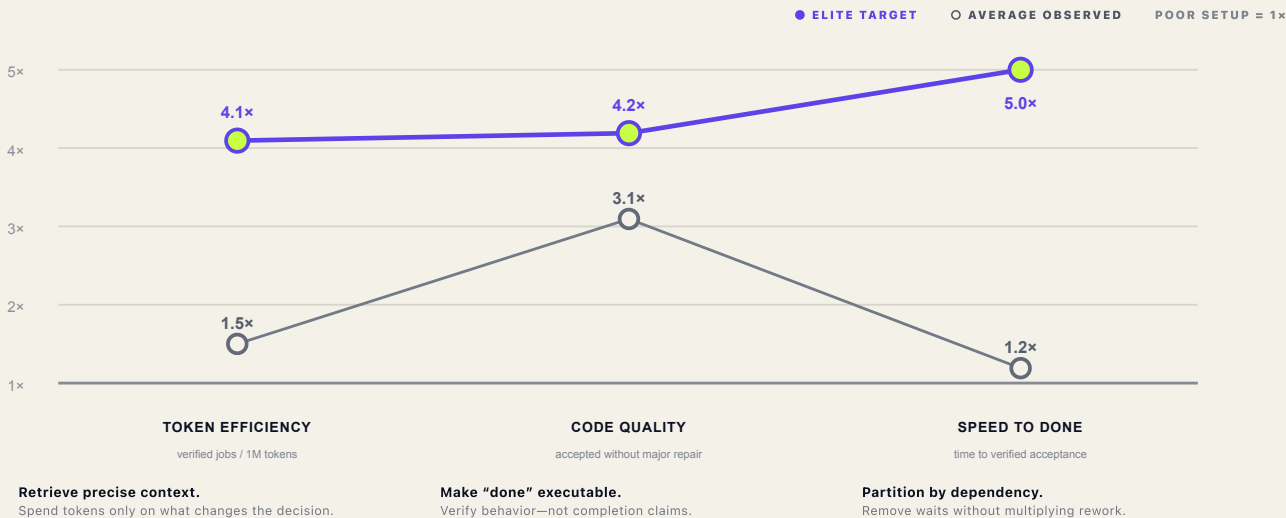
The product is no longer always software. Sometimes software is simply how the agent reaches the outcome.

EXPLORE WHAT BUILDERS ARE LEARNING → GETTRUTHAPI.COM

SEE THE GAP. CLOSE THE GAP.

What the best builders do differently

The prize is not more code. It is more verified work from every token, hour and review. Average is the baseline; the green distance is your opportunity.



RESEARCH BENEATH THE RANGE	-60% tokens FastContext improved retrieval and resolution. arXiv:2606.14066	12-15% pass ScarfBench exposed the floor on complex migrations. arXiv:2605.06754	4.8x faster MPAC removed coordination waits in multi-agent review. arXiv:2604.09744
----------------------------	--	---	--

Multiples normalize observed research ranges against the poor boundary. "Elite" is an evidence-informed target—not a measured percentile. Compare the same task class and count only verified work.

01 · GREENFIELD

Prove the whole path

- Deploy one vertical slice.**
Real data, one user job and the target environment.
- Make "done" executable.**
Build, browser and deploy checks.
- Standardize the bootstrap.**
Known auth, persistence, CI and observability.

02 · EXISTING CODE

Protect behavior

- Retrieve precise context.**
Symbols, callers, tests and recent decisions.
- Bound and regression-lock.**
Small diff, invariants, private checks and rollback.
- Review for mergeability.**
Independent correctness and maintainability gate.

03 · TEAMS

Scale the control plane

- Partition by dependency.**
Coherent ownership—not arbitrary file bundles.
- Separate authority.**
Build, verify and merge are distinct gates.
- Manage review capacity.**
Outcomes per reviewer-hour, queue age and rollback.

WHAT BEST DOES

+14%
PASS RATE

2.10x
WALL-CLOCK SPEED

-35%
API COST

Co-Coder moved all three together by partitioning work around repository dependencies.
arXiv:2606.00953

TruthAPI evidence, editorial judgment

The evidence window runs from May 18 through July 13, 2026. TruthAPI searches covered Reddit, Discord, GitHub issues and releases, YouTube, blogs and Twitter. Broad scans identified the market cohort; focused scans separated praise, surprise, regressions, cost, context and reliability; full records verified the quotations below. Cross-source waves were used to locate inflection points.

WHAT THOUSANDS OF BUILDERS ALREADY LEARNED

Six practitioner signals behind this issue

105 agents observed in one Claude session	40% lower cost attributed to better accuracy	250–353k Codex CLI context reported by users
100k+ LoC in a Cursor user's multi-service project	~8× reported Cursor headless slowdown vs IDE	~3× Copilot renewal increase reported by users

- CLAUDE [C]

C1 agent view · C2 nested subagents · C3 40% cost report · C4 malformed tool calls · C5 timeout regression · C6 quality regression
- CODEX [O]

O1 usability comparison · O2 lower-drift one-shot · O3 \$20 plan value · O4 banked resets · O5 Windows regressions · O6 context deal-breaker · O7 indefinite working state
- CURSOR [U]

U1 cloud-agent release · U2 100k-LoC workflow · U3 config awareness · U4 Windows performance fix · U5 speed/reasoning tradeoff · U6 headless slowdown · U7 plan-value concern
- COPILOT [P]

P1 plan-mode application · P2 model swarm · P3 local models · P4 redesigned CLI · P5 pricing change · P6 plan/agent UX · P7 code-review regression
- GEMINI [G]

G1 graphical-app comparison · G2 full-stack/UI praise · G3 visual quality · G4 hallucination/quota evaluation · G5 crash/state loss · G6 Antigravity downgrade · G7 rate-limit latency
- LIMITATIONS

These records are directional qualitative evidence, not a representative survey. The "top five" reflect visibility, not audited share. Model sentiment can be conflated with harness quality. Forecast probabilities express editorial confidence, not measured frequencies.

REPRODUCIBILITY

The exact TruthAPI queries, selected record IDs and processed conclusions are preserved in `findings/ai-coding-systems-truthapi-research-log.md`.