

LESSONS FROM THE CLAUDE CODE TEAM

Code verification when agents write all the code.

They didn't stop reviewing. They automated most of the reading and saved people for the risky parts.

2x+

pull requests merged, June
to September

-40%+

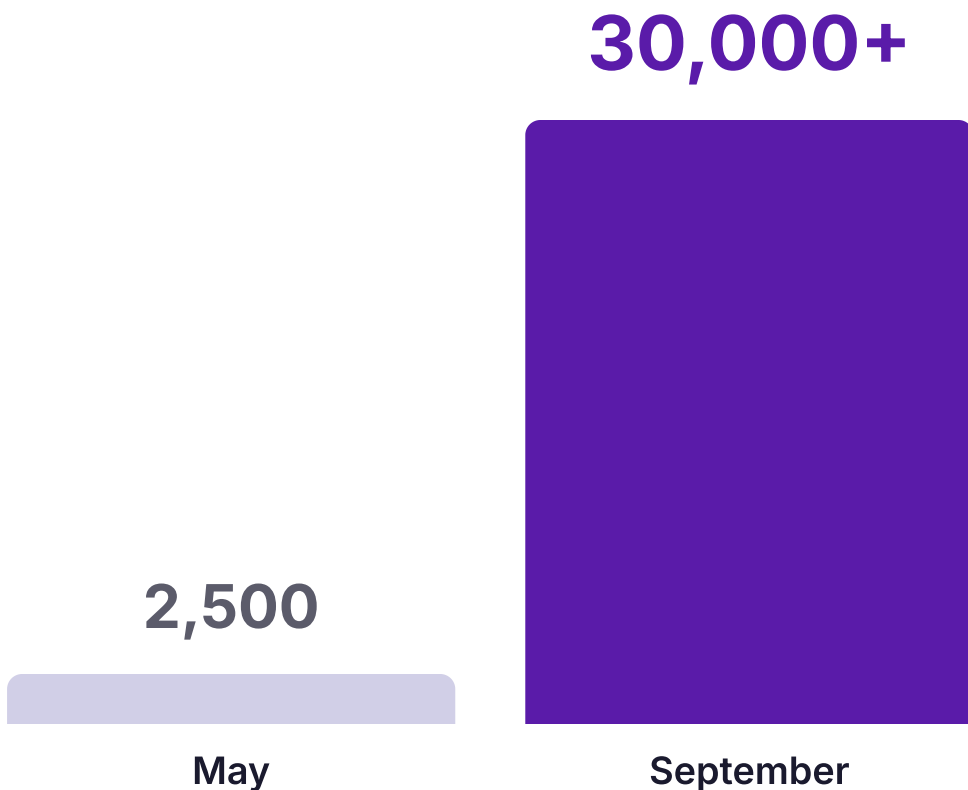
share of PRs reverted

Based on the Anthropic post shared by Addy Osmani

THE PRESSURE

Line by line stopped scaling.

Changes merged per day, Anthropic monorepo



THE NEW QUESTION

Not "did someone read it?" but "what could go wrong, and which check should catch it?"

THE GOAL

Human time has two sides.

Time reviewing

reading and approving
PRs



Time fixing

bugs that bad code let
through

Cut the first without cutting the second and you have only moved the work.

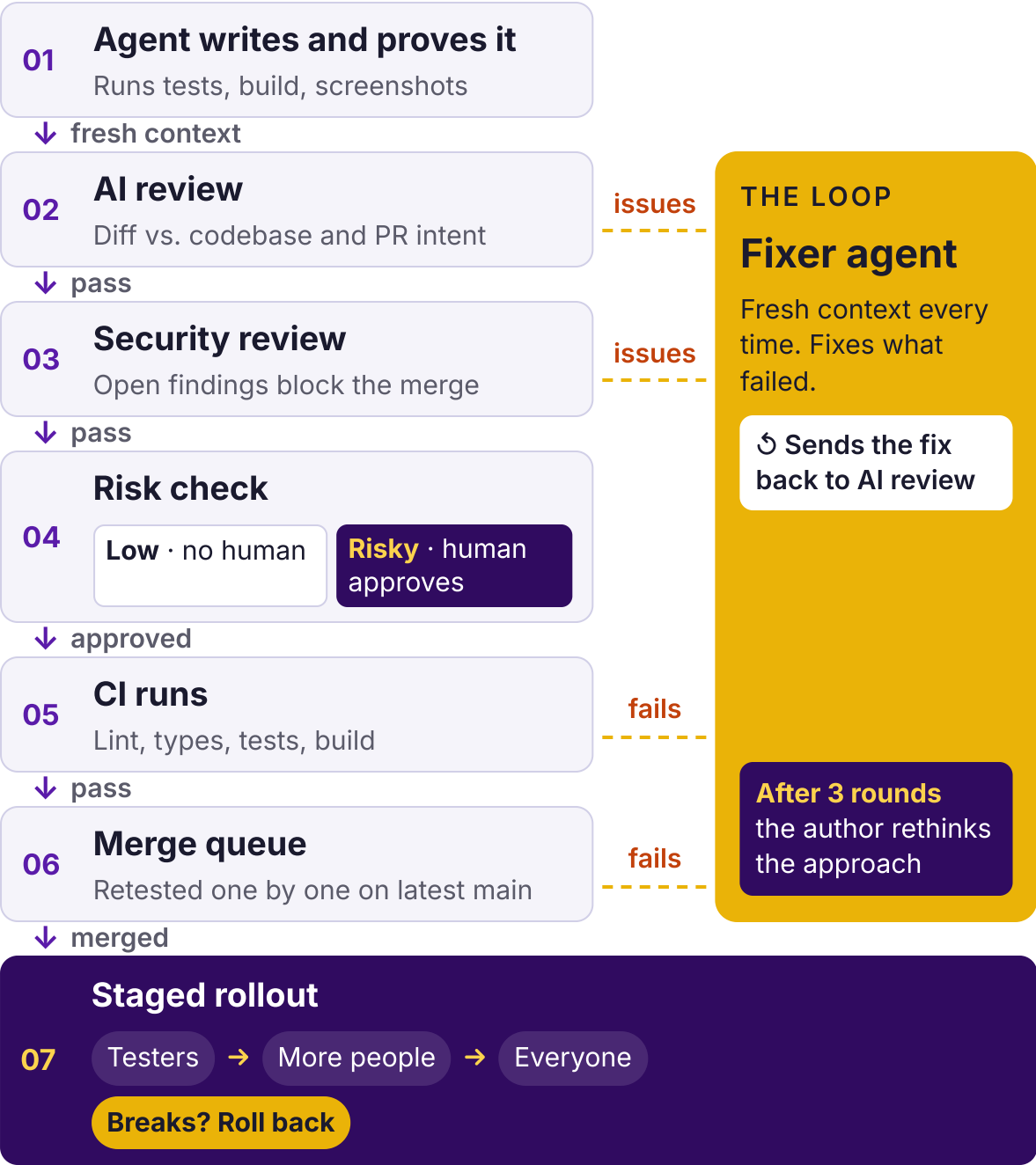
TARGET

Most PRs merge with at most one person in the loop.

Track: approvals per PR,
human comments, wait time

Guardrail: revert and
incident rates

Seven layers. One fixer loop.



Two green PRs can still **break main**.

LOW RISK

No human needed. Goes straight to CI.

RISKY

A human reviewer approves, then it goes to CI.

↓ both paths meet at CI

Many changes are ready

All of them passed CI on their own branch

↓ they wait in line

One at a time

When its turn comes, each change is next in line

↓

Retested on the latest main

With every change ahead of it already applied

↓ passes

Merged into main

Then the staged rollout starts

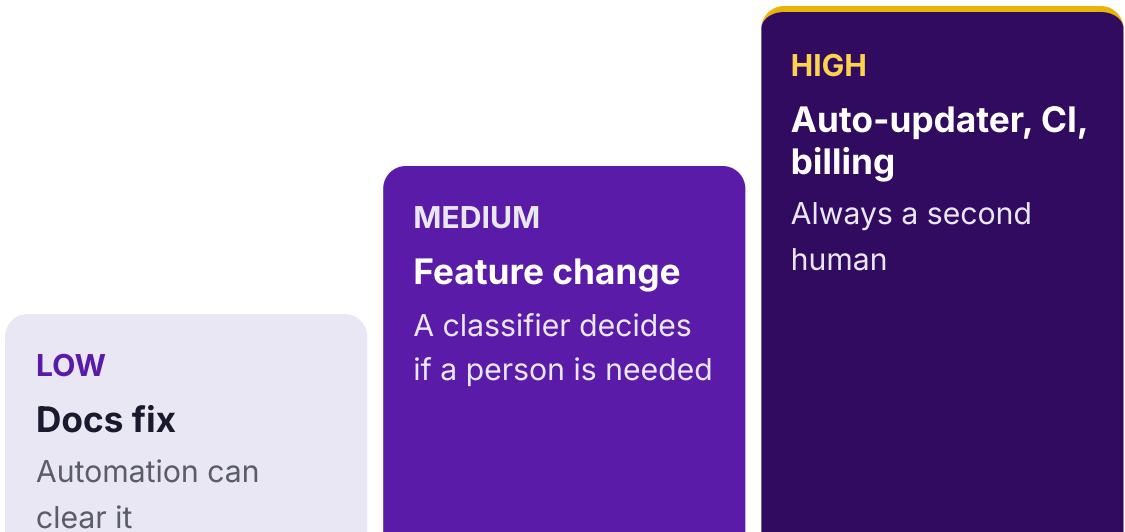
----- fails

Leaves the line

Goes back to the fixer agent, then re-enters review

Main only ever gets changes that passed together.

The change sets the bar, not the author.



-
- ✗ Agent approvals never count as human approvals.
-
- ↻ An approval covers one commit. A new untrusted commit needs a new approval.
-
- If an agent wrote the PR, the person running it is a required reviewer.
-

Agent writes. Agent reviews. **Forever.**

50+

review cycles on one PR overnight, until someone stopped it

~59%

of repeat flags were old code, not new bugs

How they made review converge

01

Cap at three cycles, then rethink

02

Re-review only the new diff

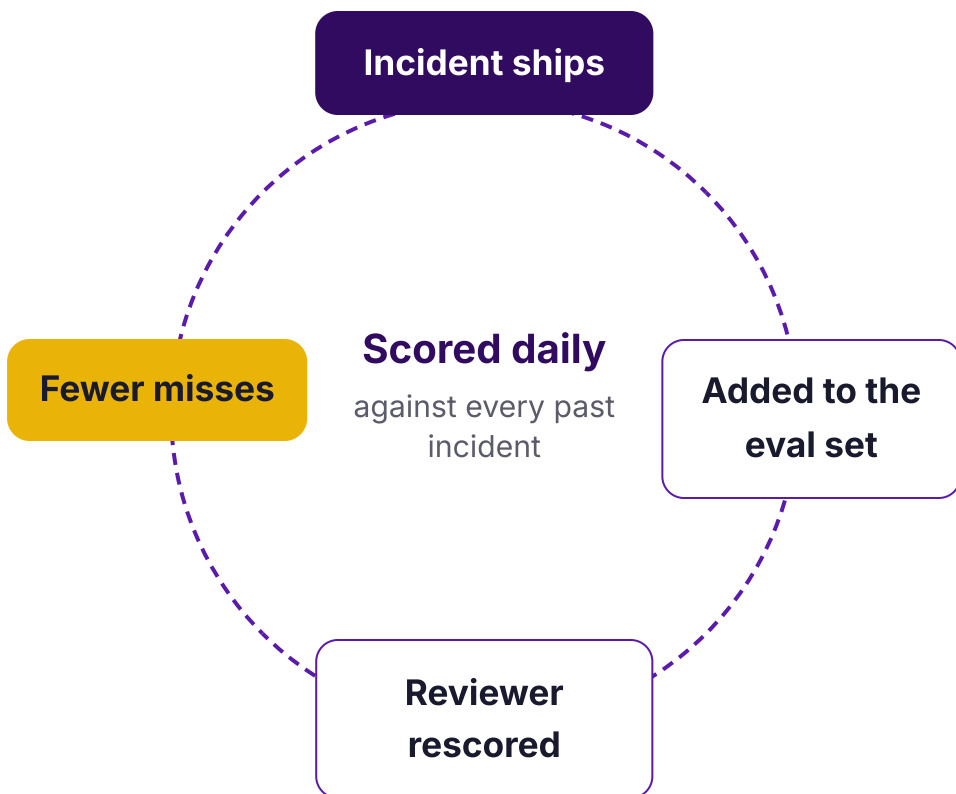
03

Name the bug class, so all of it gets fixed

04

Never re-report a known issue

Every incident becomes a test.



Noise is measured too: how people handle each comment

Fixed

Rejected

Intentional

Didn't matter

What people still own.

01 **Should we build this?**

Vision, taste, and goals

02 **Dangerous code**

Billing, auto-updater, CI workflows

03 **Architecture**

It sets the conventions others follow

04 **Checking the proof**

Test output, commands run, screenshots

You land it, you own it. A person answers for every change.

THE TAKEAWAY

Review is now a system, **not a person.**

Layers of checks, risk-based approval, a loop that converges, and incidents as the eval. Every team letting agents open PRs needs this.

One Company Brain. **Always on.**

AI Employees for Product, Engineering & Analytics.



[Learn more · xhawk.ai](https://xhawk.ai) →

Source: Anthropic, "How the Claude Code team does code review," shared by Addy Osmani on X.