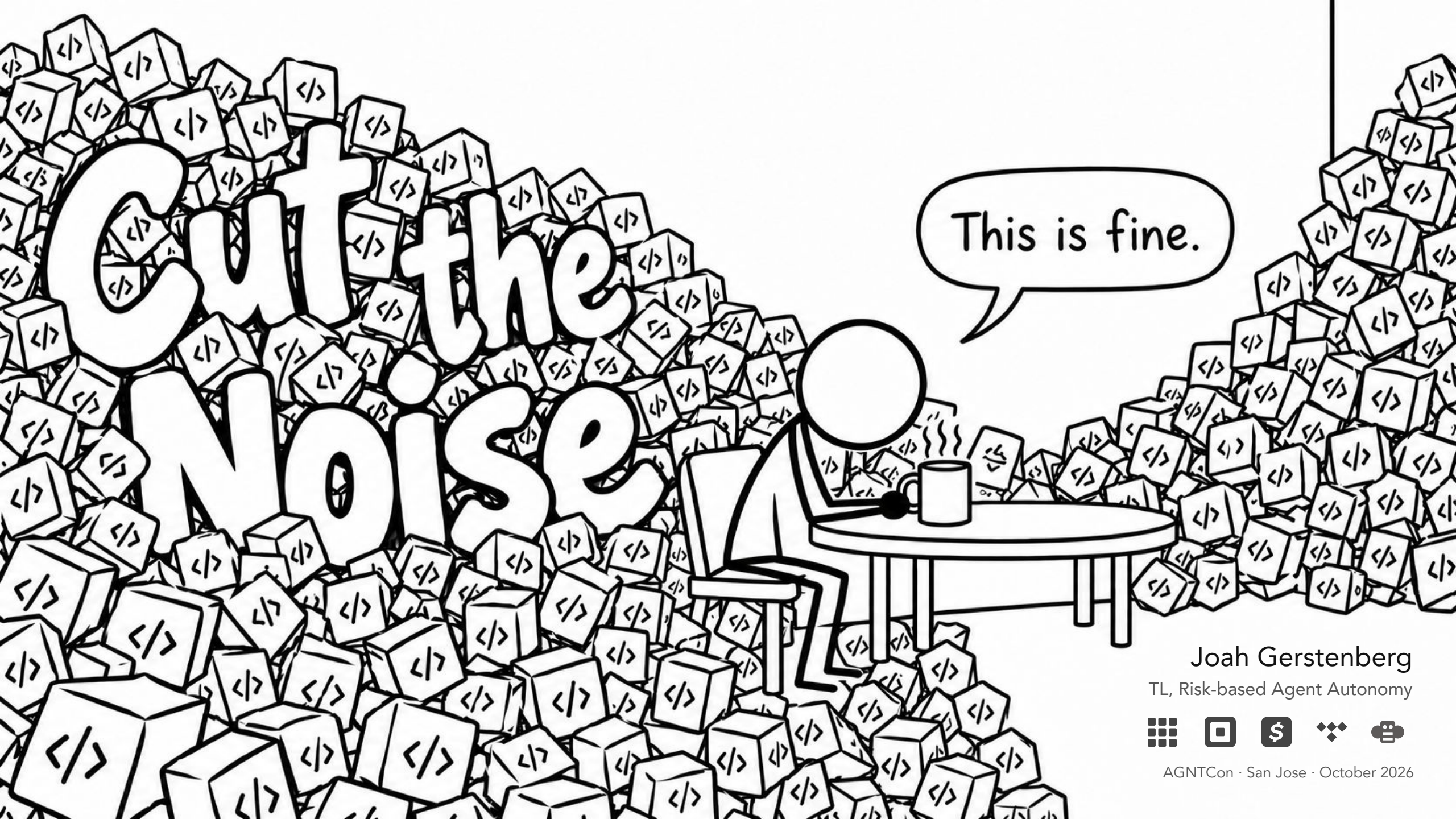


#ai-champions ▾



joah Nov 12th, 2025 at 5:47 PM

[@Angie Jones](#) We need better models for
code review!

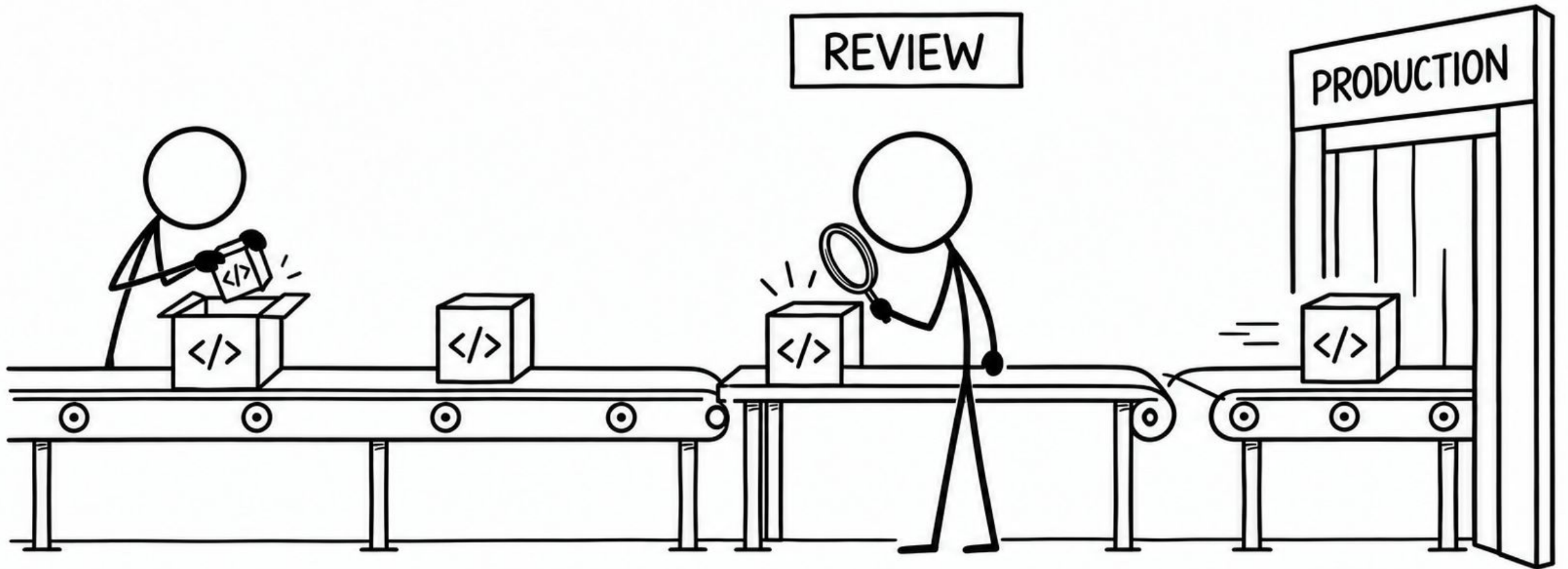


Joah Gerstenberg

TL, Risk-based Agent Autonomy



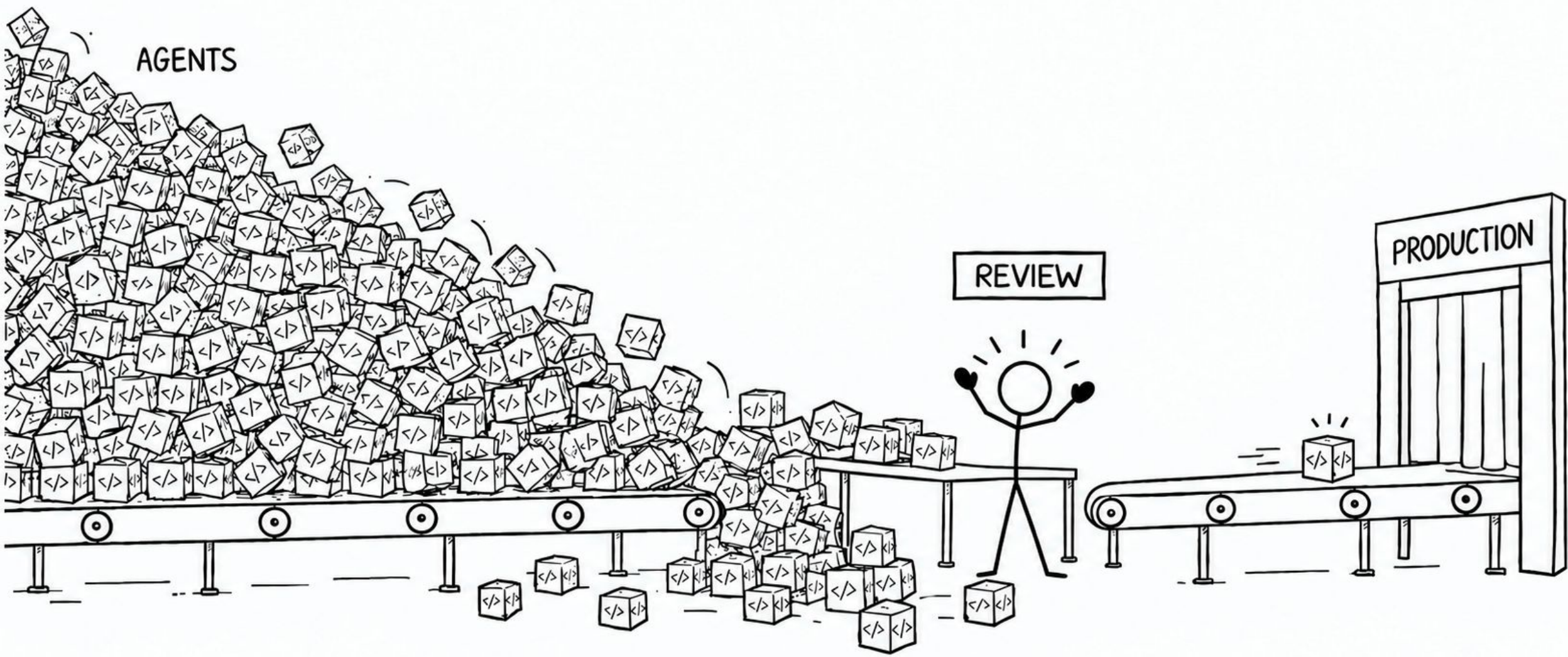
AGNTCon · San Jose · October 2026

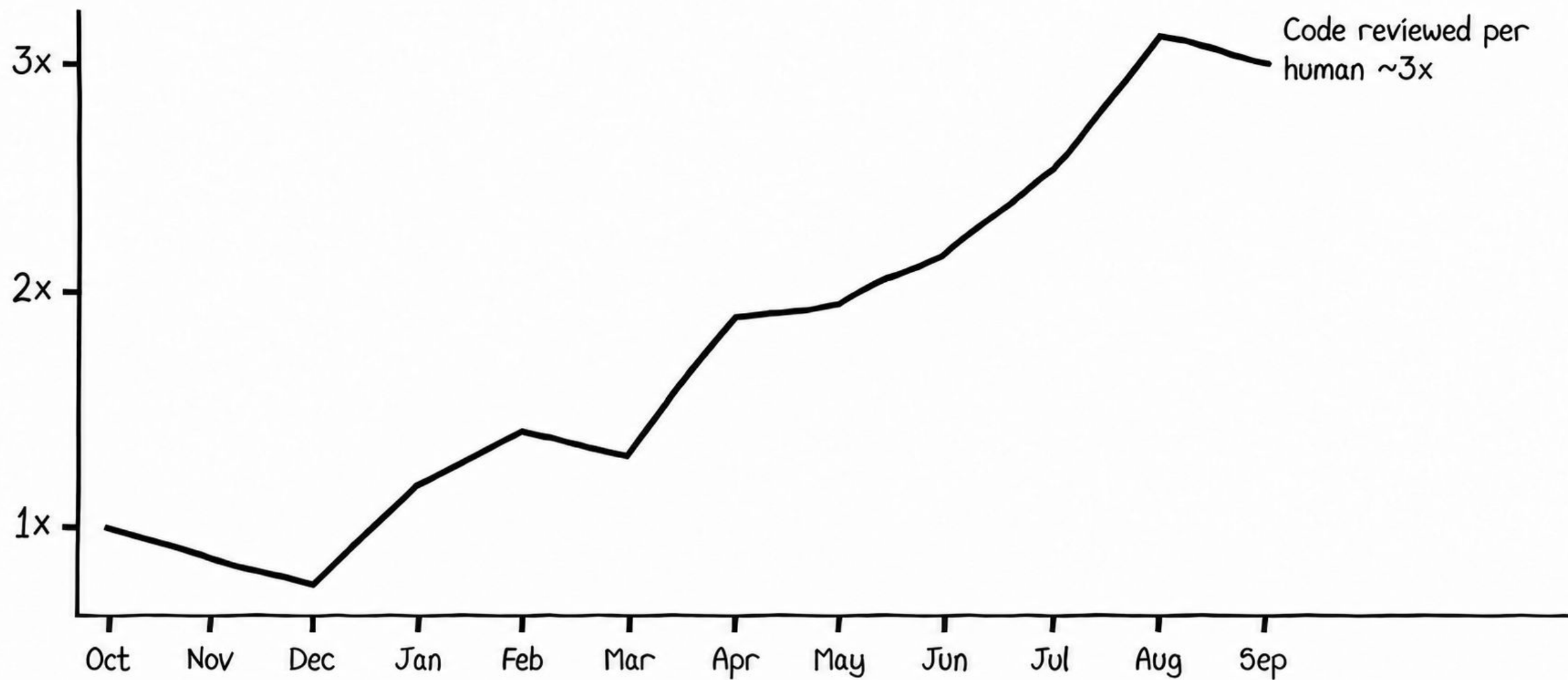


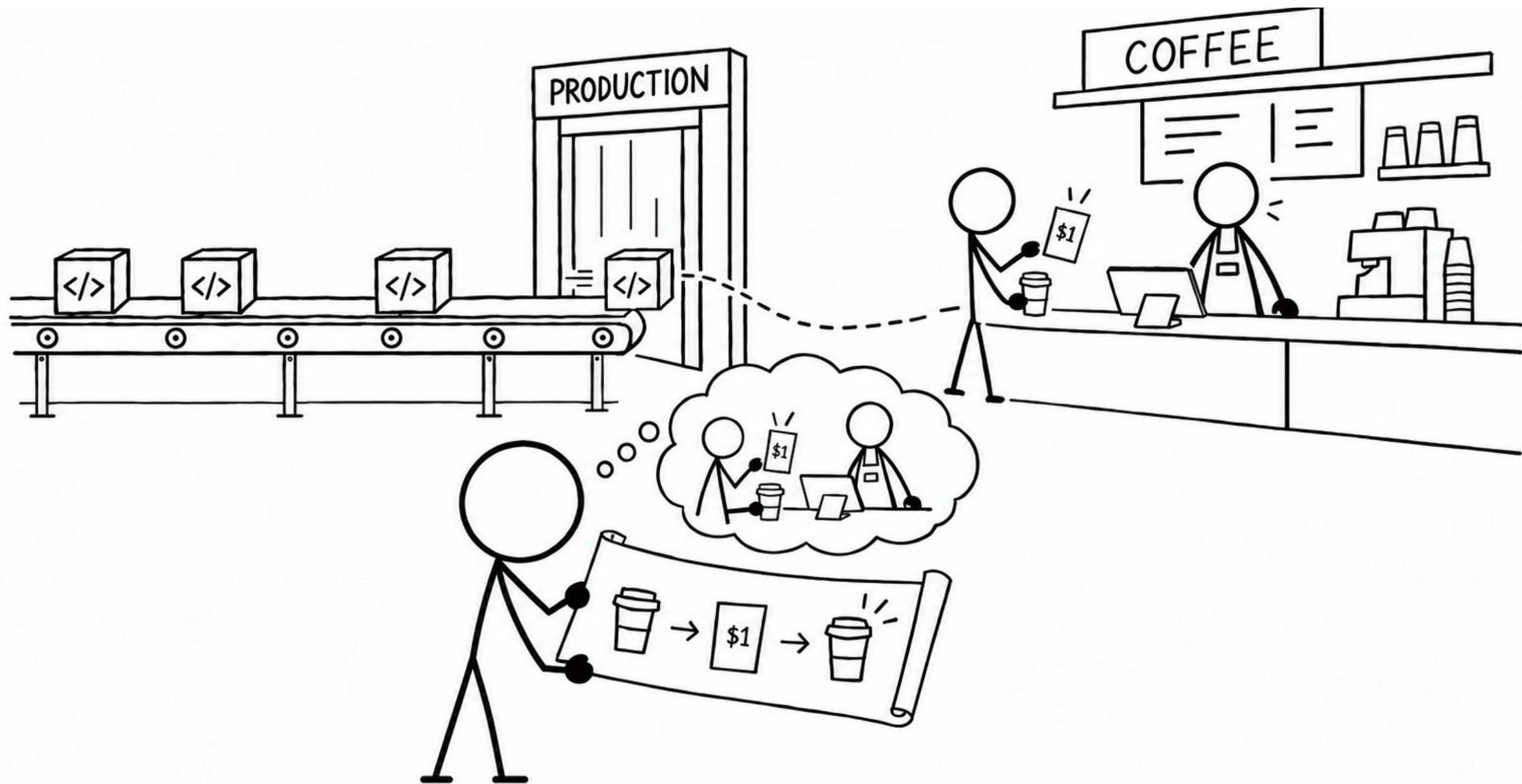
AGENTS

REVIEW

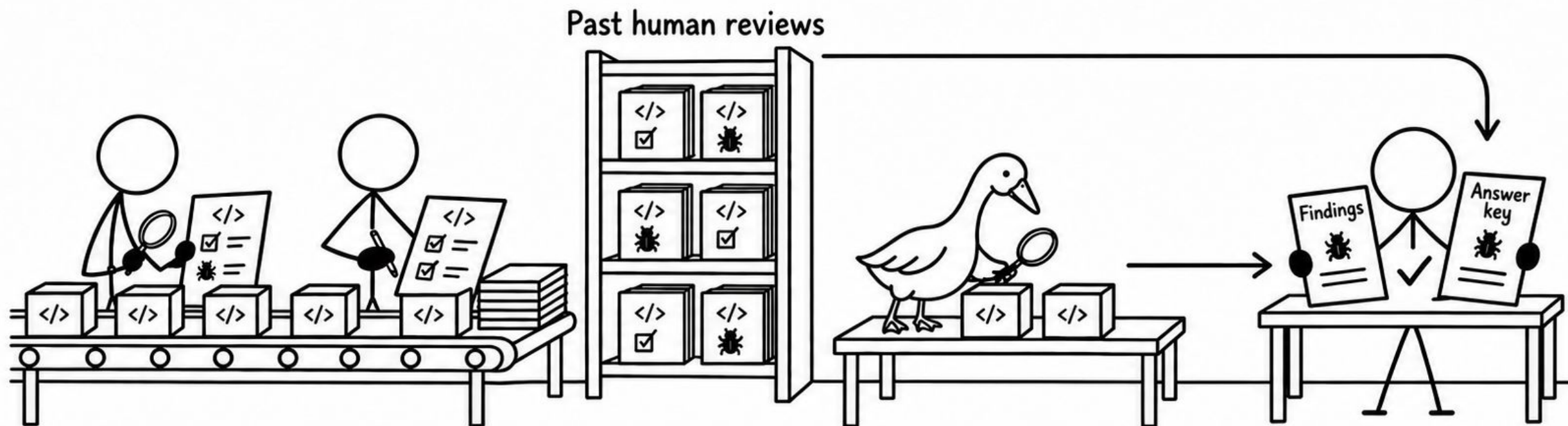
PRODUCTION











Calibrate your reviewer to past human reviews

▼ MODULES.yaml

```
5 + auto_assign:
6 +   round_robin: 2
7 +   except: ██████████
8 + exclude_author_from_core_reviewers: true
```



builderbot-code-review Bot on Apr 29



[untrusted-pr] exclude_author_from_core_reviewers: true allows PR authors to be excluded from the required core reviewer group



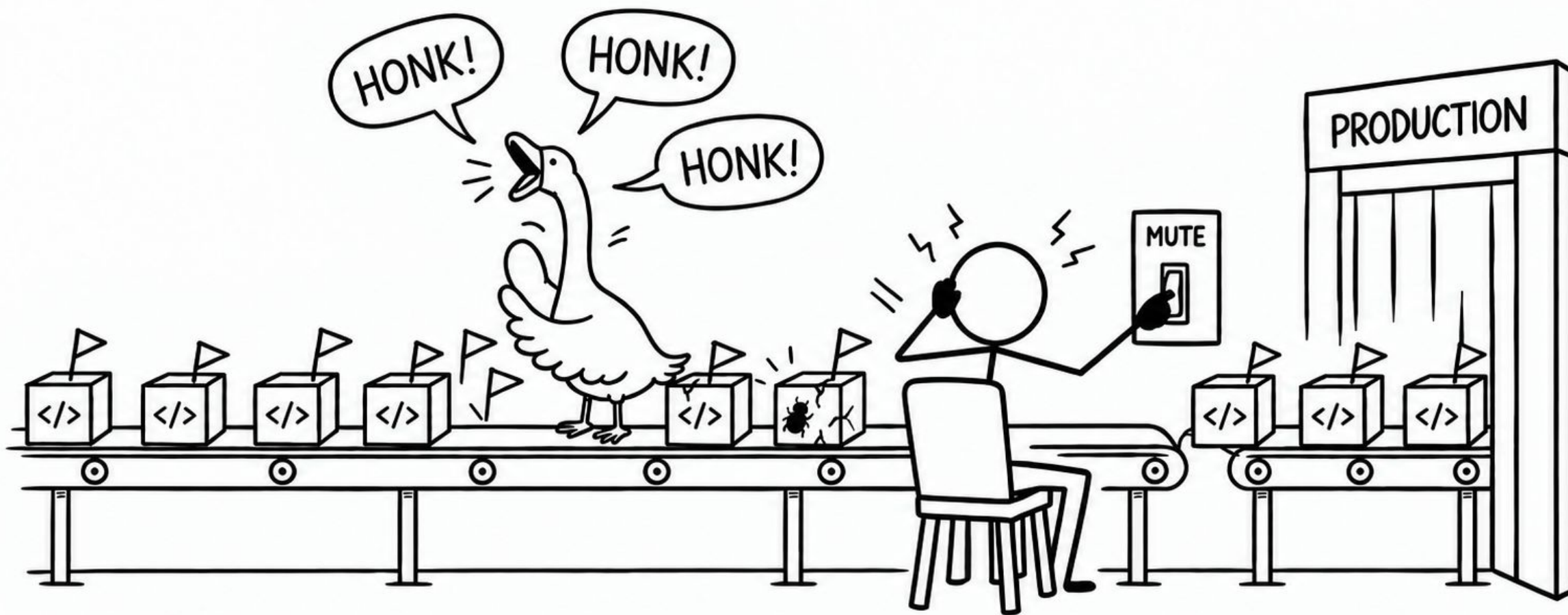
██████████ on Apr 29

Author



This is the worst comment I've ever seen





A reviewer that's muted catches nothing

OFFLINE EVALS

Compare findings
with an answer key

ONLINE EVALS

Observe what happens
in production

PRECISION

Valid findings

All findings

RECALL

Known defects found

All known defects

ADDRESS RATE

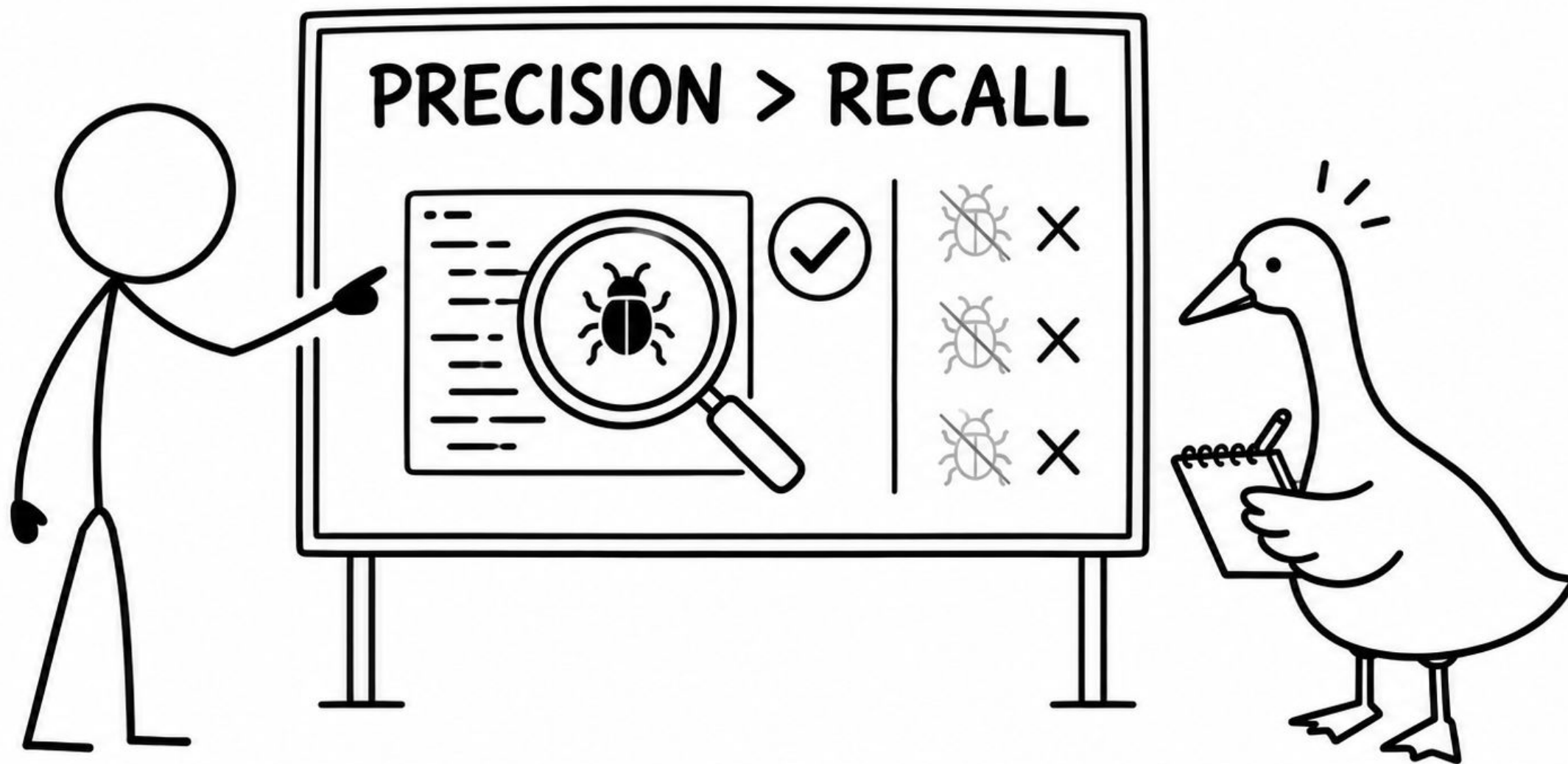
Findings acted on

All settled findings

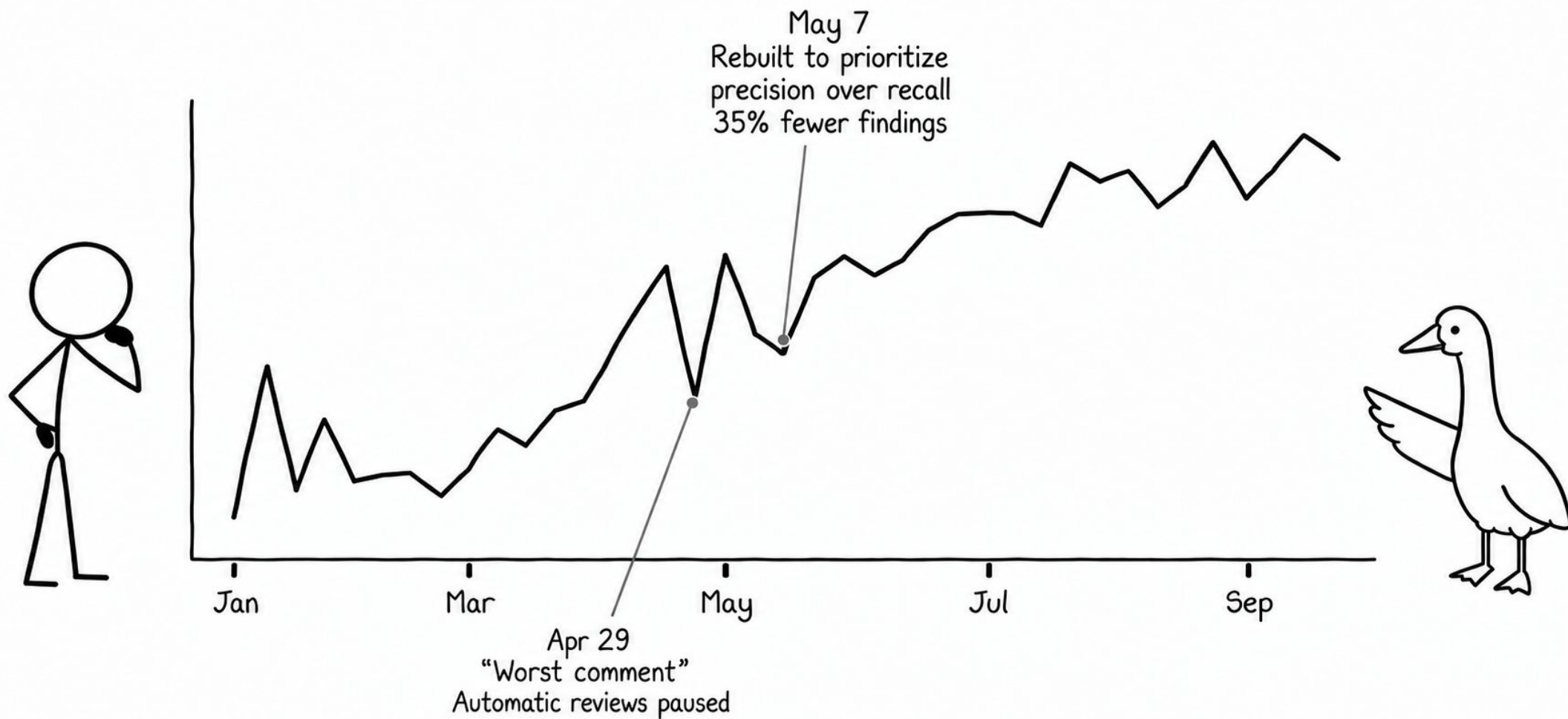
CHANGE FAIL RATE

Changes from findings
later fixed or reverted

All changes made
from findings



Prioritize precision over recall



Online evals: address rate



review-bench.ai



REVIEWBENCH

Research preview

[Leaderboard](#)

[Methodology](#)

[Submit an agent](#)



ReviewBench: An open benchmark for AI code review.

Measuring how well AI agents find issues in real-world pull requests against human-validated ground truth. [Run the benchmark](#) yourself with your own reviewer and compare the results.

Leaderboard

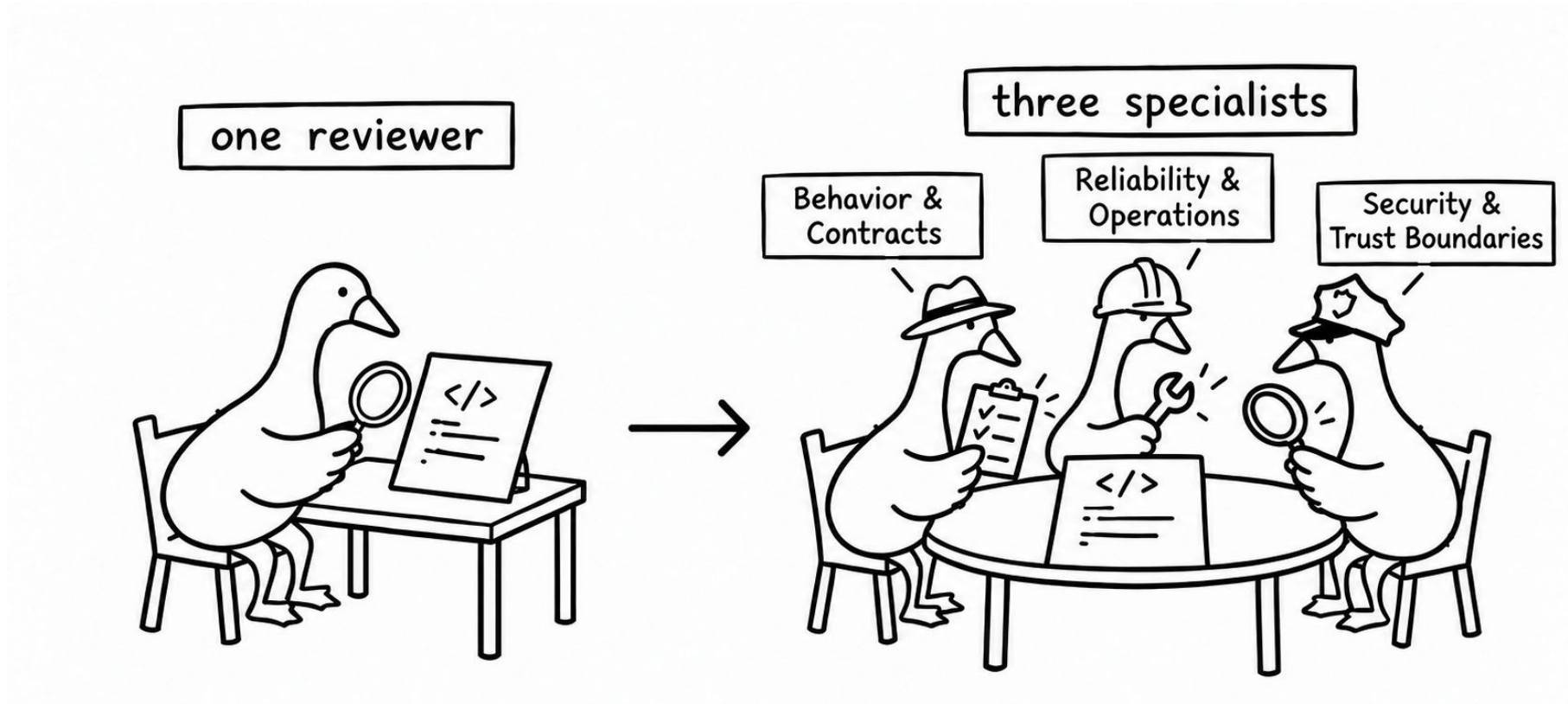
Filters

Grounded vs. augmented

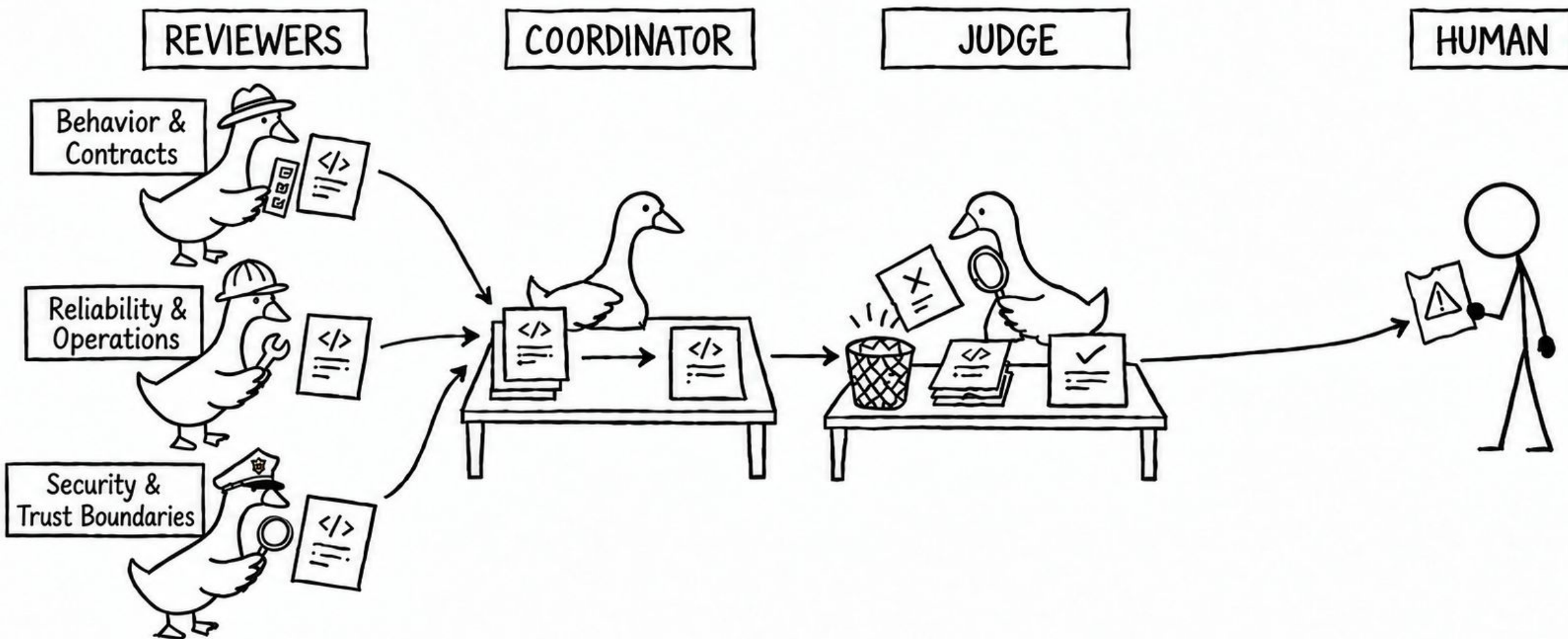
[Compare agents](#)

Rank	Agent	Grounded F1	Grounded Precision	Grounded Recall	Augmented F1	Augmented Precision	Augmented Recall	Duration	Da
1	Copilot Code Review Balanced	40.1%	87.8% ± 0.5	26.0% ± 0.6	49.7%	87.4% ± 2.0	34.7% ± 0.2	4m 43s ± 18s	Oct 1, 4



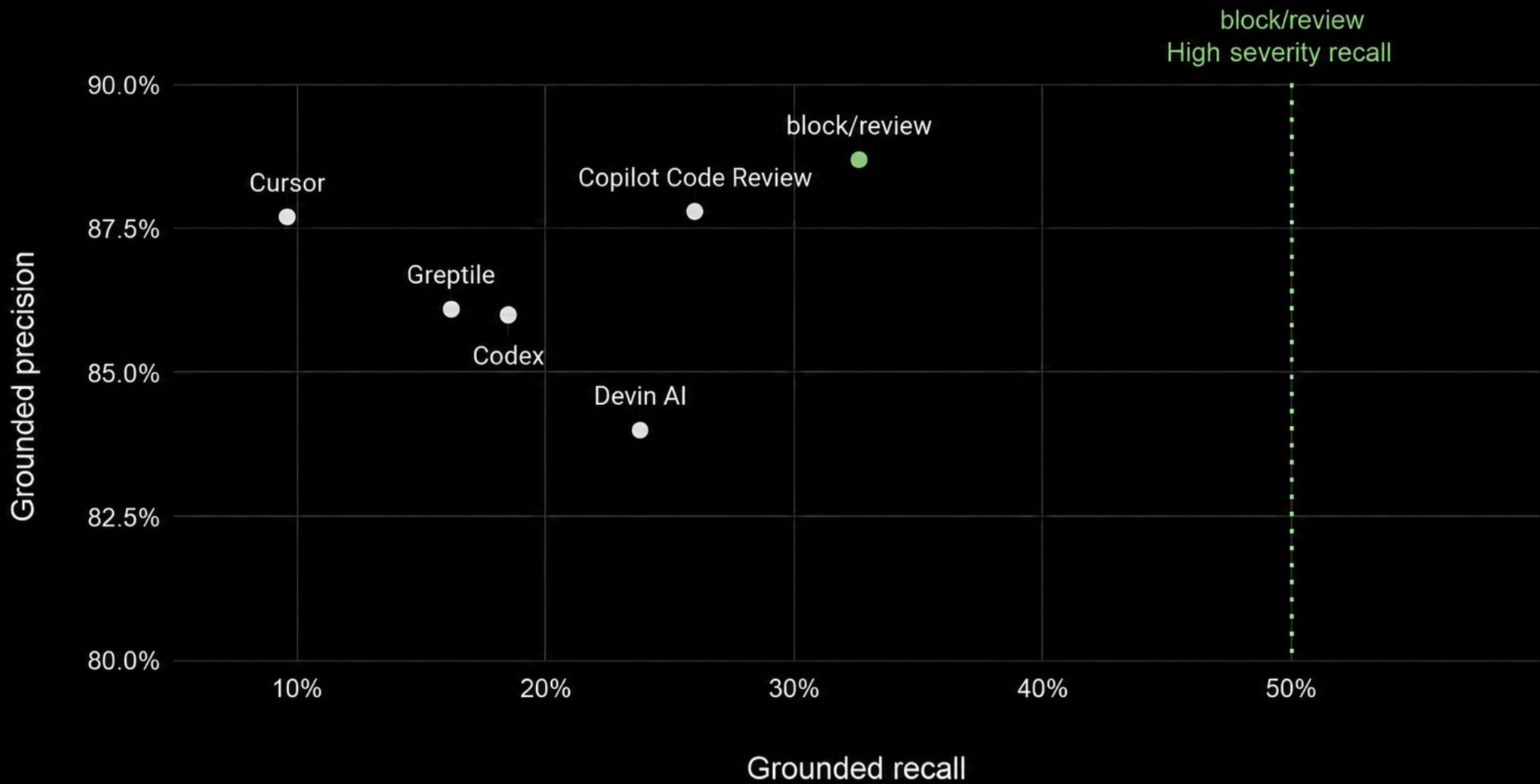


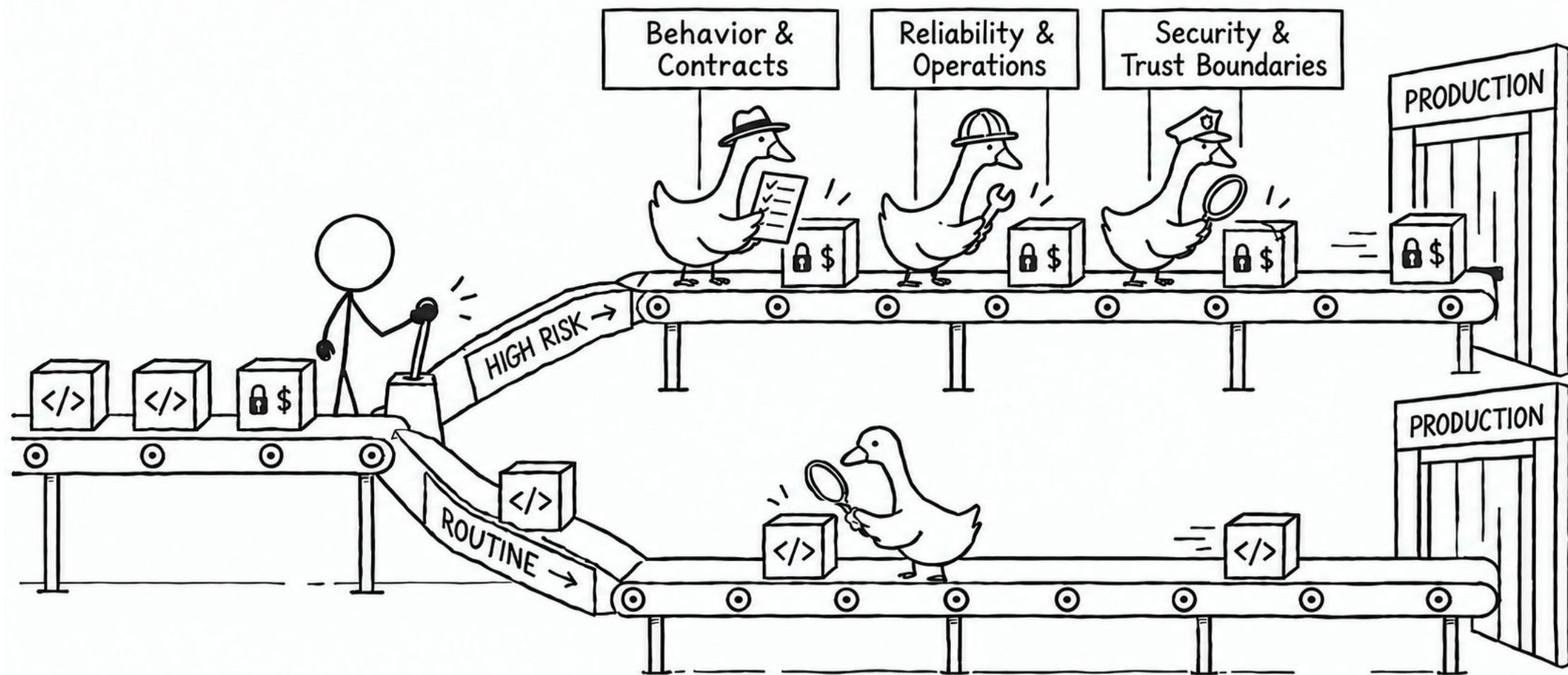
Three specialists improved recall by 11% over one reviewer



More layers increase precision, more reviewers increase recall

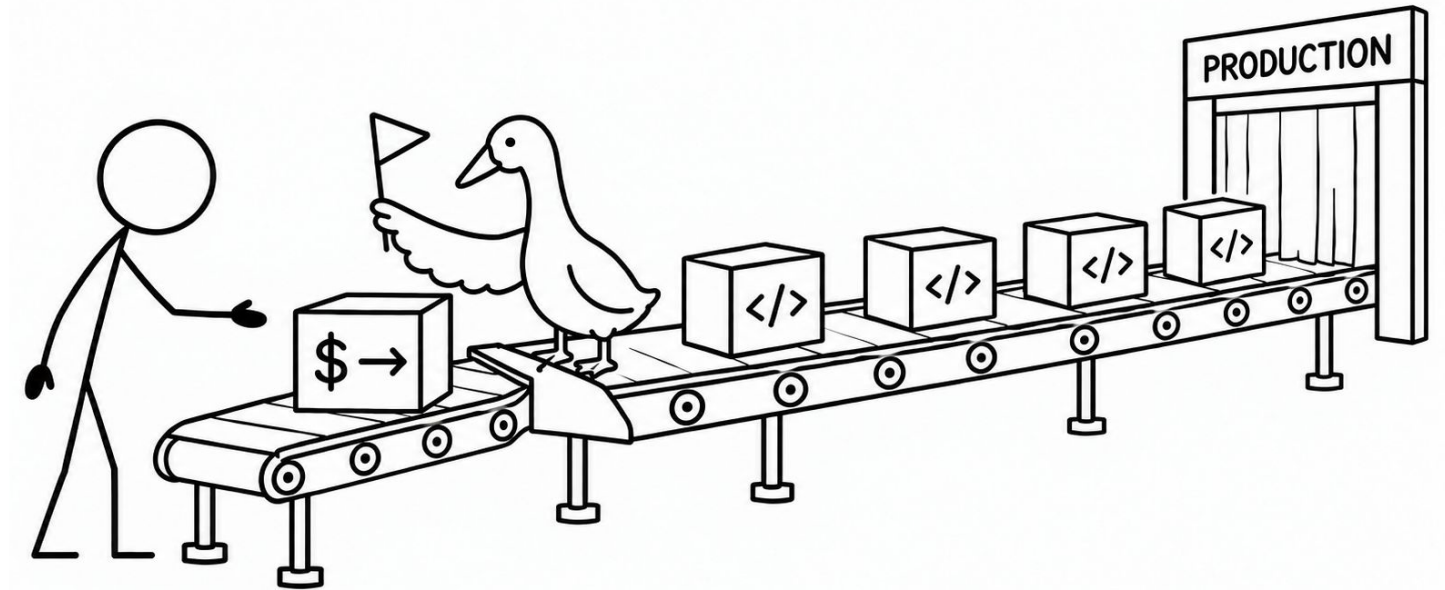
Grounded precision and recall



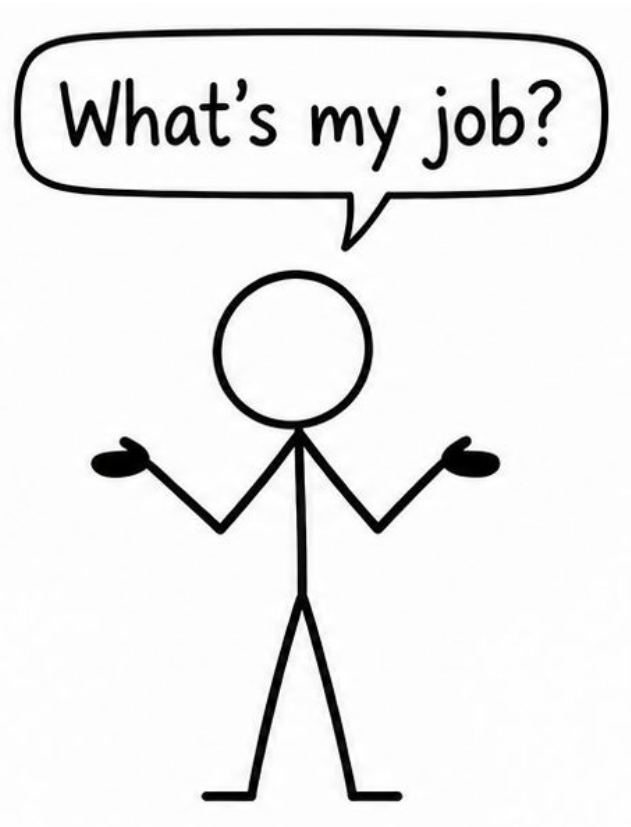


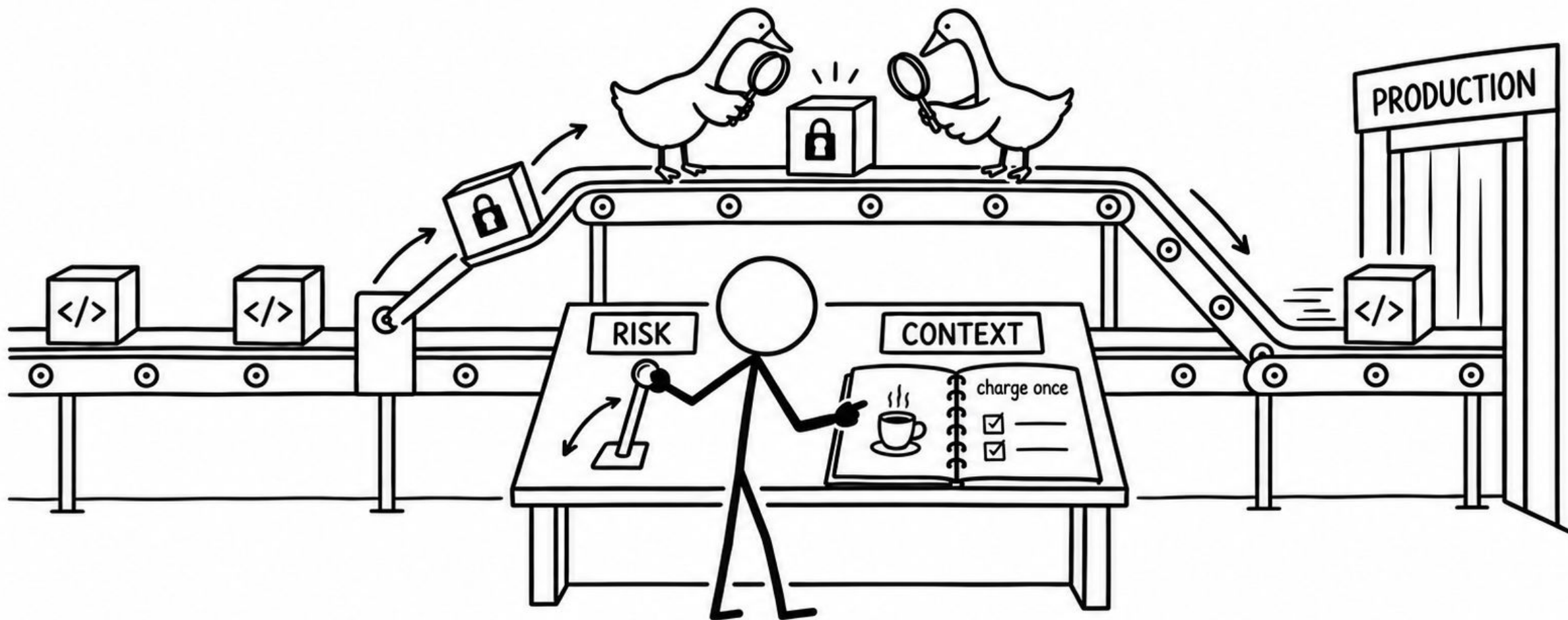
Risk-based review tiering

- Irreversible changes
- Public product claims
- SOX compliance
- Money movement
- Personally identifying information

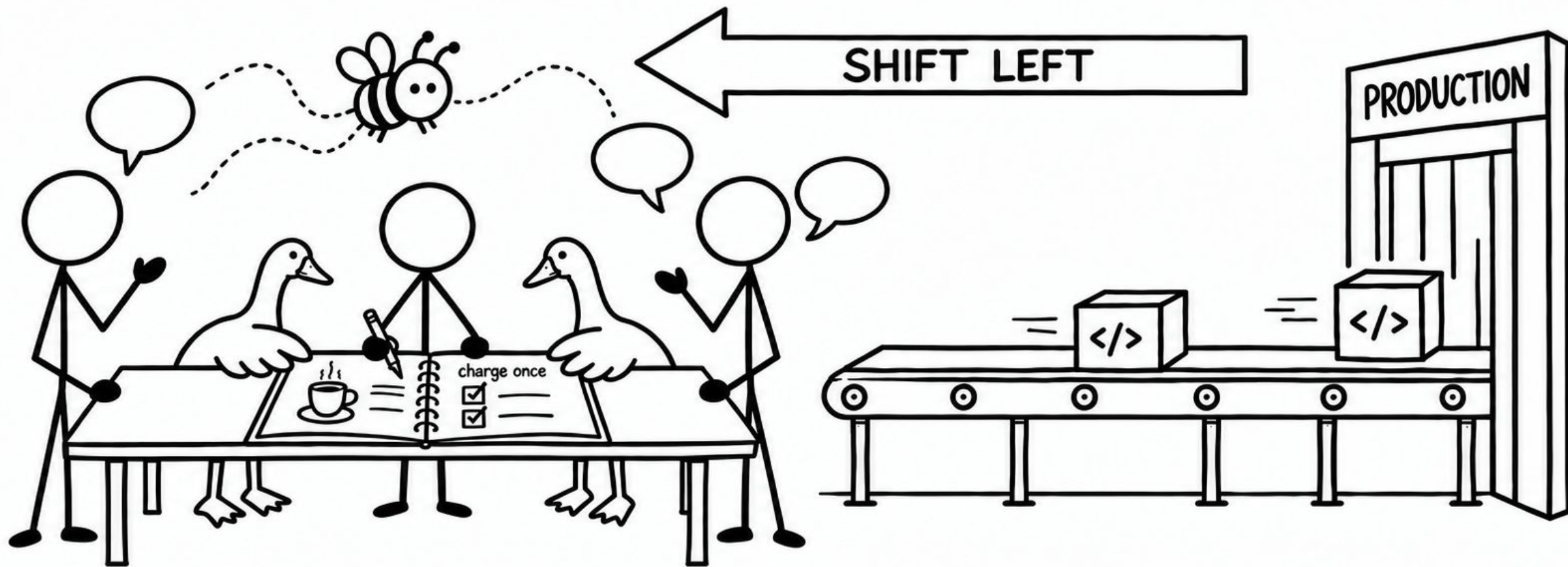


Humans only review the most sensitive changes





Align and contain



Get review before the code is written

Prioritize precision
over recall

Give your agents
the big picture

Build alignment
before the code is written

