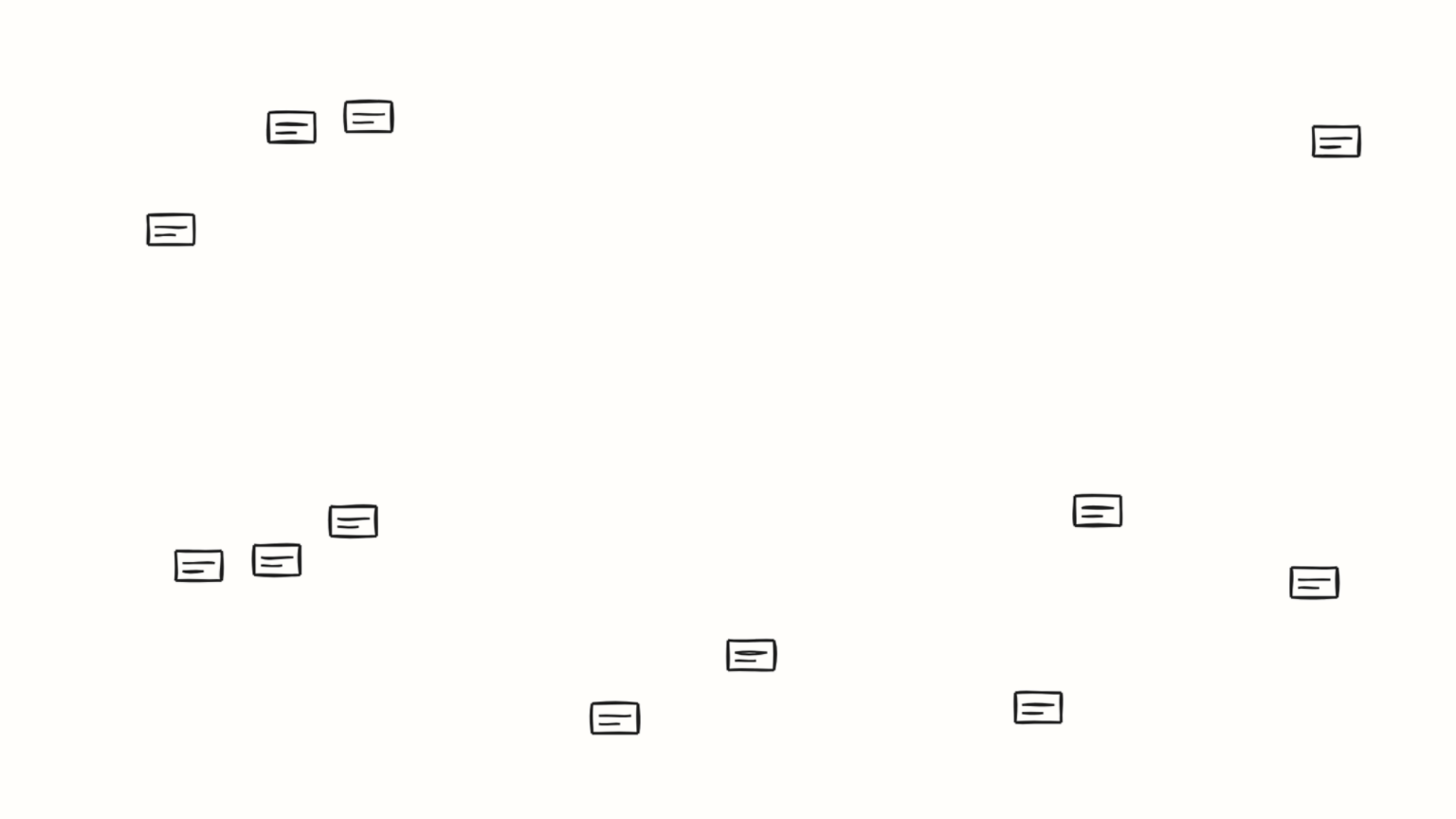
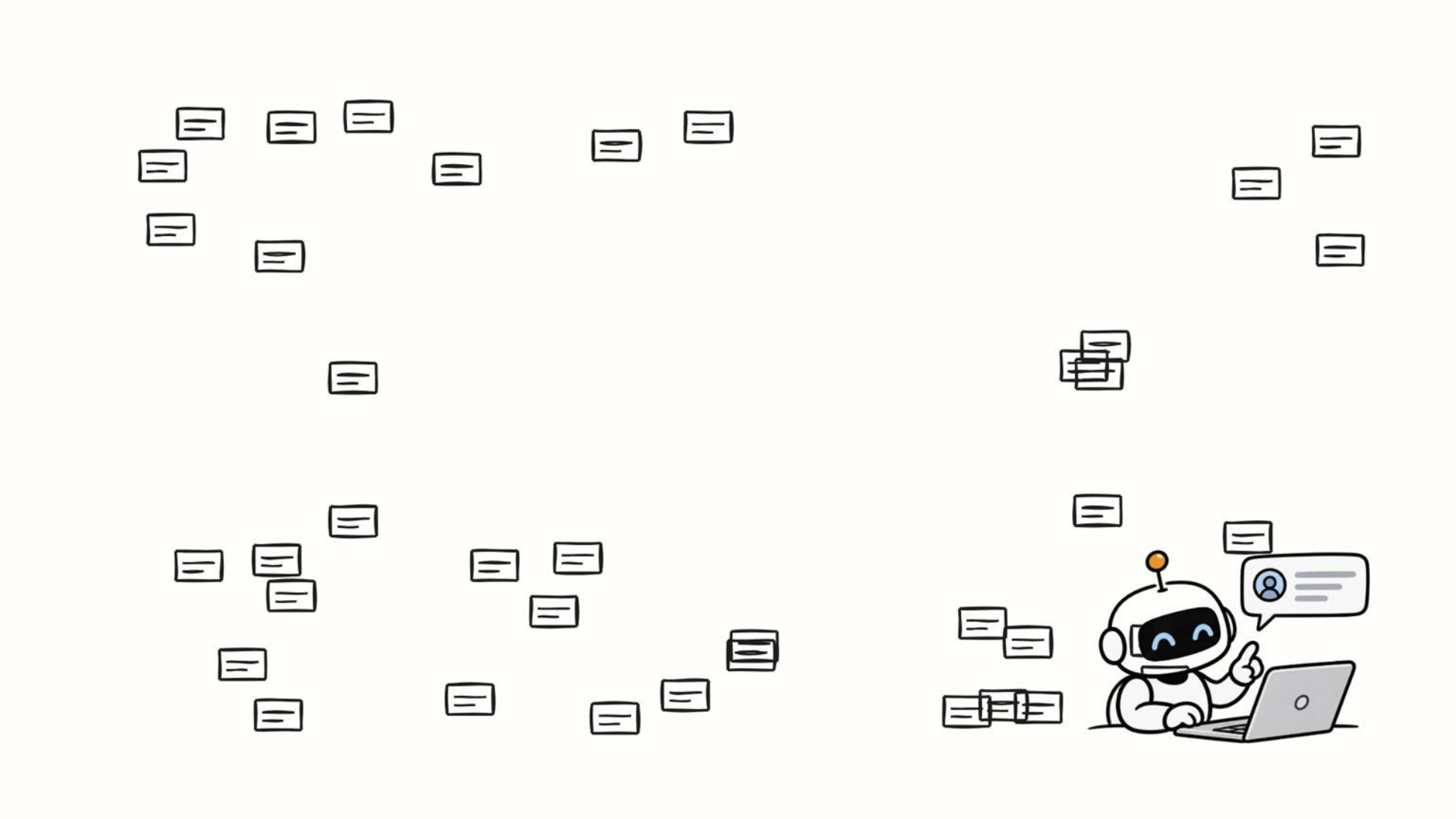
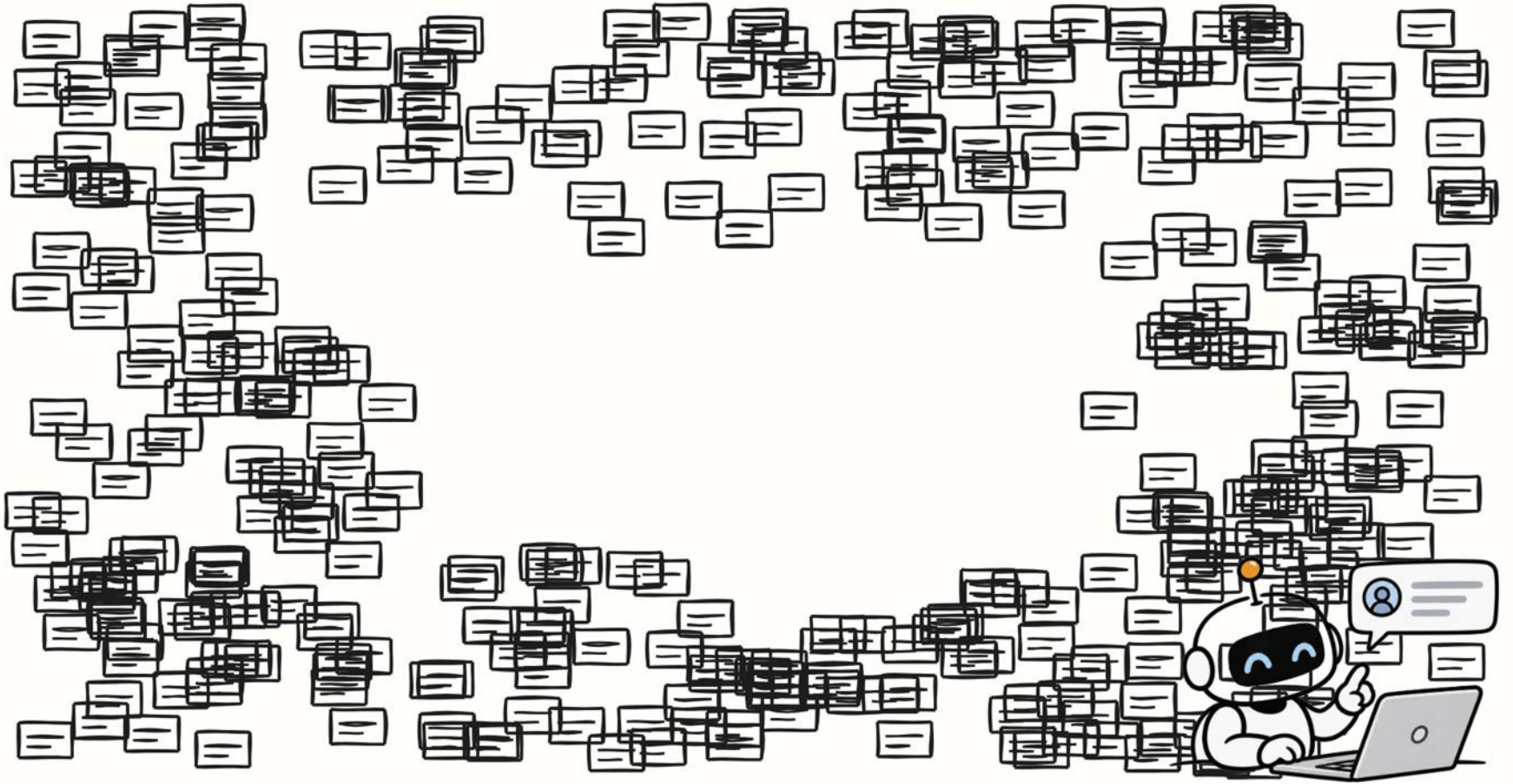
AI-Powered Code Review at Scale

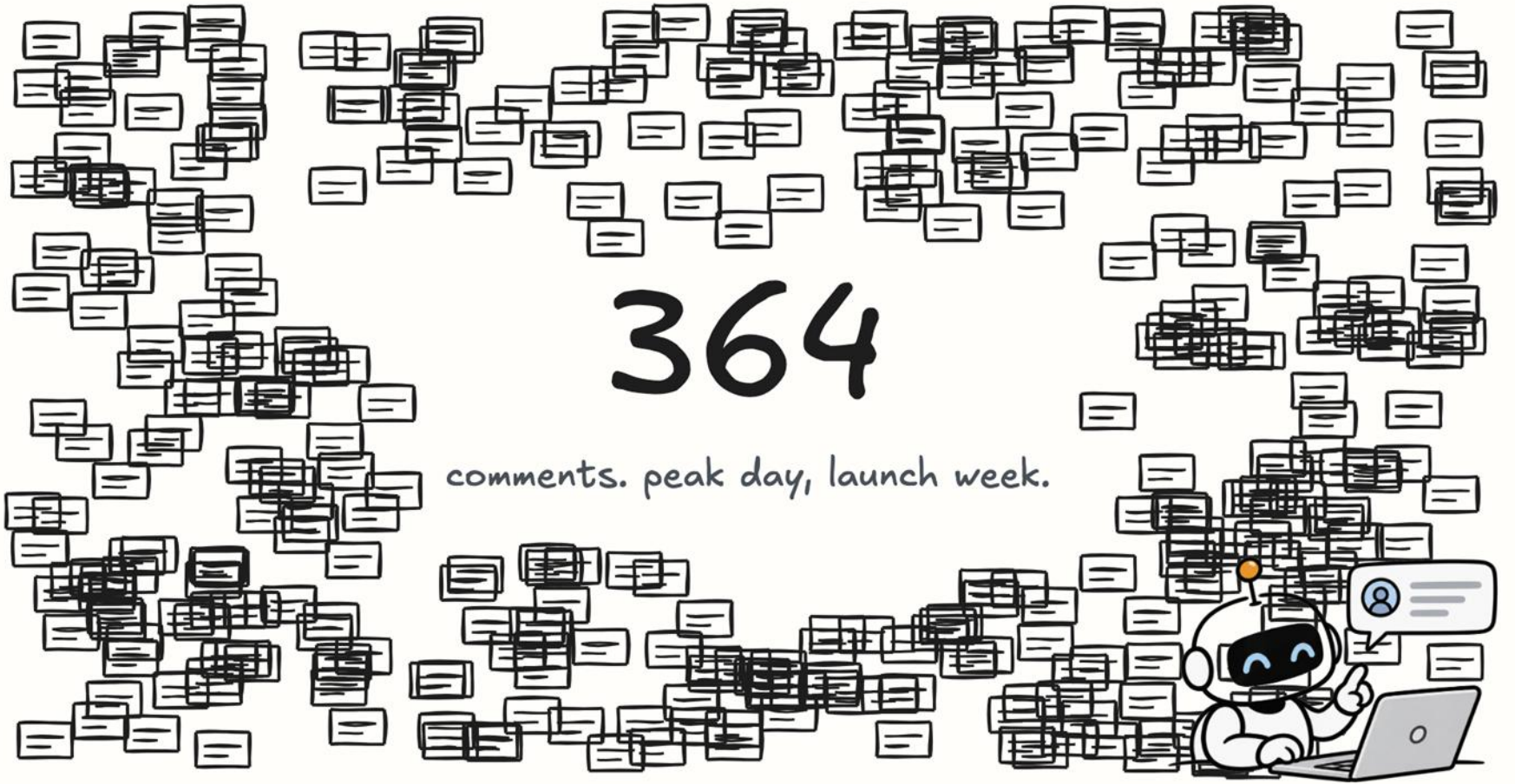


Serhii Yakovenko · LeadDev Berlin 2026








364
comments. peak day, launch week.



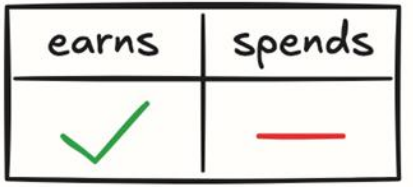
the team's reaction, week one

week two:



correct ≠ worth reading

correct ≠ worth reading





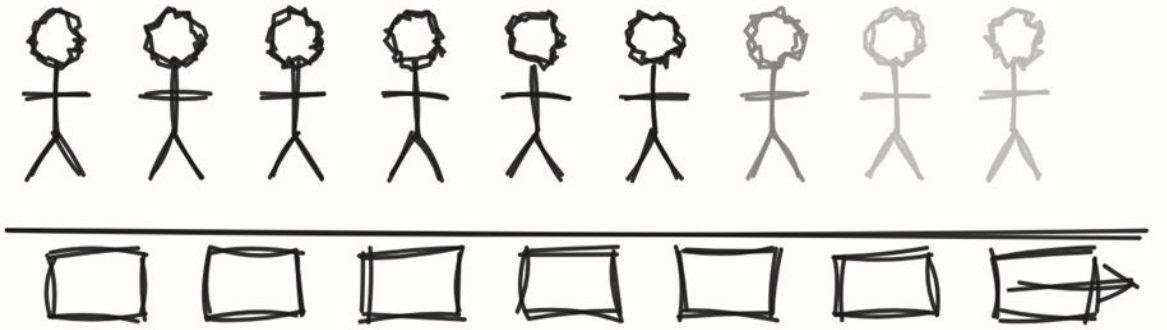
Serhii Yakovenko

engineering manager

platform · DevEx · AI tooling

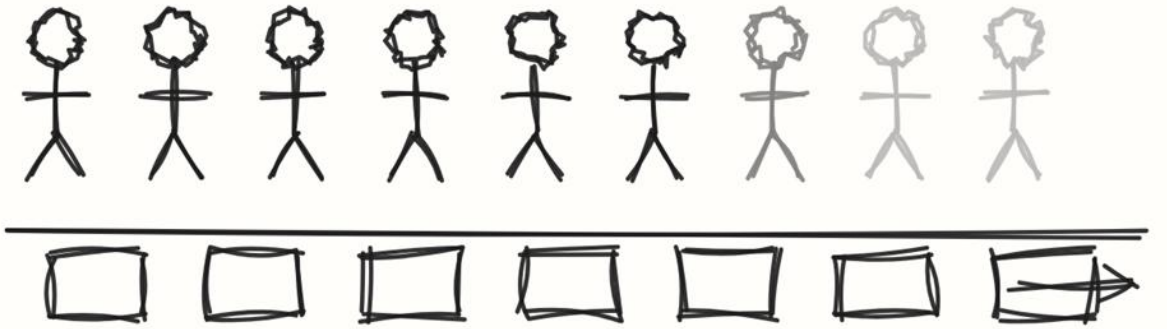
chapter 1

what we built



~200 developers

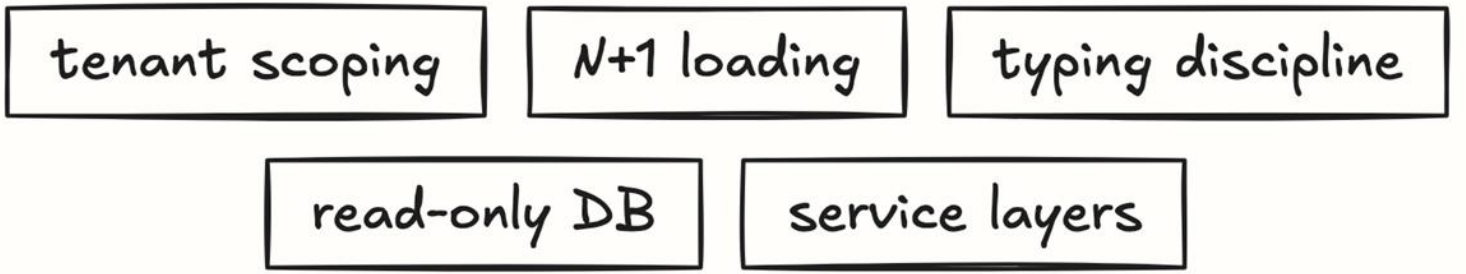
~50 MRs merged a day



~200 developers

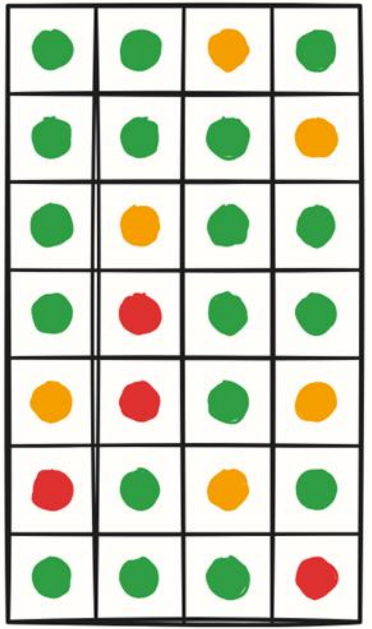
~50 MRs merged a day

1/3 of a senior's week, on the same comments

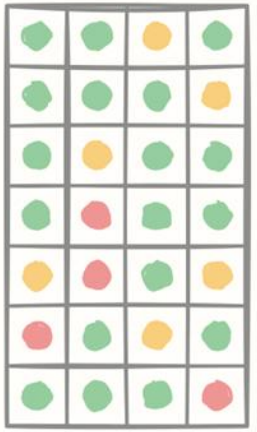


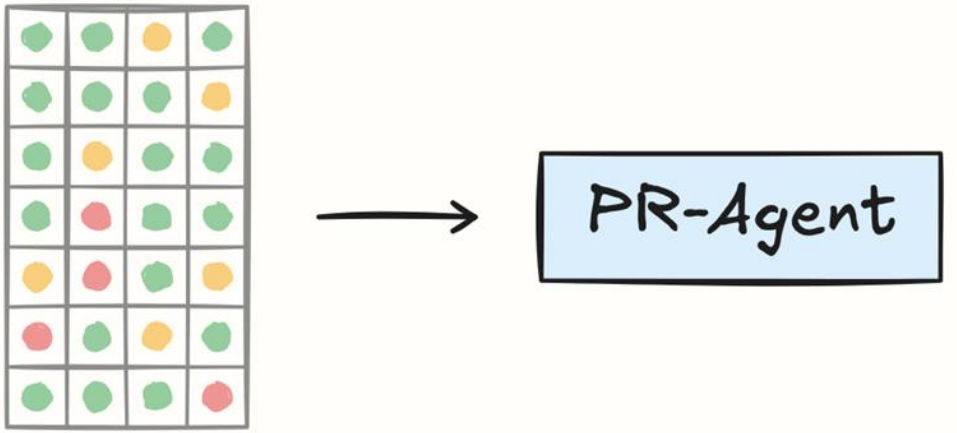
> half of all review comments

4 tools · 18 dimensions

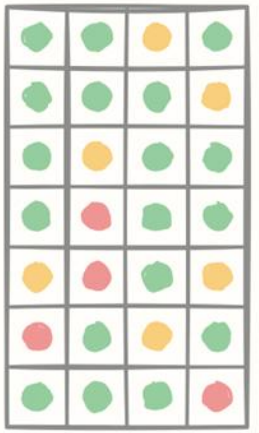


"We recommend Greptile"





we shipped the other one



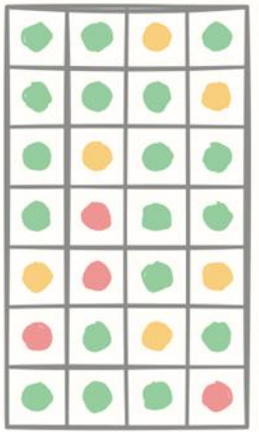
PR-Agent

we shipped the other one

self-hosted

model freedom

ownable



PR-Agent

we shipped the other one

self-hosted

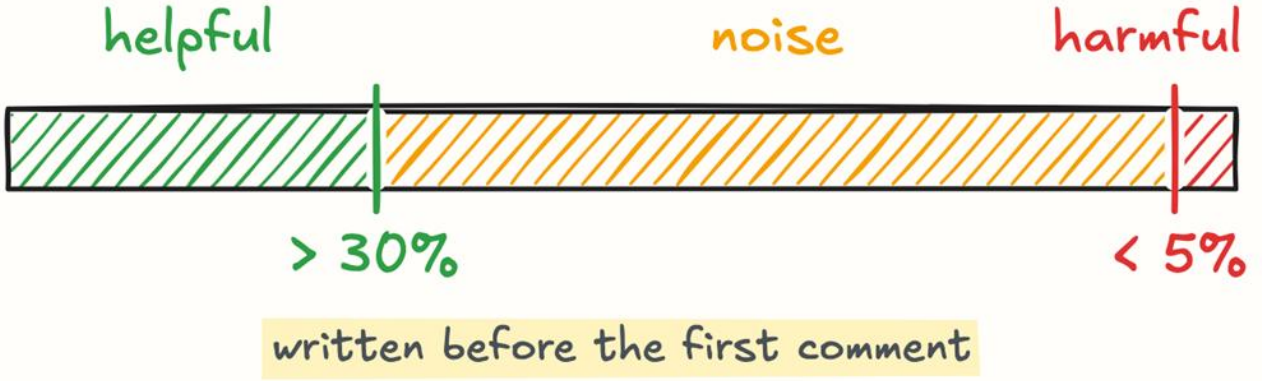
model freedom

ownable

~$300/mo vs $3.3-6.9K/mo quotes







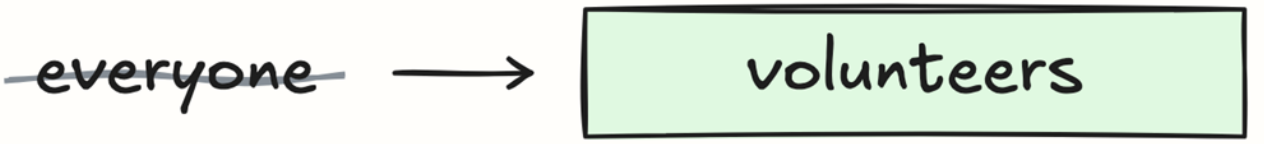
written before the first comment



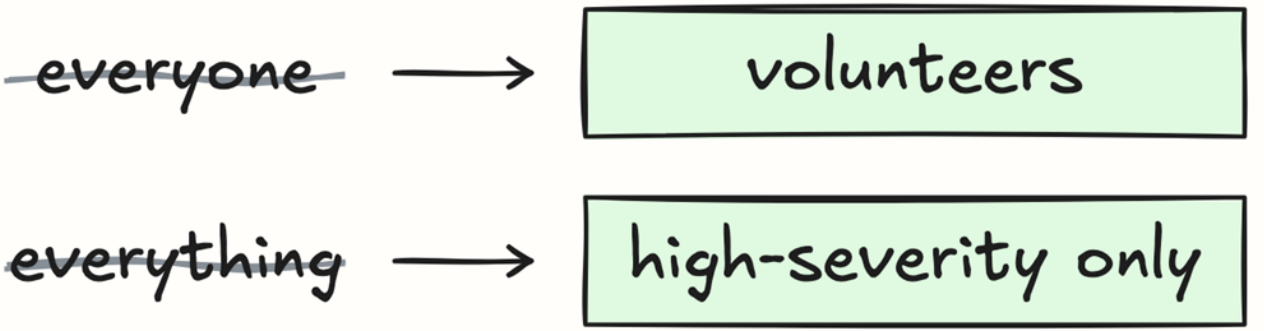
"the feedbacks were 50/50"

everyone

everything



everything


everyone
volunteers
everything
high-severity only

was: a CI job inside every pipeline



webhook → context → score gate → publish

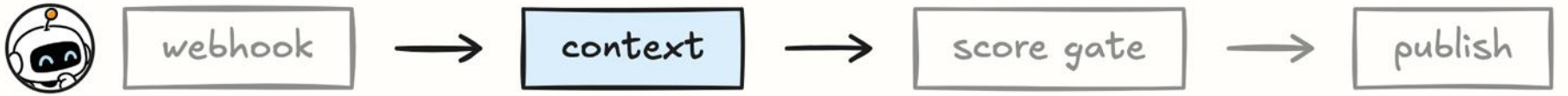
~~was: a CI job inside every pipeline~~

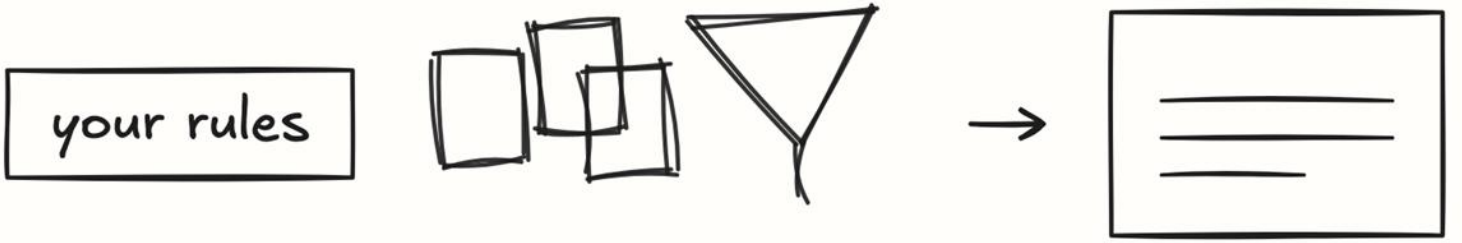
MR event
/review · /ask

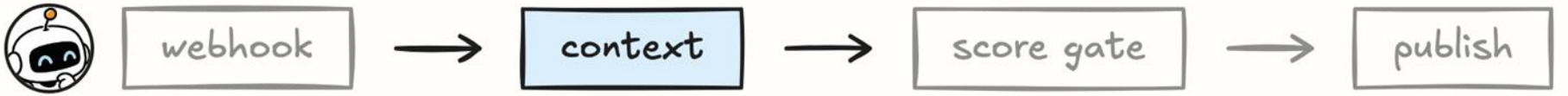


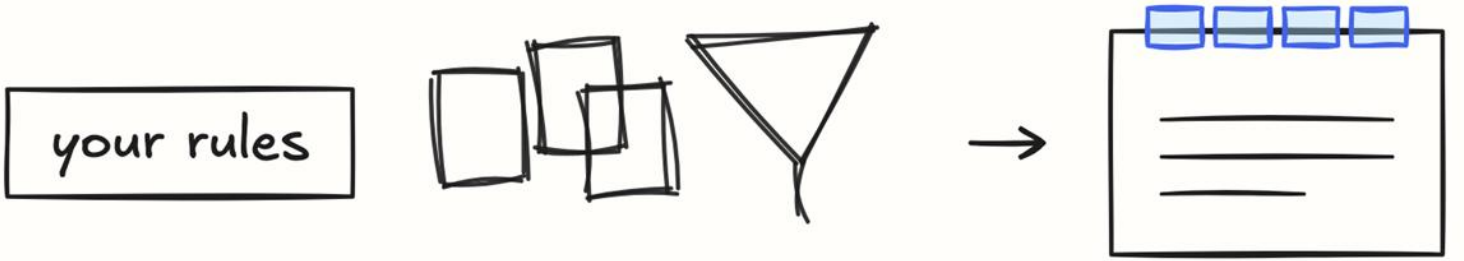
webhook → context → score gate → publish

push and pull, one pipeline

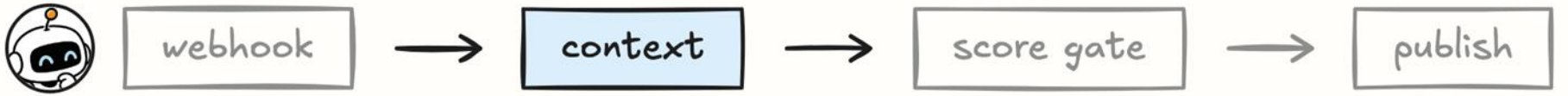

webhook
context
score gate
publish

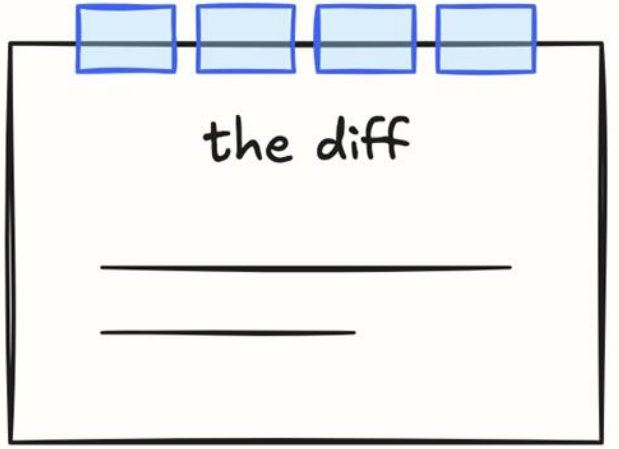

your rules



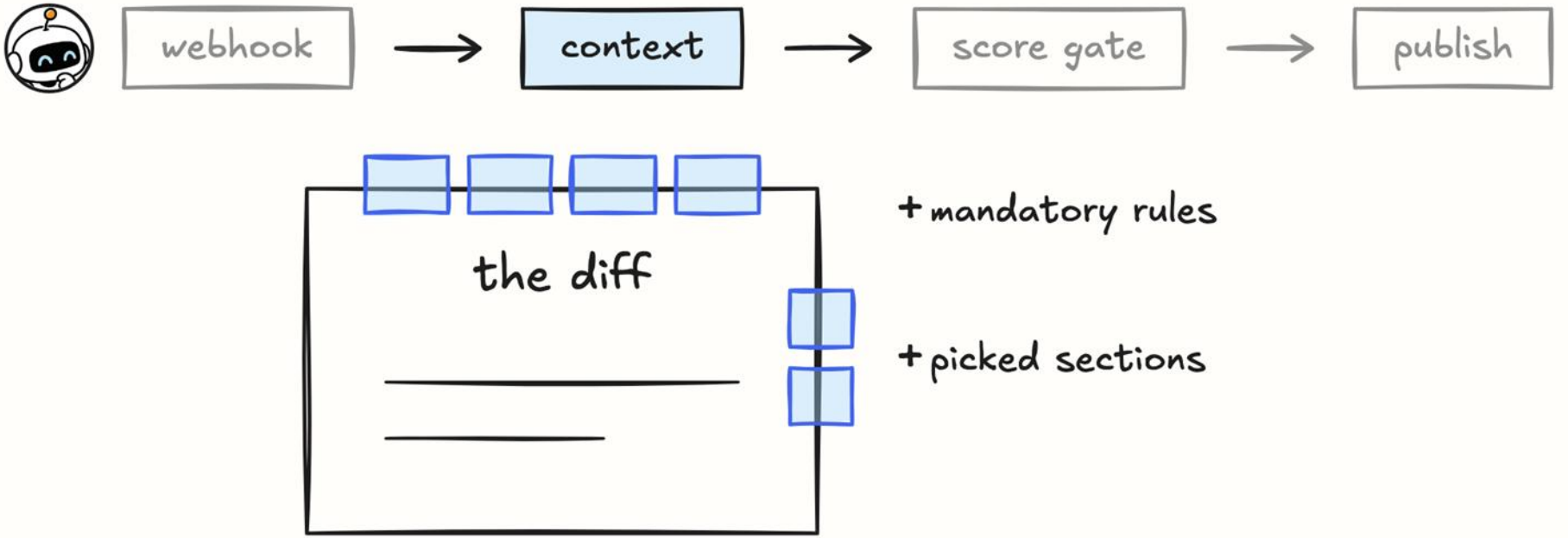


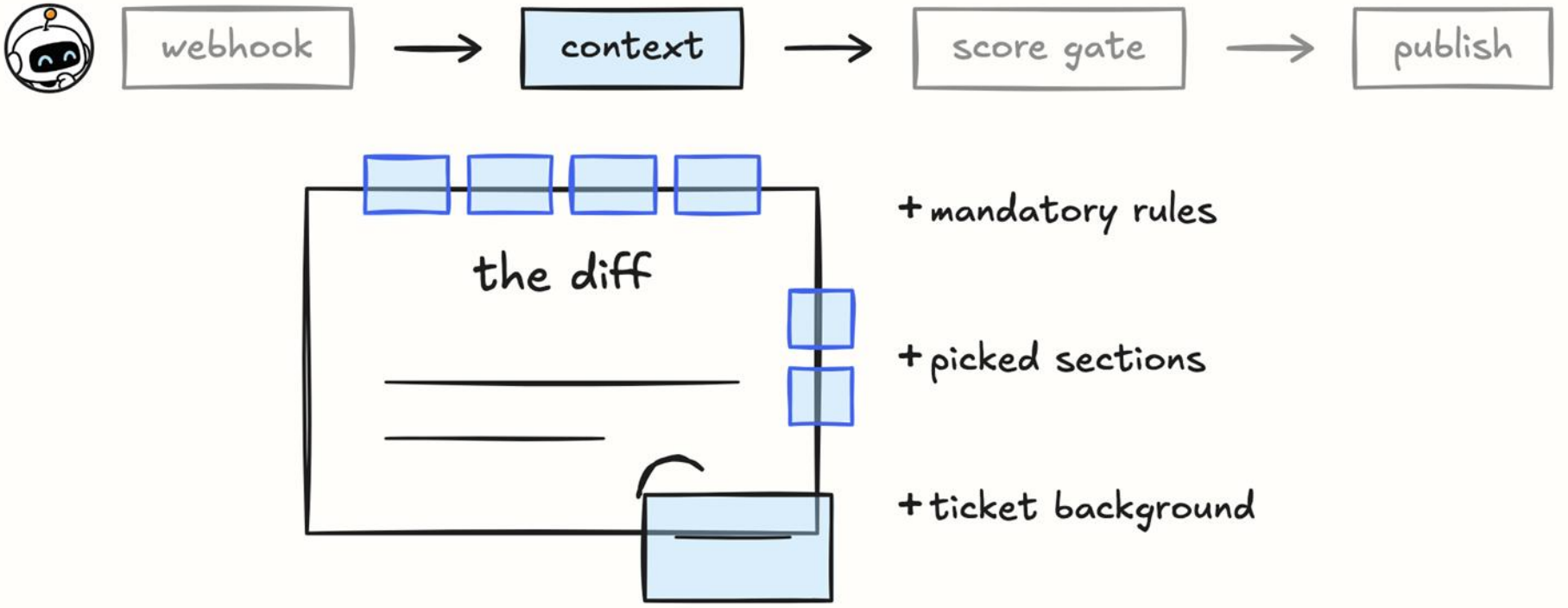
no vector DB

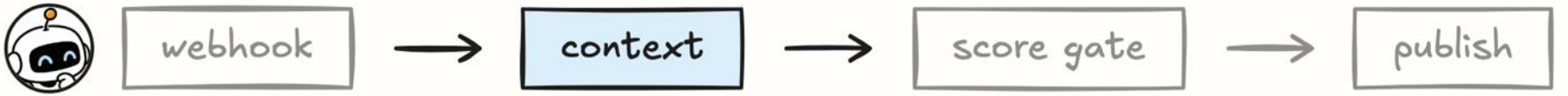


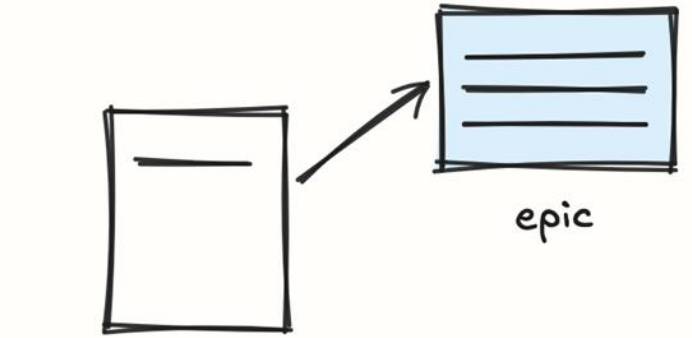


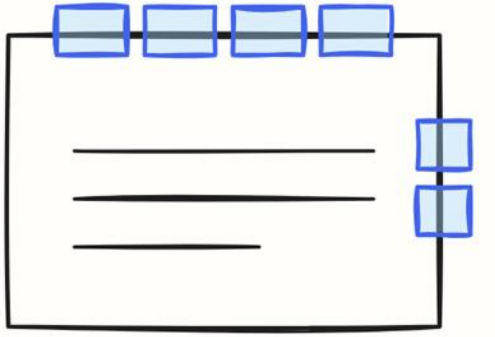
+ mandatory rules

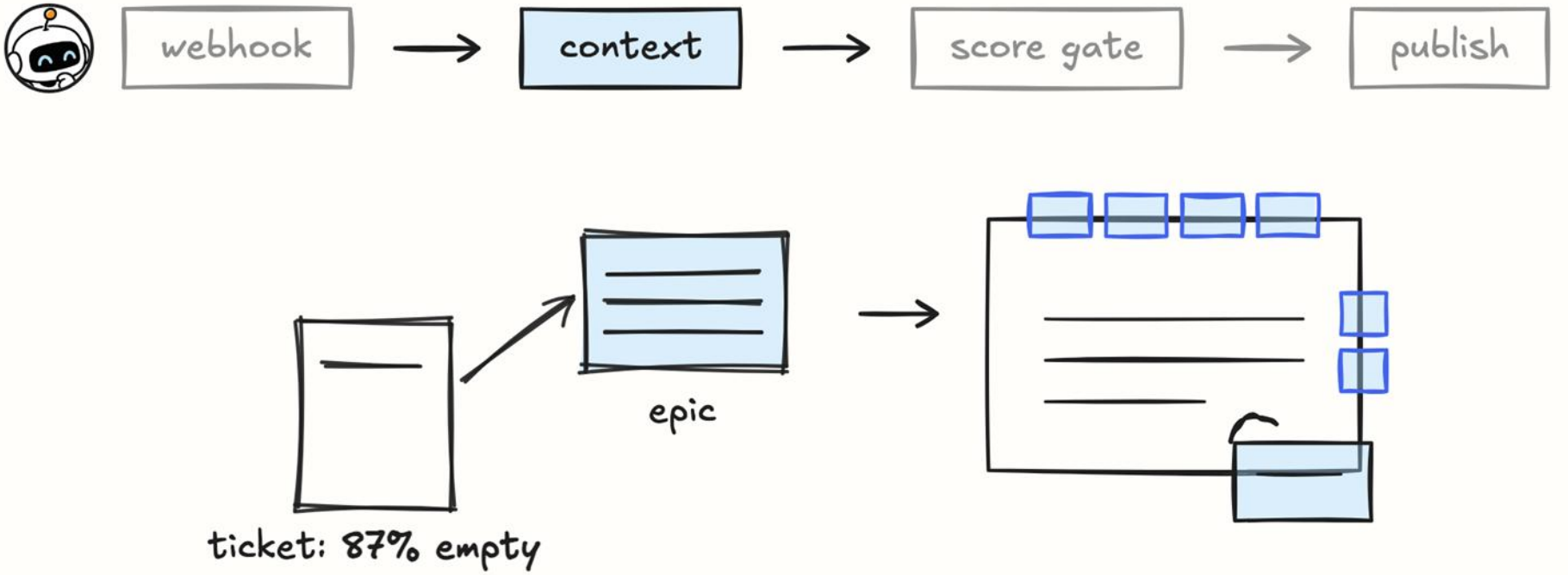

webhook
context
score gate
publish
the diff
+ mandatory rules
+ picked sections


webhook
context
score gate
publish
the diff
+ mandatory rules
+ picked sections
+ ticket background
= what the model sees

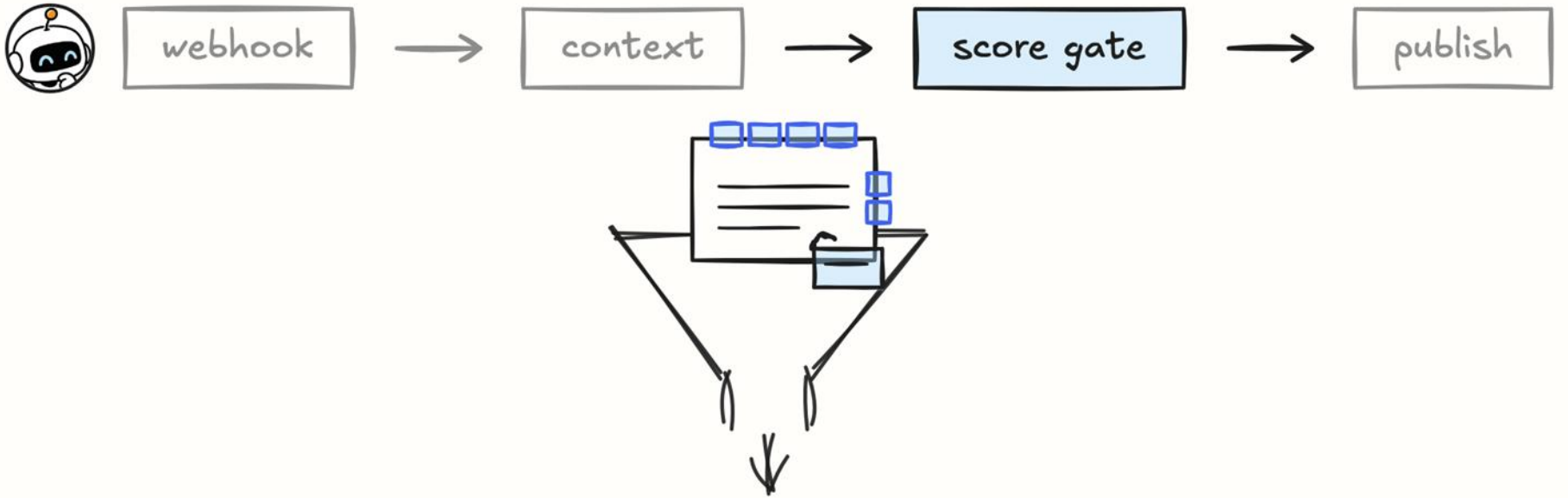

webhook
context
score gate
publish


epic
ticket: 87% empty



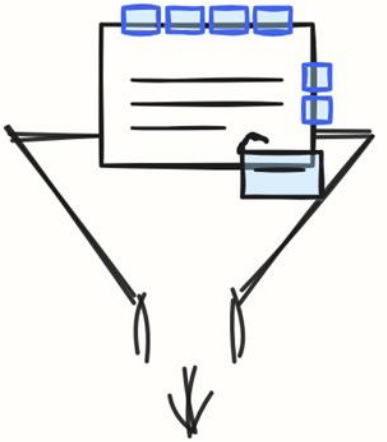


background ≠ requirements


webhook
context
score gate
publish

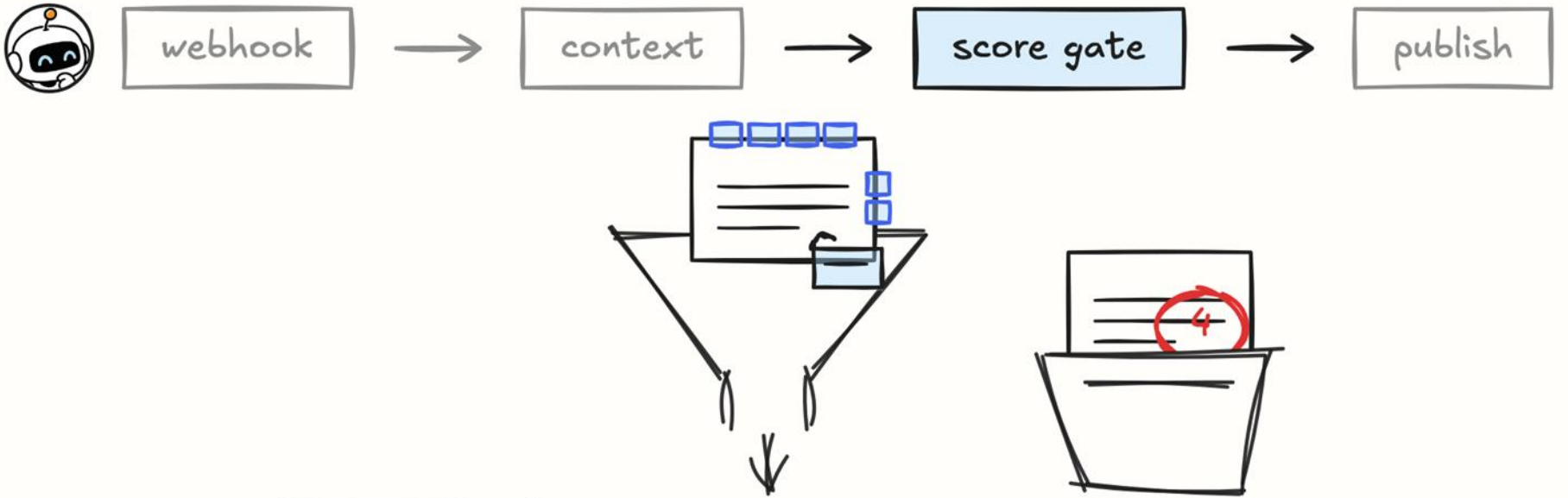


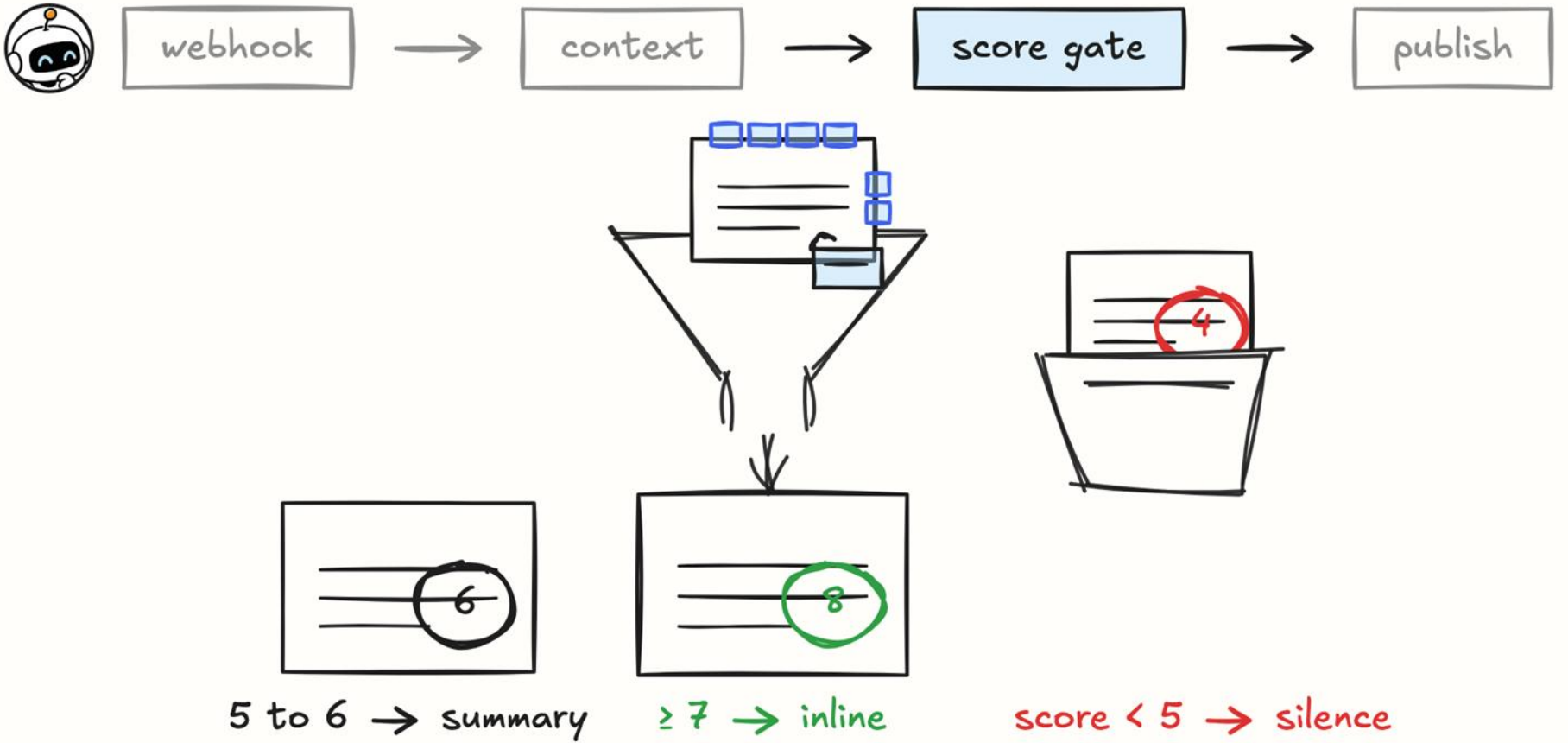
webhook → context → **score gate** → publish



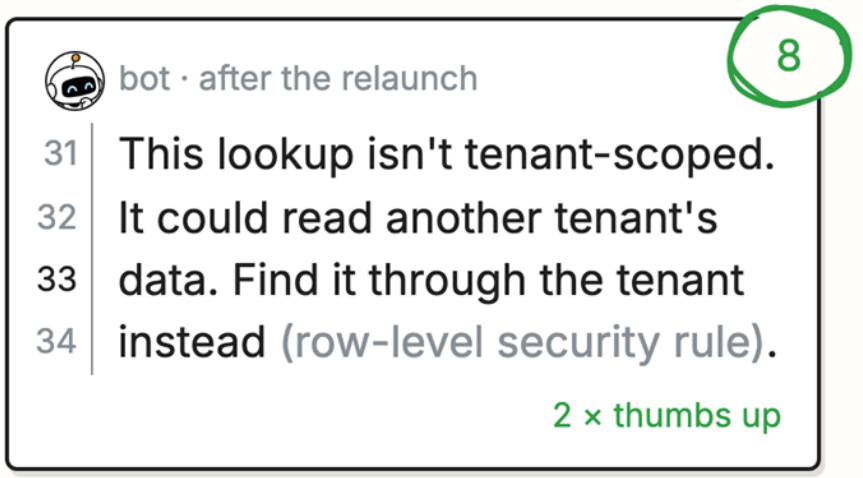


score < 5 → silence


webhook
context
score gate
publish
4
6
5 to 6 → summary
score < 5 → silence

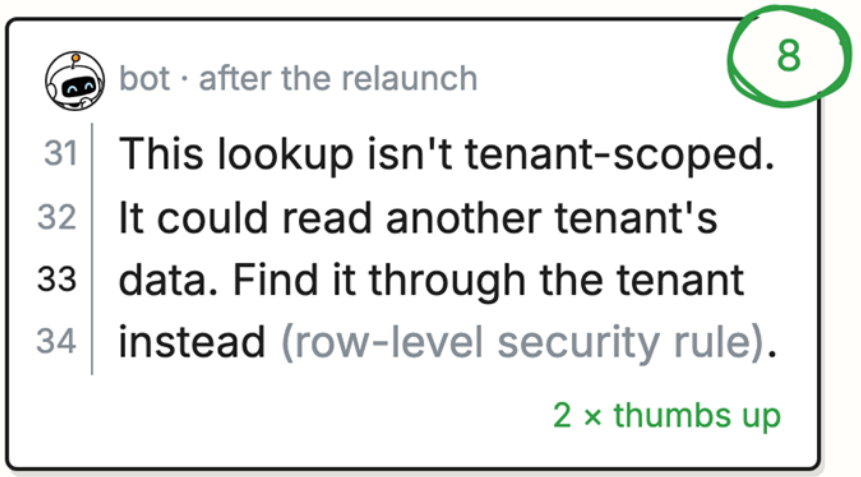

webhook
context
score gate
publish
4
6
8
5 to 6 → summary
≥ 7 → inline
score < 5 → silence

real comments · production merge requests · paraphrased

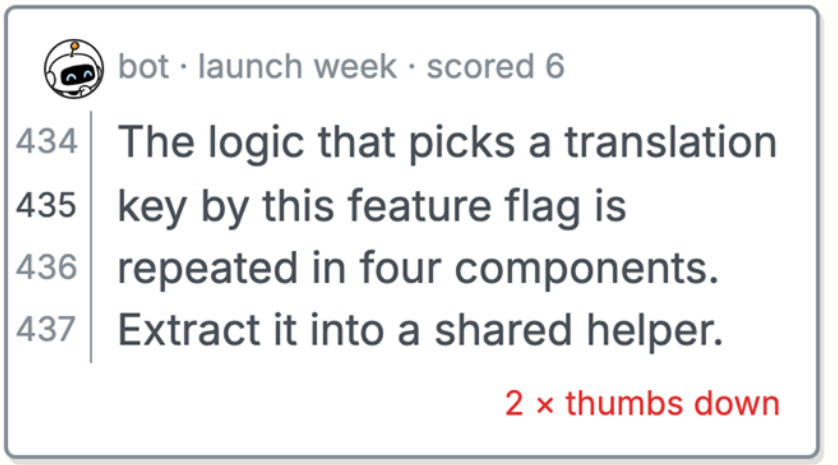


earned its place

real comments · production merge requests · paraphrased

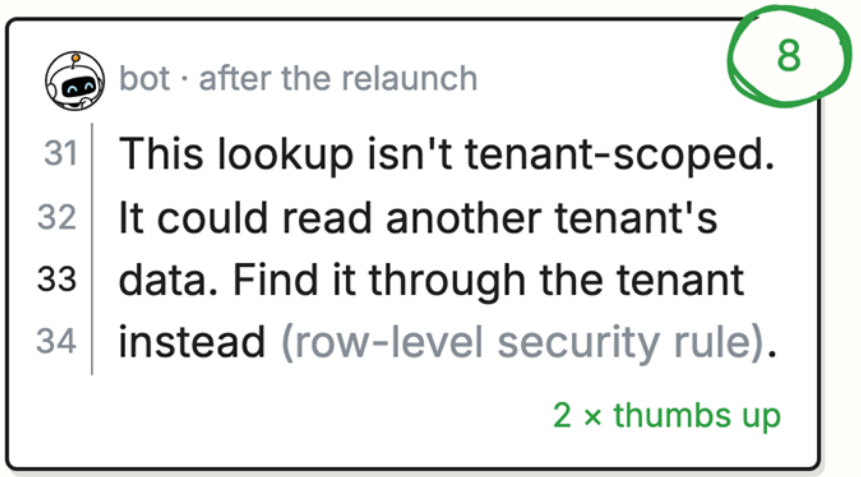


earned its place

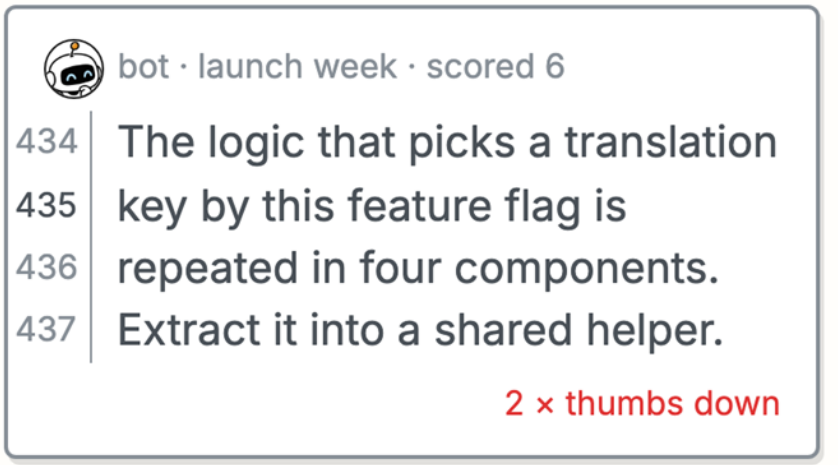


correct. nobody needed it.

real comments · production merge requests · paraphrased



earned its place



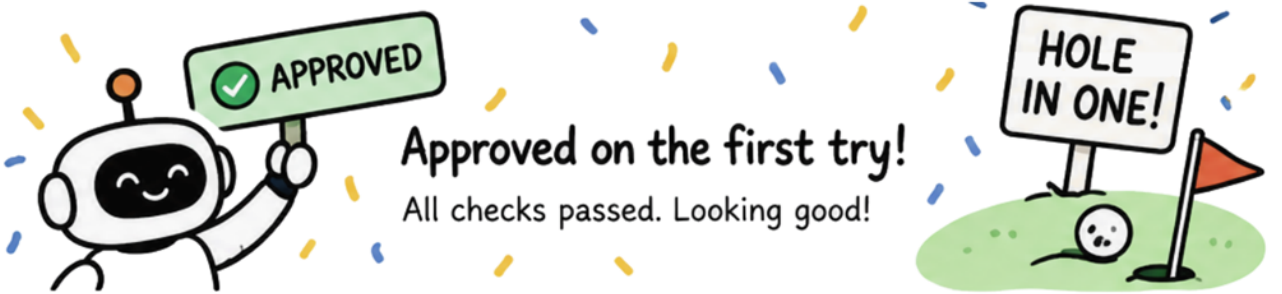
correct. nobody needed it.

the whole talk, in two cards

zero findings → approve

.../-/merge_requests/...

Add request validation for the export endpoint



✓ approved by the AI reviewer · zero findings
a real approval: one of the two the merge needs

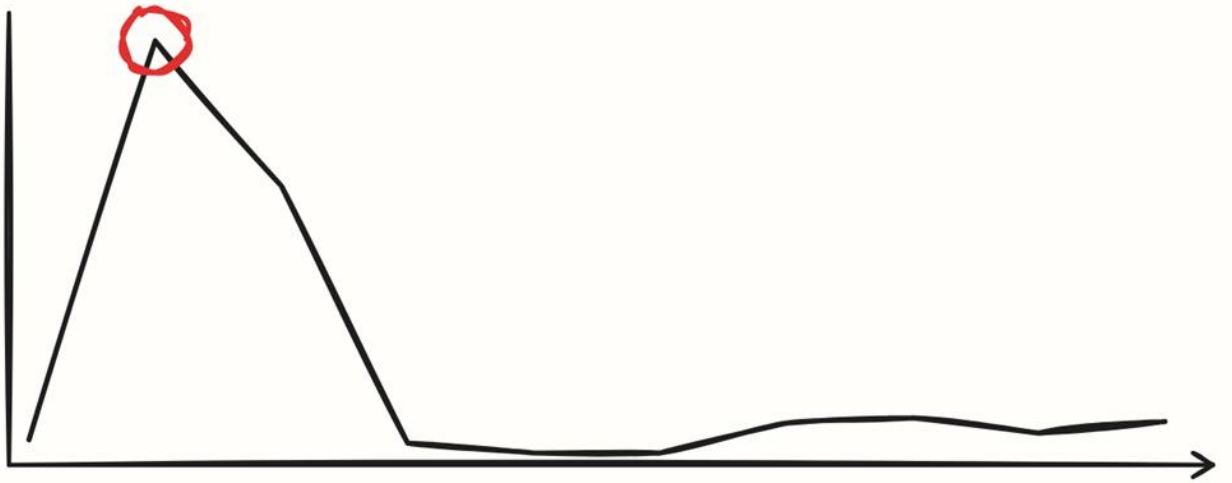
pipeline: ✓ build ✓ test ✓ lint

Merge

CI still gates · a human still merges

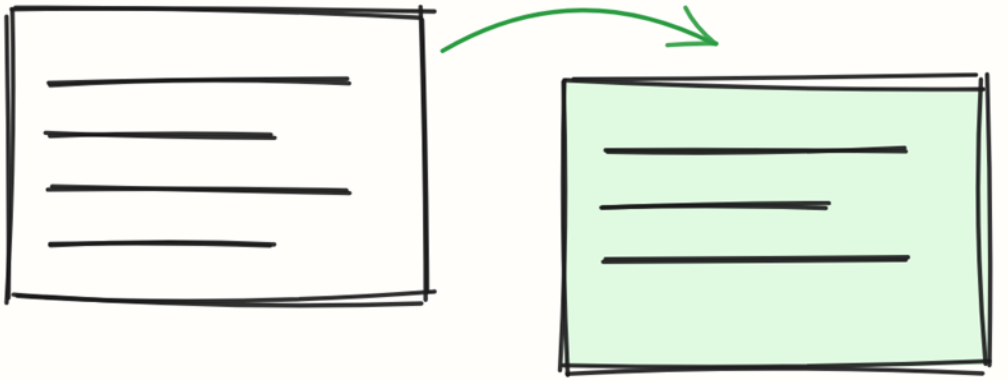
chapter 2

where it is now



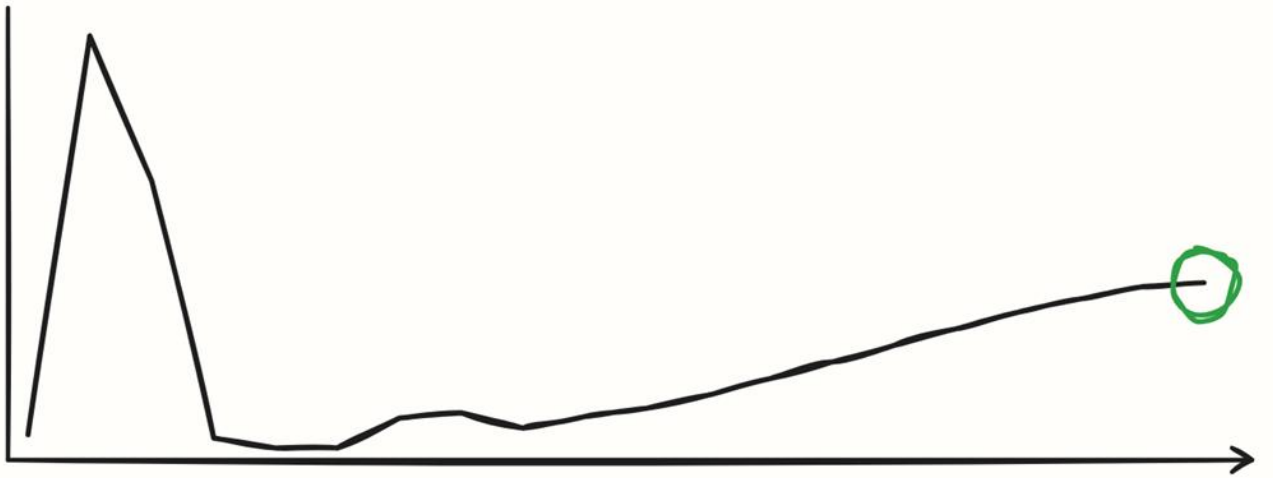
1,200/wk → ~100/wk

pilot months: 153 thumbs up · 103 thumbs down



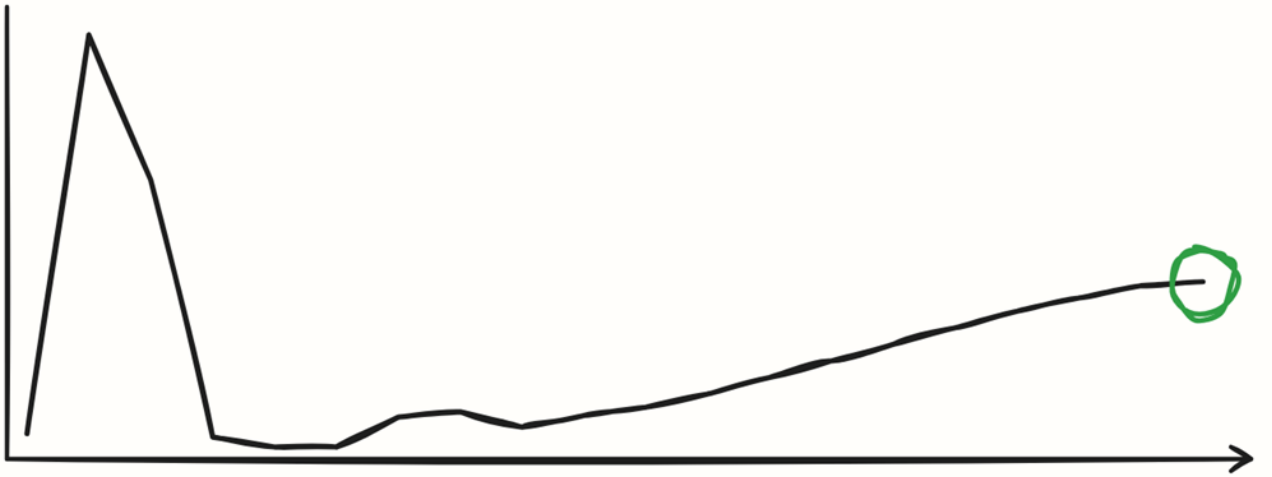
no mandate

the whole install: one config file + one CI include



today · 500+/wk · 6+ repos

busiest weeks ever



today · 500+/wk · 6+ repos

busiest weeks ever

the bill now: ~$650/mo · bill ~2x, runs ~4x

one outcome we can read

✓volume	
✓thumbs	
✓green-rate alarm	

one outcome we can read

✓volume	✓cycle time ~20 h, flat under +40% merges
✓thumbs	line changed?
✓green-rate alarm	

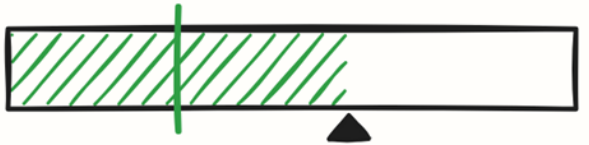
38.7%

of AI review comments changed the code. Atlassian, in production, 2026.

chapter 3

what we learned

helpful



> 30%

helpful

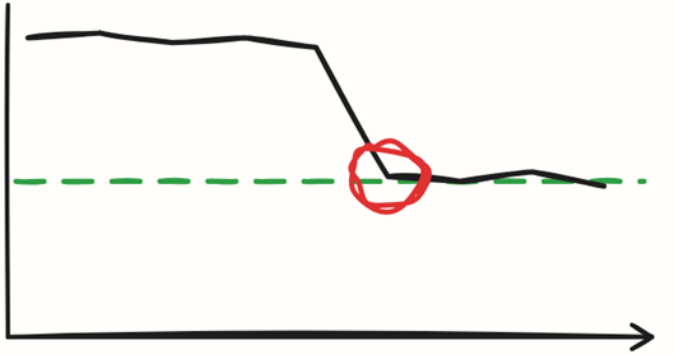


> 30%

helpful



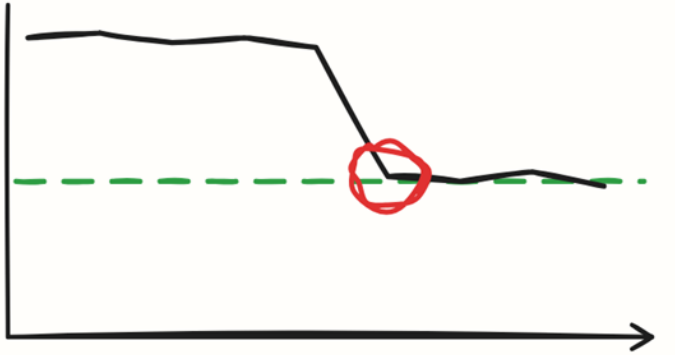
> 30%



helpful



> 30%



~60% → ~30%

48 hours. pinned the version.

credibility not capability

credibility	not capability
operability	beats features

credibility	not capability
operability	beats features
pre-define	the metric

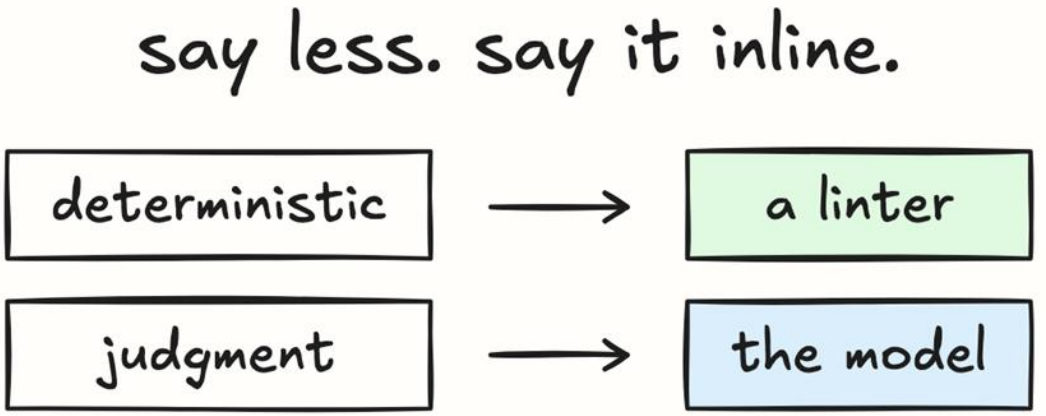
credibility	not capability
operability	beats features
pre-define	the metric
at 50/50	invert the defaults

credibility	not capability
operability	beats features
pre-define	the metric
at 50/50	invert the defaults
say less	say it inline

credibility	not capability
operability	beats features
pre-define	the metric
at 50/50	invert the defaults
say less	say it inline
adoption	you can grep for

say less. say it inline.

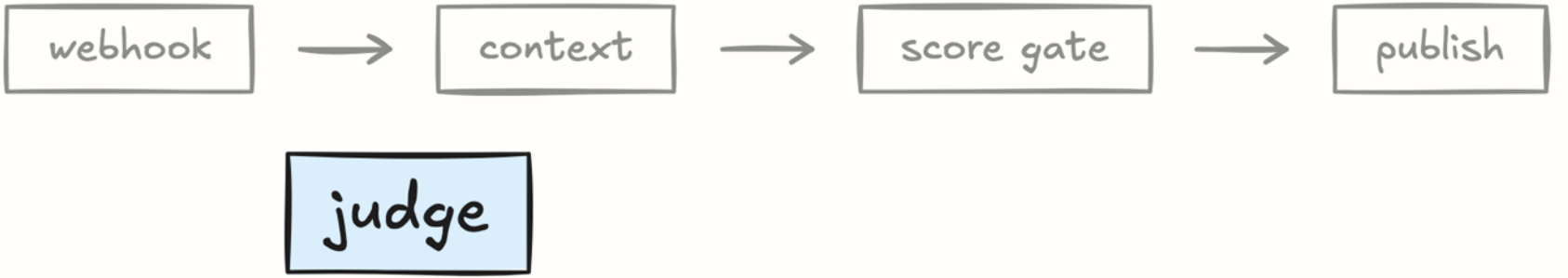


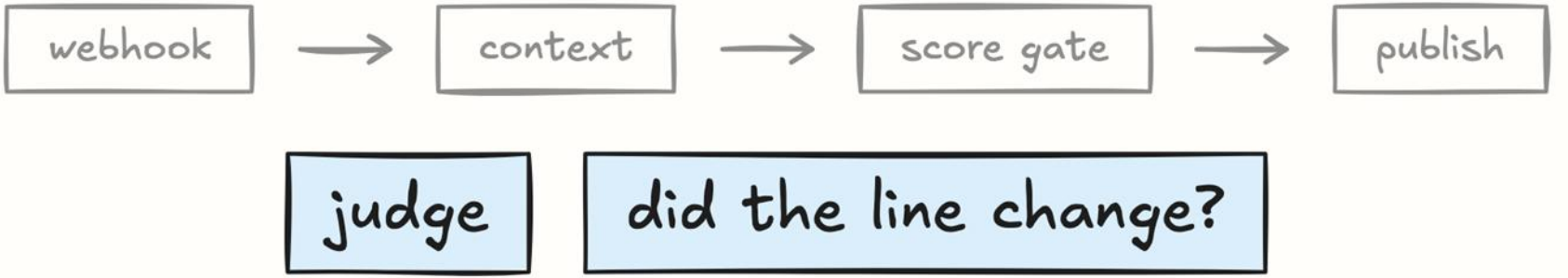

say less. say it inline.
deterministic
a linter
judgment
the model

chapter 4

if we started again


webhook
context
score gate
publish


webhook
context
score gate
publish
judge

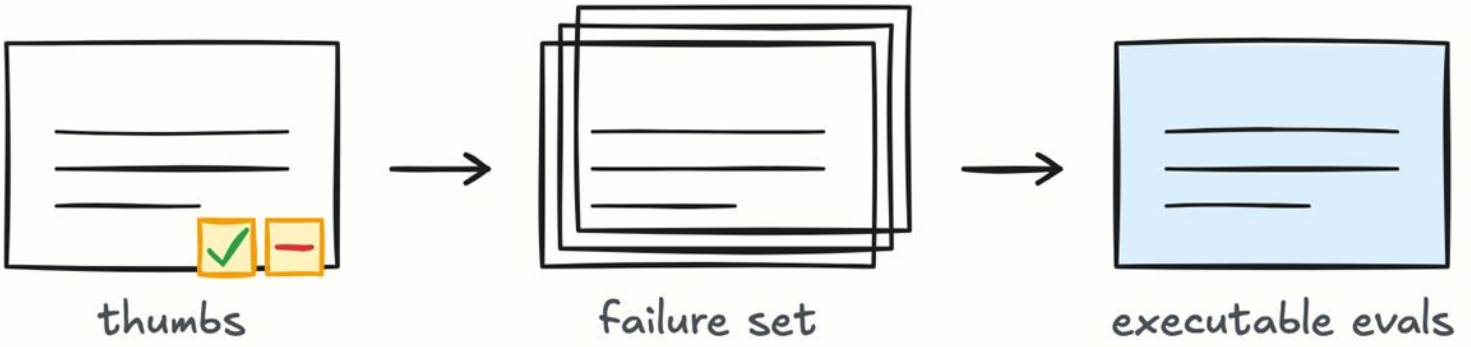


today: add a judge · track resolution



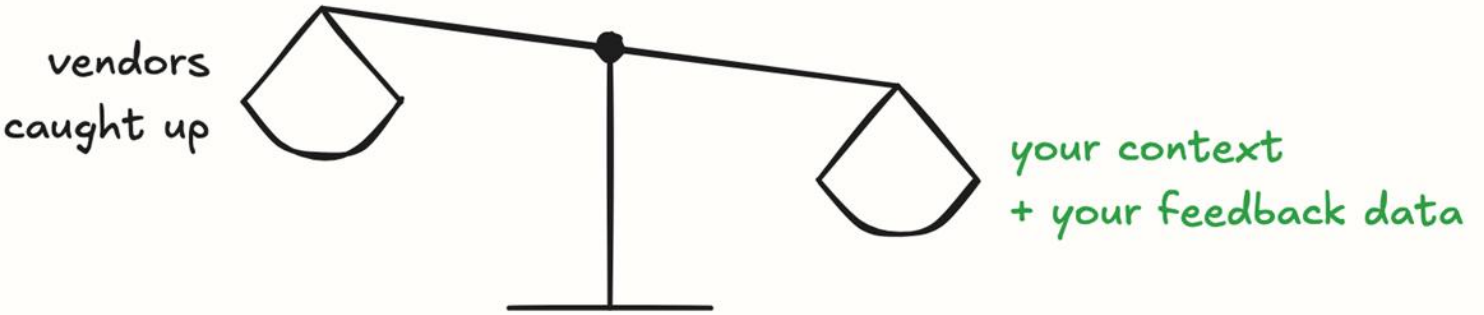
thumbs


thumbs
failure set

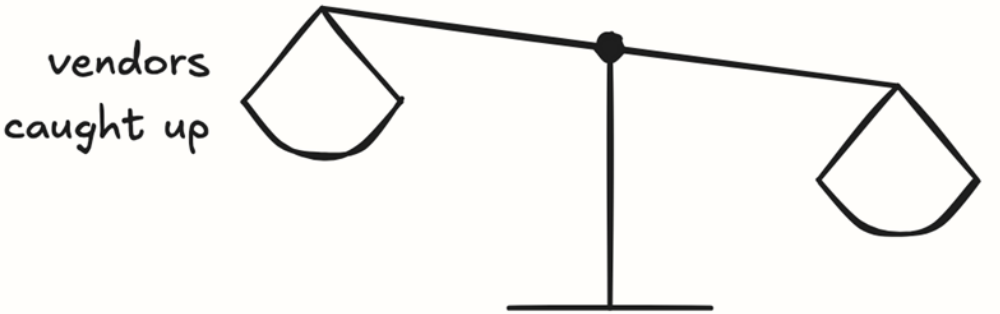


start the eval set on day one

would we build it today?



would we build it today?



forks inherit vulnerabilities, not patches

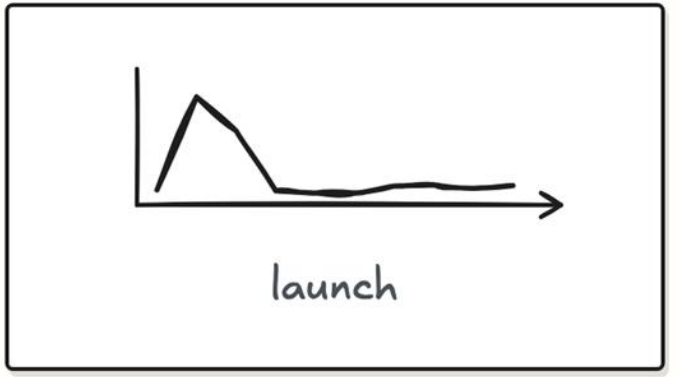
build only where you
own the context,
and the feedback data

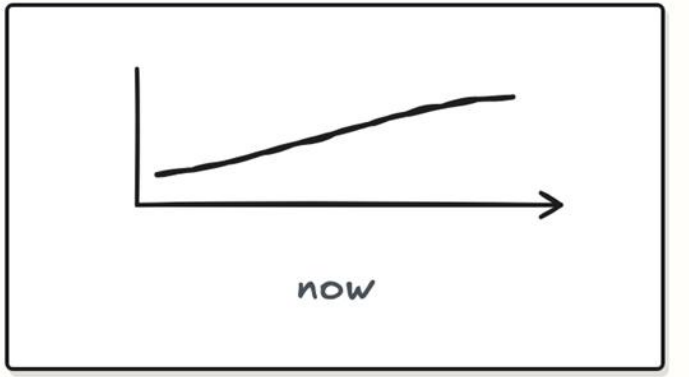
automate	deterministic to linters, patterned to the model

automate	deterministic to linters, patterned to the model
human	intent, architecture, mentoring

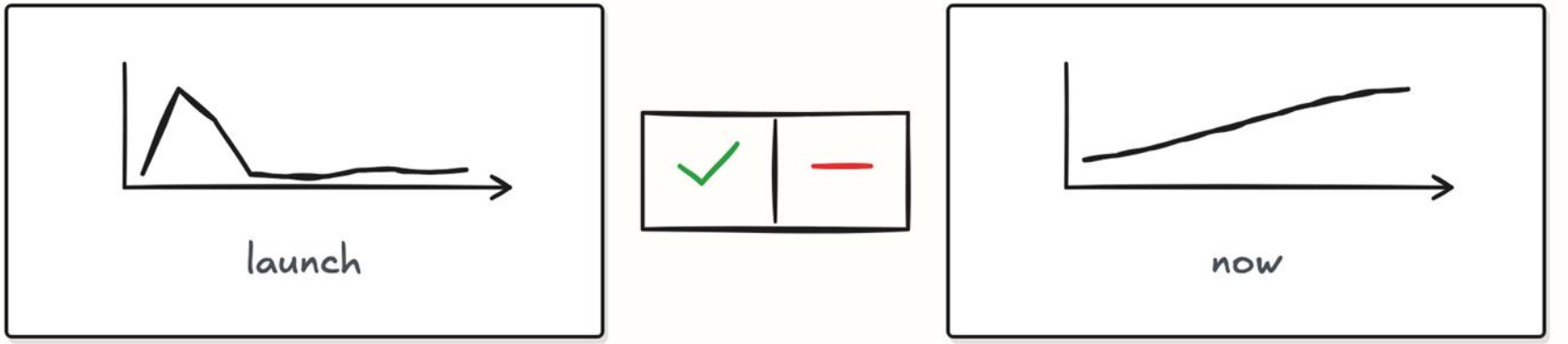
automate	deterministic to linters, patterned to the model
human	intent, architecture, mentoring
measure	pre-defined green / red + resolution

automate	deterministic to linters, patterned to the model
human	intent, architecture, mentoring
measure	pre-defined green / red + resolution
adopt	volunteers → copyable → watch the copies





same volume. opposite meaning.



same volume. opposite meaning.

thank you



follow me on LinkedIn