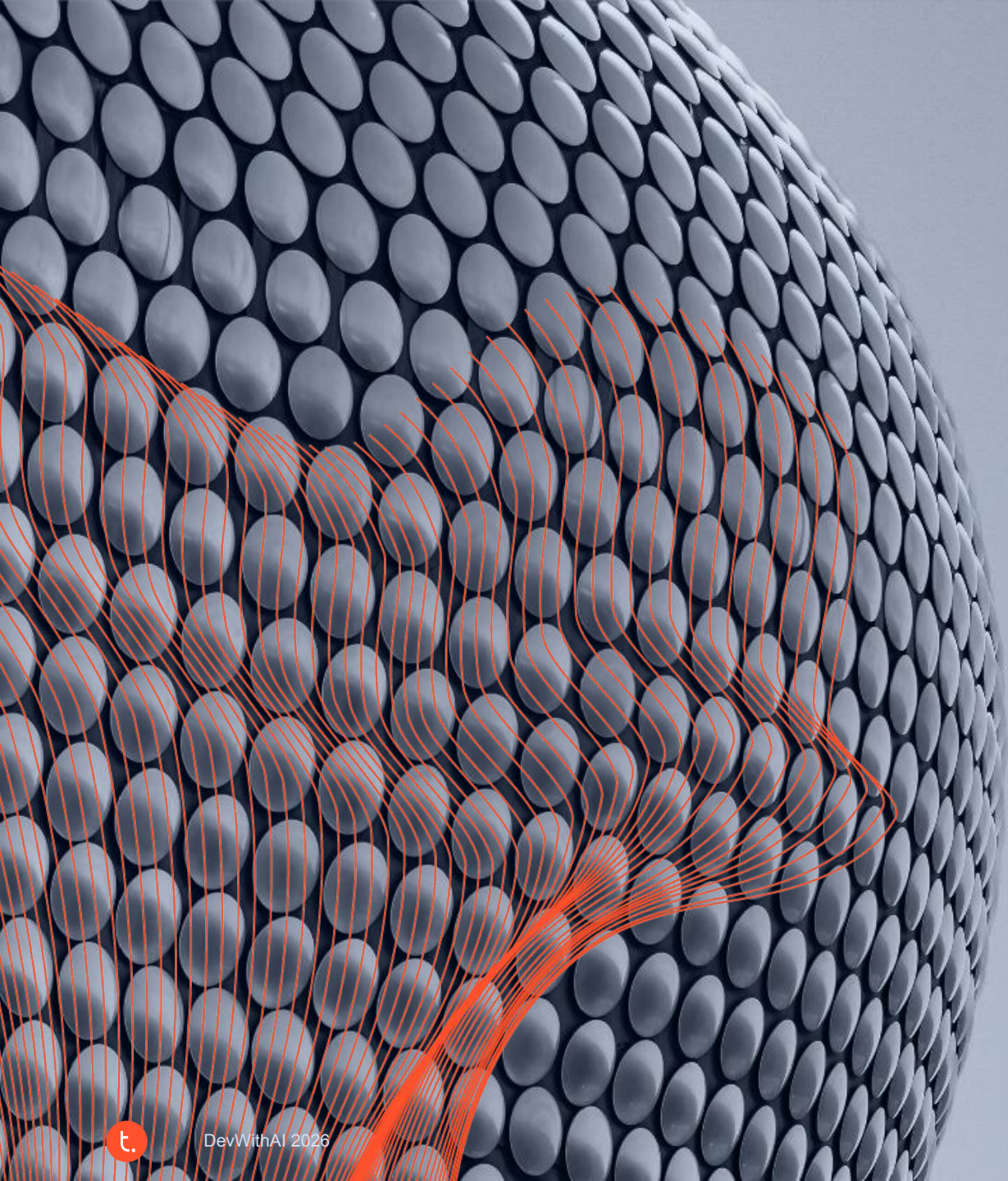




theodo.

Conversation review

How to learn when AIs write code

AGENT SESSION

Implement ticket #203

✧ Generating changes...

Allow command execution?

Deny

Always allow

AGENT SESSION

Implement ticket #203

✧ Generating changes...

Allow command execution?

Deny

Always allow

AGENT SESSION

Implement ticket #203

✧ Generating changes...

Allow command execution?

Deny

Always allow

✓ 18 files changed, +2,847 / -391 lines

AGENT SESSION

Implement ticket #203

✧ Generating changes...

Allow command execution?

Deny

Always allow

✓ 18 files changed, +2,847 / -391 lines

Accept all

- **At what point was the
engineer actually
engineering?**

● FIFTY YEARS OF REVIEW

We have reviewed code for 50 years



1976

Formal software inspection

Michael Fagan, IBM

We have reviewed code for 50 years



1976

Formal software inspection

Michael Fagan, IBM

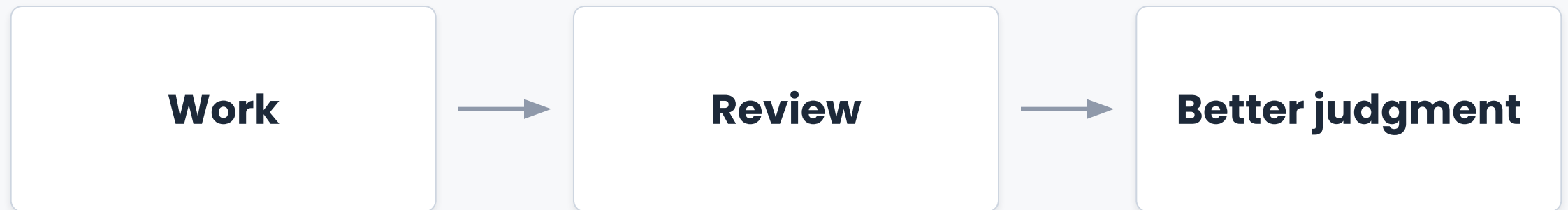


2008

Pull requests

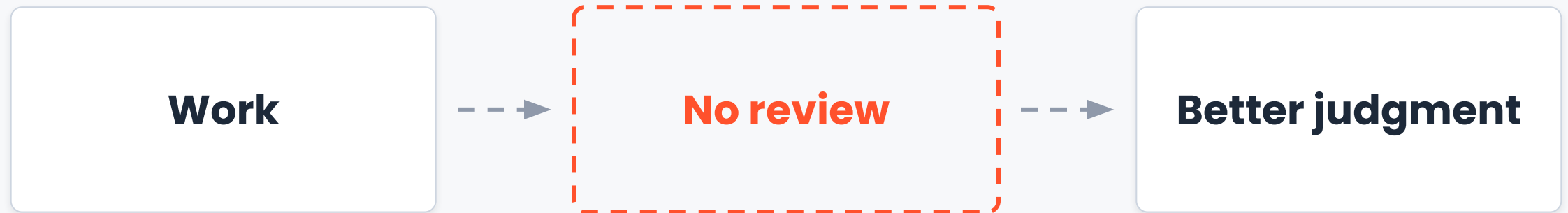
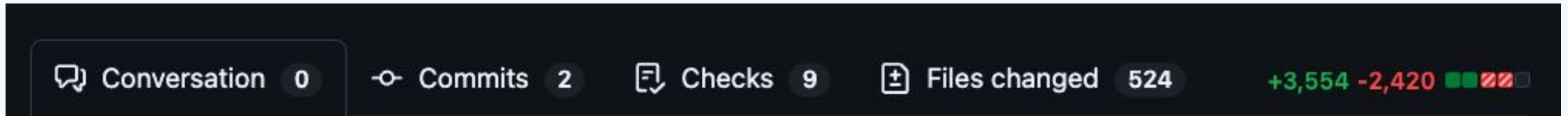
Review becomes everyday workflow

Code review matters



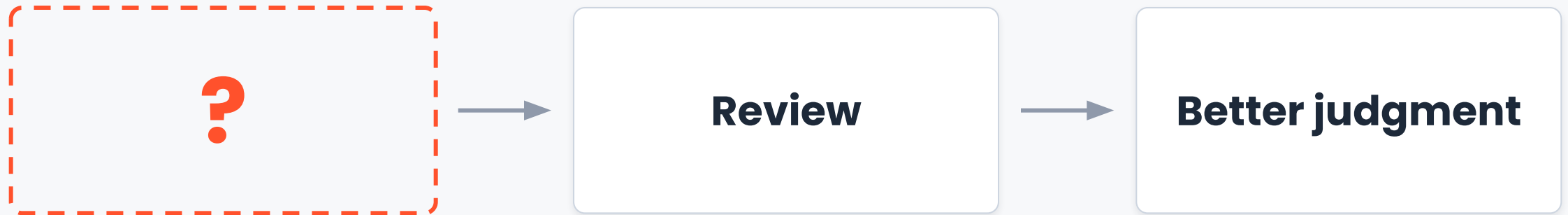
● THE FEEDBACK LOOP DISAPPEARS

But now it's sometimes no longer possible



- THE FEEDBACK LOOP DISAPPEARS

Or the work is something entirely different



**Pov: You're a developer
In 2026 🤖**

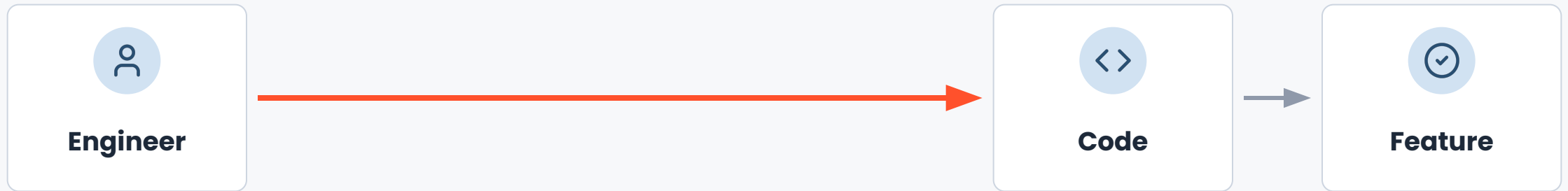


Source: r/singularity

● HOW THE WORK CHANGED

We used to write code

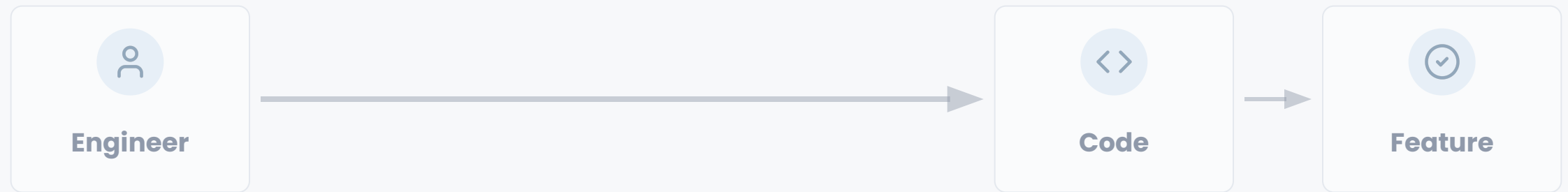
BEFORE



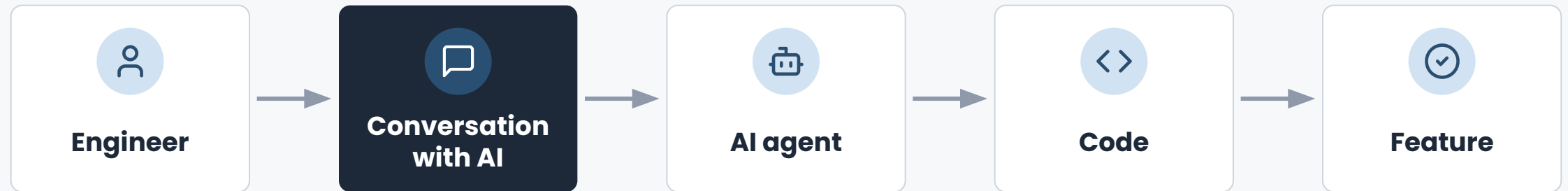
● HOW THE WORK CHANGED

Now we create conversations

BEFORE



NOW



- THE ARGUMENT

How to review conversations ?

● THE SKILL OF WRITING A CONVERSATION

Get the largest sensible task done

● THE SKILL OF WRITING A CONVERSATION

Get the largest sensible task done



Correctly

● THE SKILL OF WRITING A CONVERSATION

Get the largest sensible task done



Correctly



At the lowest cost

● THE SKILL OF WRITING A CONVERSATION

Get the largest sensible task done



Correctly



At the lowest cost



With minimal human dependency

Get the largest sensible task done



Correctly



At the lowest cost



With minimal human dependency



Using as few steps and as little context as possible

Get the largest sensible task done



Correctly



At the lowest cost



With minimal human dependency



Using as few steps and as little context as possible



With a high success rate

Get the largest sensible task done



Correctly



At the lowest cost



With minimal human dependency



Using as few steps and as little context as possible



With a high success rate



While remaining in control of the decisions that matter

● A REAL SESSION

Let's take a sample conversation

☒ I would like to integrate Laya model to triage my emails. I want a page that ... discussion dashboard simulation		\$17.46	248.1k	1h 6m	+1867 -10
L SUBAGENT 2026-09-27-01-28-53-a4f92cb8043badd67 discussion dashboard simulation		—	—	—	—

<https://github.com/theodo-group/coding-conversation-analyzer>

Coding Conversation Analyzer

Collapse All

Expand All

Toggle Tool Results

BRANCH main STARTED 2026-09-27 00:27:02 +0200 ENDED 2026-09-27 01:33:52 +0200 DURATION 1h7m SESSION eaed4937

👤 Input

15 👤 · 1 🔔

Assistant

60 🍰 · 25 🧠

Tools

135 ● · 130 ● · 4 ● · 3 📖

▼ Turn 1

 User

43k #1/15 ▼

I would like to integrate Laya model to triage my emails. I want a page that says "urgent" / "not urgent" and "action needed" , "no action". It should run using local Laya. How would you go about this ?

↓ Next User

Assistant

I'll look at the existing triage route and the LLM wiring first, so the plan fits the code you already have.

↓ Next Assistant

● **Tool Call: Bash**

43k #1/135 ▼

List project structure and search for LLM providers

```
ls; ls src src/api src/api/routes src/api/shared; grep -rniE '
```

↓ Next Tool Call

● **Tool Result: Bash**

44k #1/130 ▶

● **Tool Call: Bash**

44k #2/135

Find LLM usage and read triage route

```
grep -rniE 'laya|ollama|lmstudio|llama|openai|anthropic|claude
```

↑ Prev Tool Call

↓ Next Tool Call

● **Tool Result: Bash**

48k #2/130

● **Tool Call: Bash**

48k #3/135 ▼

Read classifier and check for local LLM runtimes

```
sed -n 1,140p src/agents/inbox-triage/classifier.ts; ls src/ac
```

↑ Prev Tool Call

↓ Next Tool Call

● **Tool Result: Bash**

○ 50k #3/130 ▶

- **Tool Call: Bash**

50k #4/135 ▼

List locally installed Ollama models

● A REAL SESSION

Let's take a sample conversation

I would like to integrate Laya model to triage my emails. I want a page that says "urgent" / "not urgent" and "action needed", "no action". It should run using local Laya. How would you go about this?

COST

\$17.46

TOKENS

248.1k

DURATION

1h 6m

LINES CHANGED

+1,867 -10

What is the task?

TASK	COST	TOKENS	TIME	LINES
Technical strategy for model integration	\$17.46	248.1k	1h 6m	+1,867 -10

USER PROMPT

I would like to integrate Laya model to triage my emails. I want a page that says "urgent" / "not urgent" and "action needed", "no action". It should run using local Laya. **How would you go about this?**

🔍 Is it worth delegating?

🔍 What does a success look like?

- DID THEY APPROACH THE PROBLEM CORRECTLY?

What should a conversation look like?

Centaur: delegation

Hand the AI one clearly bounded task

Human

AI

Cyborg: co-thinking

Work in short, interleaved turns

Human

AI

Human

AI

Source: Ethan Mollick, Centaurs and Cyborgs on the Jagged Frontier

What has been produced?

TASK	COST	TOKENS	TIME	LINES
Integrate Laya model to triage my emails	\$17.46	248.1k	1h 6m	+1,867 -10

🔍 Does it actually work?

🔍 Has it been engineered according to best practices?

● HOW WAS THE RESULT VERIFIED

How do you check a “technical strategy”?



```
Tool Call: Bash 154k #71/75
Typecheck, lint, full tests
npx tsc --noEmit 2>&1 | grep -i inbox; npx biome check src/api/routes/inbox.ts s

↑ Prev Tool Call ↓ Next Tool Call

Tool Result: Bash 154k #71/75
===TESTS===
Test Files 10 passed (10)
Tests 155 passed (155)
Start at 09:44:22
Duration 357ms (transform 799ms, setup 0ms, import 1.37s, tests 271ms, environment 1

↑ Prev Result ↓ Next Result
```

?

TYPECHECK
tsc --noEmit

?

LINT
biome check

?

TESTS
155 passed in 10 files

What went on during the conversation?

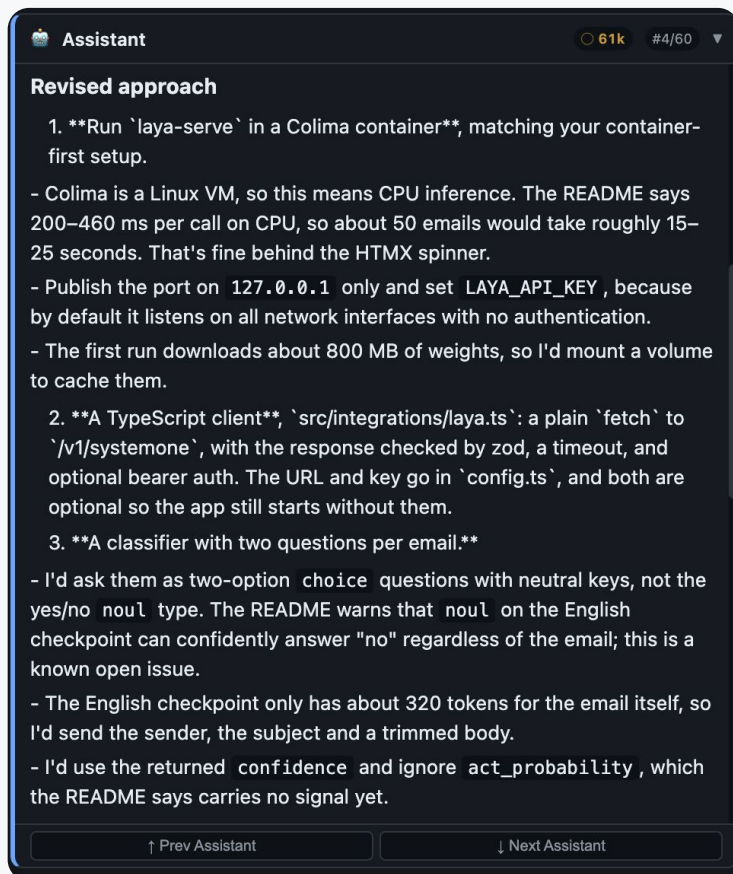
TASK Integrate Laya model to triage my emails	COST \$17.46	TOKENS 248.1k	TIME 1h 6m	LINES +1,867 -10
15 user inputs	60 assistant messages		135 tool calls	

Is the cost in tokens justified?

TASK Integrate Laya model to triage my emails	COST \$17.46	TOKENS 248.1k	TIME 1h 6m	LINES +1,867 -10
---	------------------------	-------------------------	----------------------	----------------------------

● WHAT DID THE USER AND THE AGENT DO

When did the task slip?



“Should I go ahead? And do you want the evaluation step with your own labelled emails in this first pass, or the page first and evaluation afterwards?”

An LLM question that flipped the mode

● WHAT DID THE USER AND THE AGENT DO

Could they reduce the effort necessary?



**“Are you really looking in my
node_modules?”**

More than a half of this sessions tool calls (75) was avoidable.
It could have yielded 15% economy in tokens and 50k less context.

The review framework

1

INTENT

**Did they approach
the problem
correctly?**

2

EVIDENCE

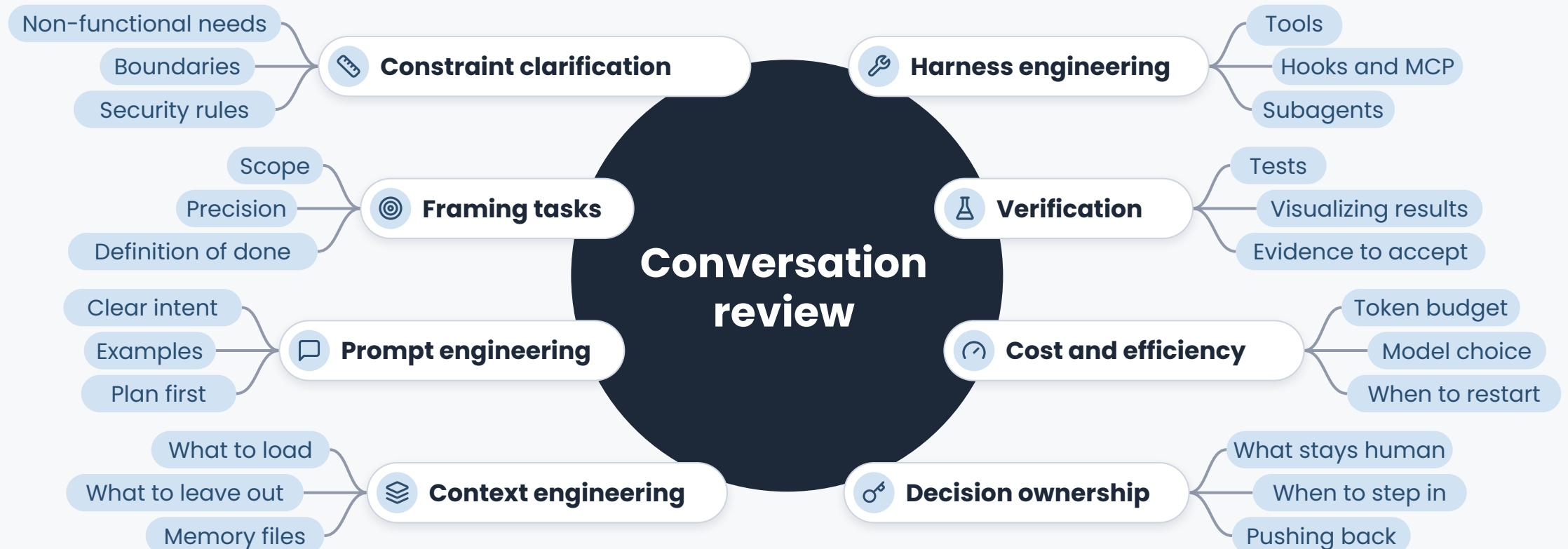
**What justified
acceptance?**

3

CONTROL

**Did they manage scope,
context, tools, permissions,
and consequential
decisions?**

What a conversation review teaches



How the team is evolving

1

TASKS

**Become
standardized and
reusable**

2

AGENTS

**Are created and
improved constantly
to respond to needs**

3

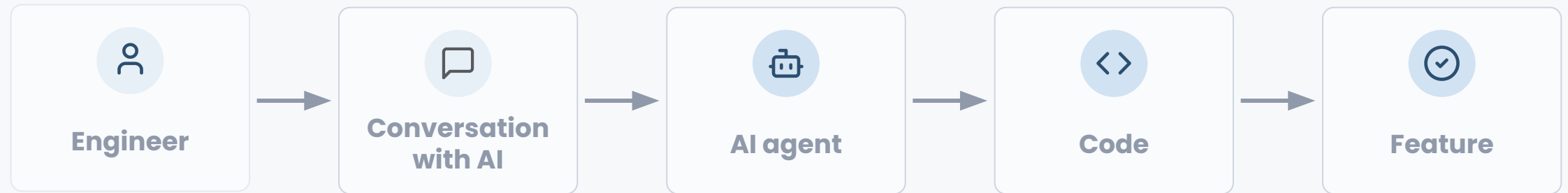
HARNESS

**Environnement
improves to delegate
with confidence**

● THE IMPACT

We are building something else

NOW



NOW



AGENT SESSION

Implement ticket #203

✧ Generating changes...

Allow command execution?

Deny

Always allow

✓ 18 files changed, +2,847 / -391 lines

Accept all

Conversation review as a team leader ritual

10

minutes

- 1 **Analyze one real session**
- 2 **Identify one consequential step**
- 3 **Try one improvement**

**Conversation review:
reintroduce discussions
precisely where we need
collective learning most**

theodo.

Marek Kalnik

CTO, Agentic Delivery Advisory

