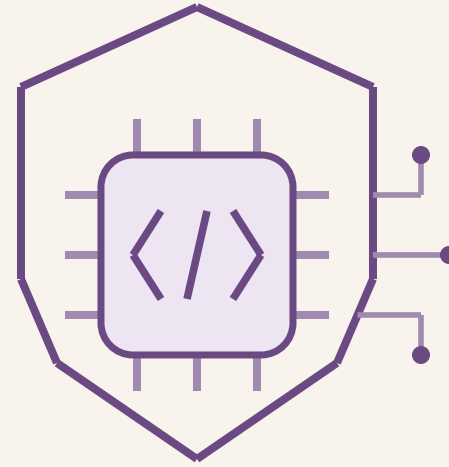


# Before the Agent Writes Code: Policy-Aware Engines for AI Coding



Nnenna Ndukwe

AI Developer Relations Lead & Software Engineer at Qodo



[Resources + about me](#)

# We'll trace one task from policy to reviewed code

- 01 Start with a coding task  
What does the agent need before it writes code?
- 02 Open and measure the policy system  
Stack → runtime flow → latency + source fidelity
- 03 Test what reaches the first draft  
One task, three context conditions, twelve implementations
- 04 Connect it to a team workflow  
Codebase context, coding standards, and Qodo review

# Token cleanup must handle retries and cancellation

## Implement a refresh-token cleanup background job

---

### Repeated runs

Safe to run again

---

### Cancellation

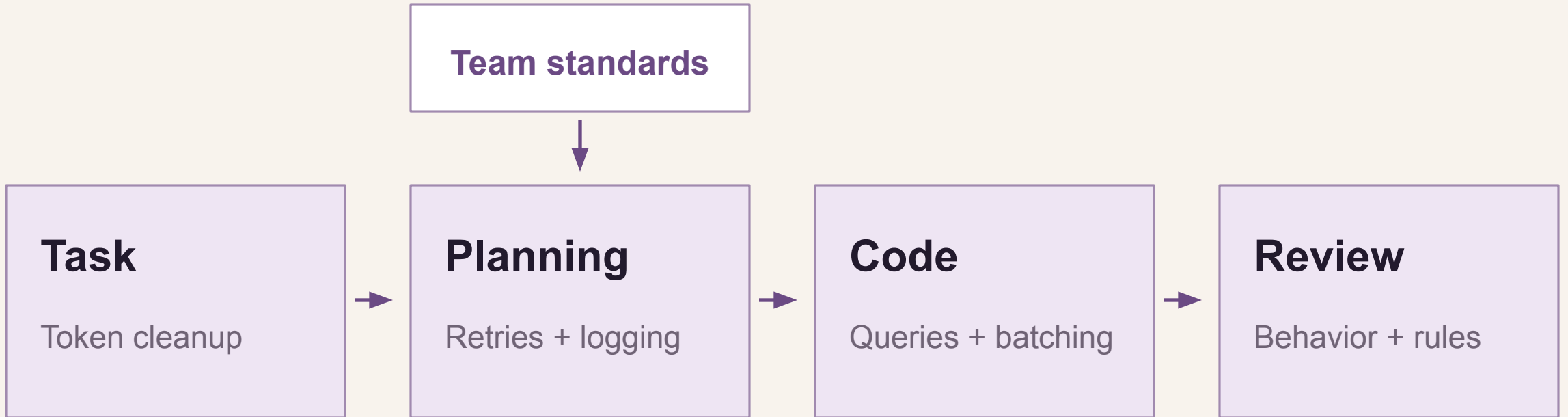
Stop without corrupting state

---

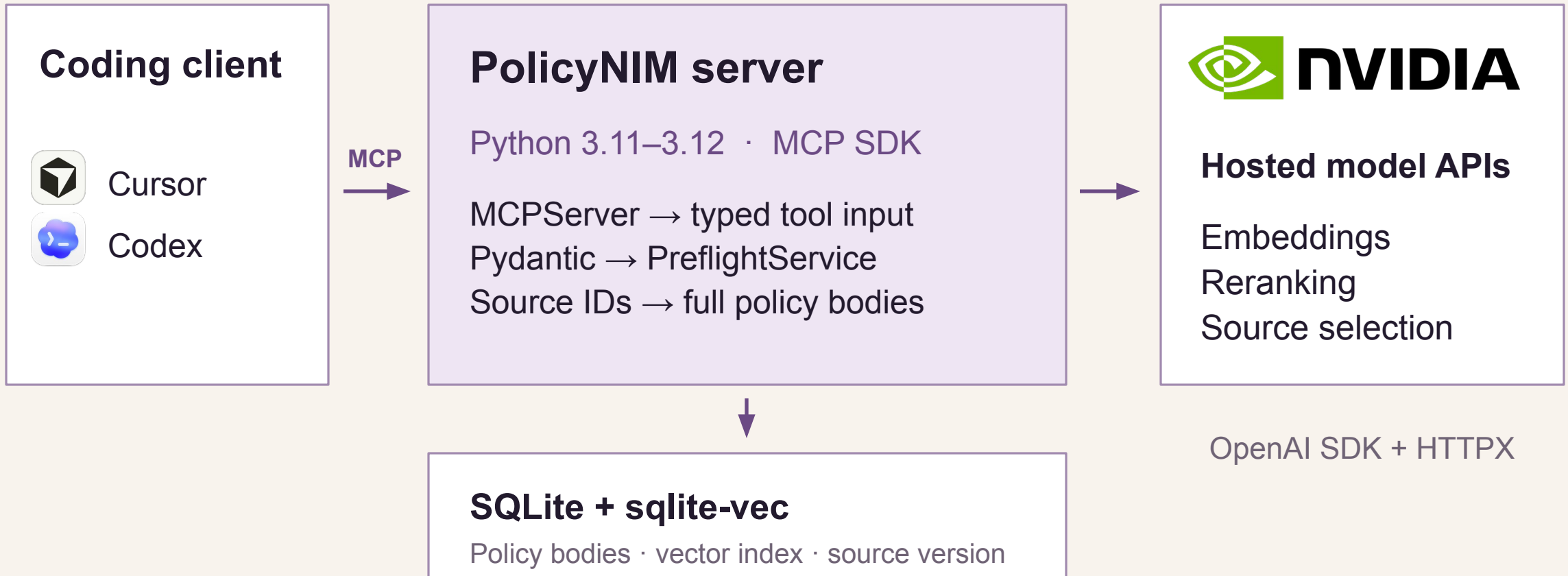
### Safe logging

Keep token values out of logs

# The coding plan needs the team's rules



# PolicyNIM serves engineering rules through MCP

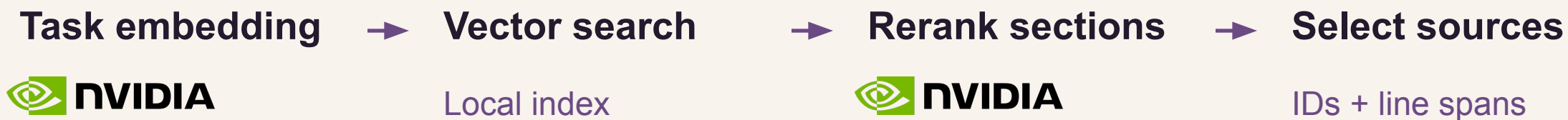


# PolicyNIM ranks the rules a coding task needs

## BUILD THE INDEX



## SERVE THE CODING TASK



# PolicyNIM preflight: 2.9 seconds at the median

**Ranked search**

1.36 s

Observed range: 1.34–1.73 s

**Full-source preflight**

2.91 s

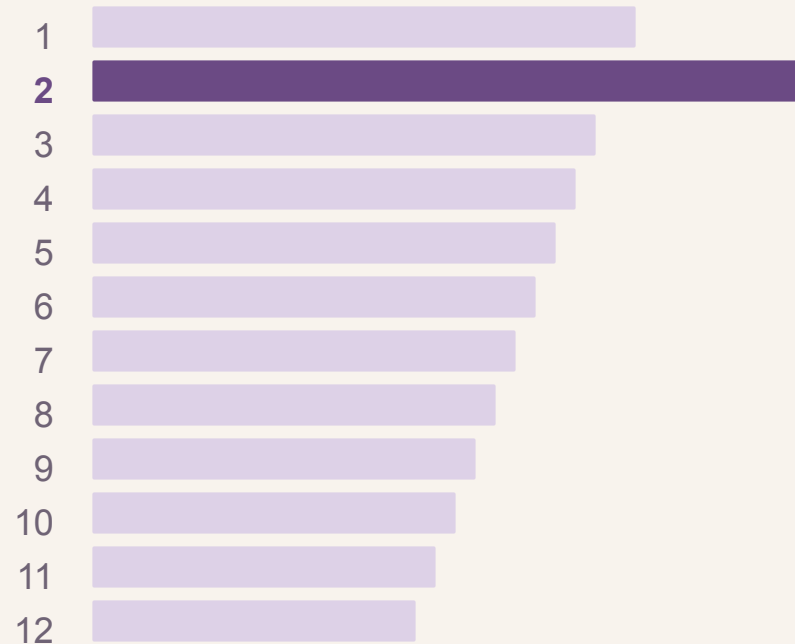
Observed range: 2.61–3.15 s

**5 / 5 preflight responses matched full bodies + source coordinates**

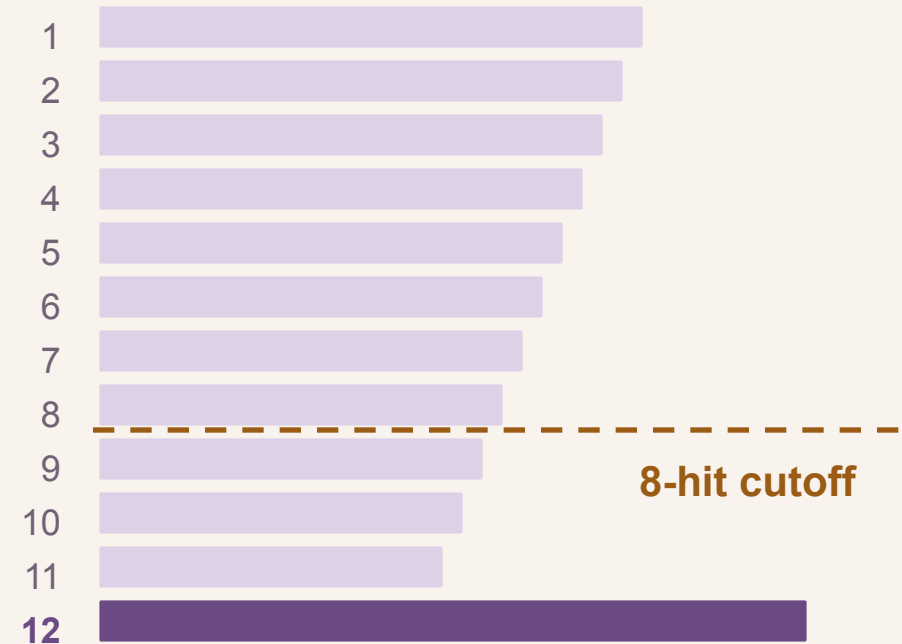
4 upstream calls · 18,576 measured tokens + unreported reranker usage

# Reranking dropped the cleanup job's required rules

## Dense retrieval



## After reranking



Required-rule section: rank 2 → rank 12

# Full policy sources keep requirements together

## Selected policy

Purpose

**Required rules**

Conditions

Exceptions



Complete source  
within the budget

If it won't fit, narrow the request.

# A valid citation can still carry an invented rule

## Valid citation

A real policy ID  
and source location

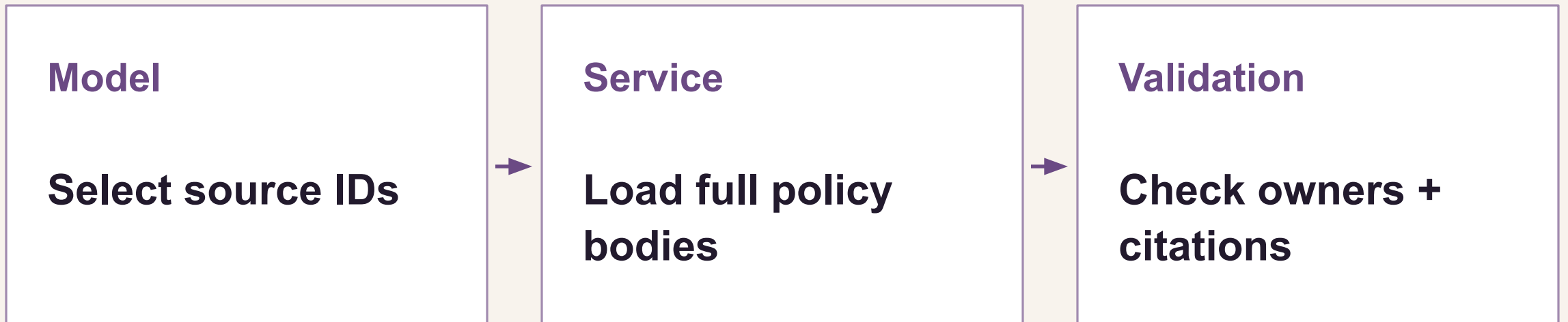
## Unsupported detail

Batch sizes

API choices

Token lifecycle

# PolicyNIM returns the policy text the model selects



The response retains the source's rules, conditions, and exceptions.

# Across 90 attempts, 110 seconds less latency

MCP pilot · same 30 tasks, three passes · 90 requests per variant

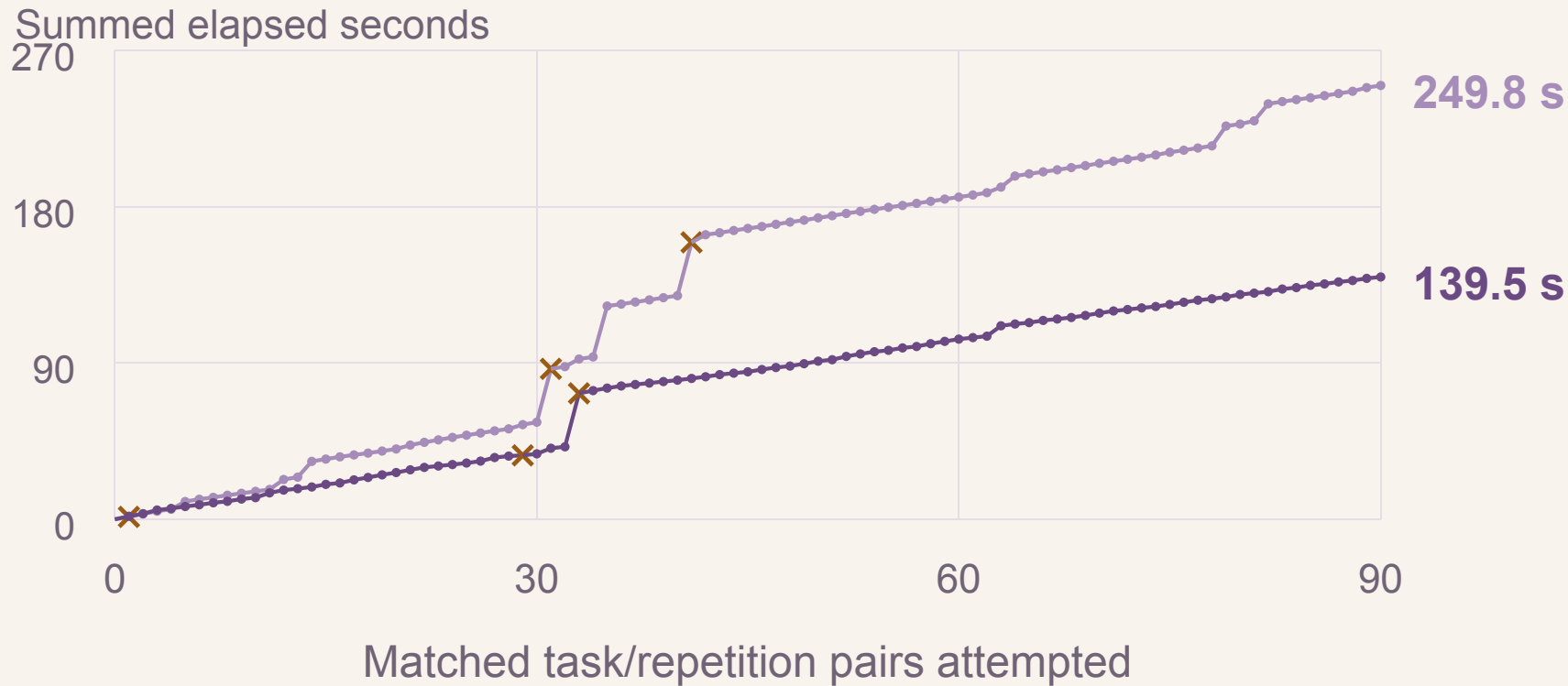
— **Current · NVIDIA**

87/90 complete

— **Jev + fallback**

88/90 complete

✗ Failed call



**110.3 s**  
less summed  
request latency  
**44.1% lower**

75 paired requests  
**60.1 s less**  
39.9% lower

# One policy version keeps text and citations aligned

## One immutable generation

**Source text**

**Catalog**

**Search index**

**Runtime rules**

Each operation holds the source version it reads.

# One task, three context conditions, twelve first drafts

**Same model · same task · frozen behavior checks**

## **Task only**

Task + public interfaces

## **8 ranked chunks**

Task + retrieved sections

## **Full policy sources**

Task + PolicyNIM preflight

**4 repeats**

**4 repeats**

**4 repeats**

14 checks: cancellation · retries · duplicate effects · safe logging

# All twelve first drafts still needed review

**Task only**

**8 ranked chunks**

**Full policy sources**

0 / 4

0 / 4

0 / 4

complete passes

---

complete passes

---

complete passes

---

**27 / 56 checks passed**

**22 / 56 checks passed**

**25 / 56 checks passed**

Measure the implementation against the rules it received.

# Inspect cleanup rules before the agent writes code

```
policy_preflight_evidence
```

**Ask for the task**

Refresh-token cleanup



**Open the sources**

Rules + line references



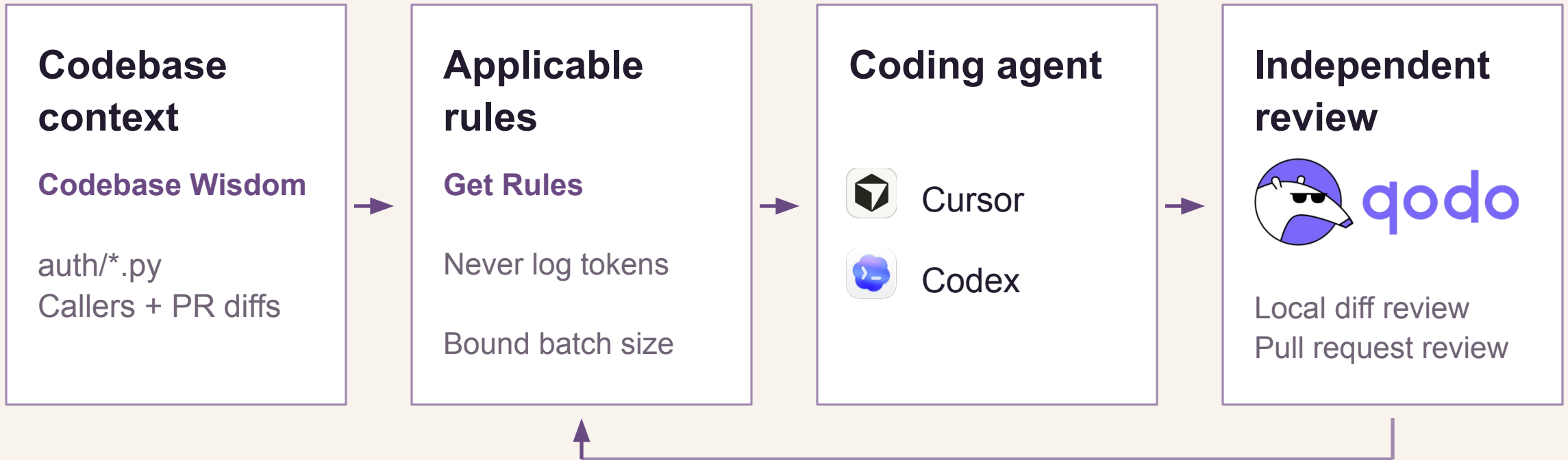
**Plan the change**

Explain each decision

Inspect the policy text and source version before implementation.

# Qodo connects standards with code and review

## Qodo Agentic Toolbox · MCP or CLI



Teams own the rules · scope by organization and repository · update from review

Bring standards  
into planning.

Review the  
resulting code.

Keep exploring

The code. The evidence.  
The open standards.

PolicyNIM · Qodo · MCP · AAIF

[nnennandukwe.com](https://nnennandukwe.com)



Scan for all resources