Elizabeth Fuentes Leone

Developer Advocate

elifuentes.tech

bit.ly/4cRzkvJ

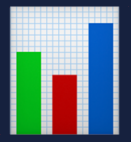
Your AI Agent Isn't Crashing. It's Bleeding Tokens

3 silent failures that cost money and time

IBM Research • Octopus • The Decoder

Three Silent Failures

None of these throw errors. They just waste tokens.



Context Overflow

214KB
→ wrong results

MCP Tools Hanging

424 error
17.2s wait

Reasoning Loops

14 calls
21 seconds

Fix 1: Memory Pointer Pattern

IBM Research: Solving Context Window Overflow

Problem

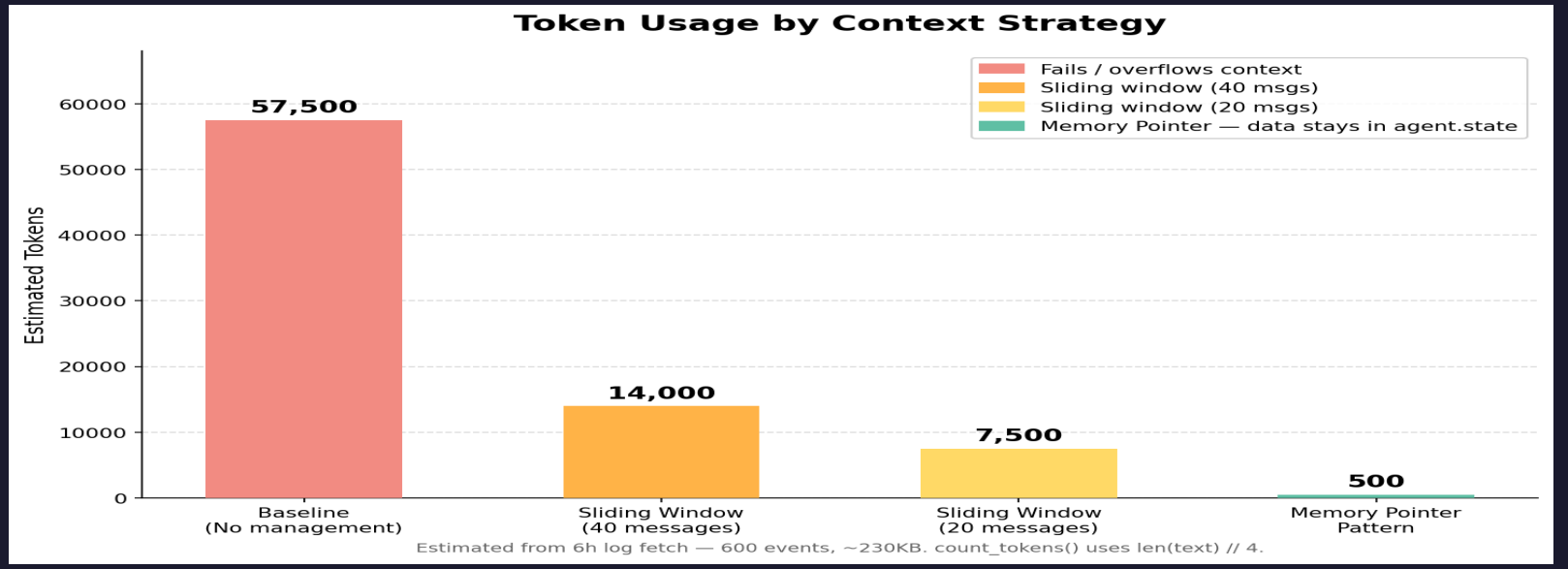
- Tool returns 214KB of logs

Context window fills up
Agent produces wrong results
No error thrown

Solution

- Store data in agent.state

Return 52-byte pointer
Agent queries when needed
7x token reduction



Demo Time

Memory Pointer: 600 events, 7x token reduction

Fix 2: Async HandleId for MCP

Octopus Research: Resilient AI Agents With MCP

Problem

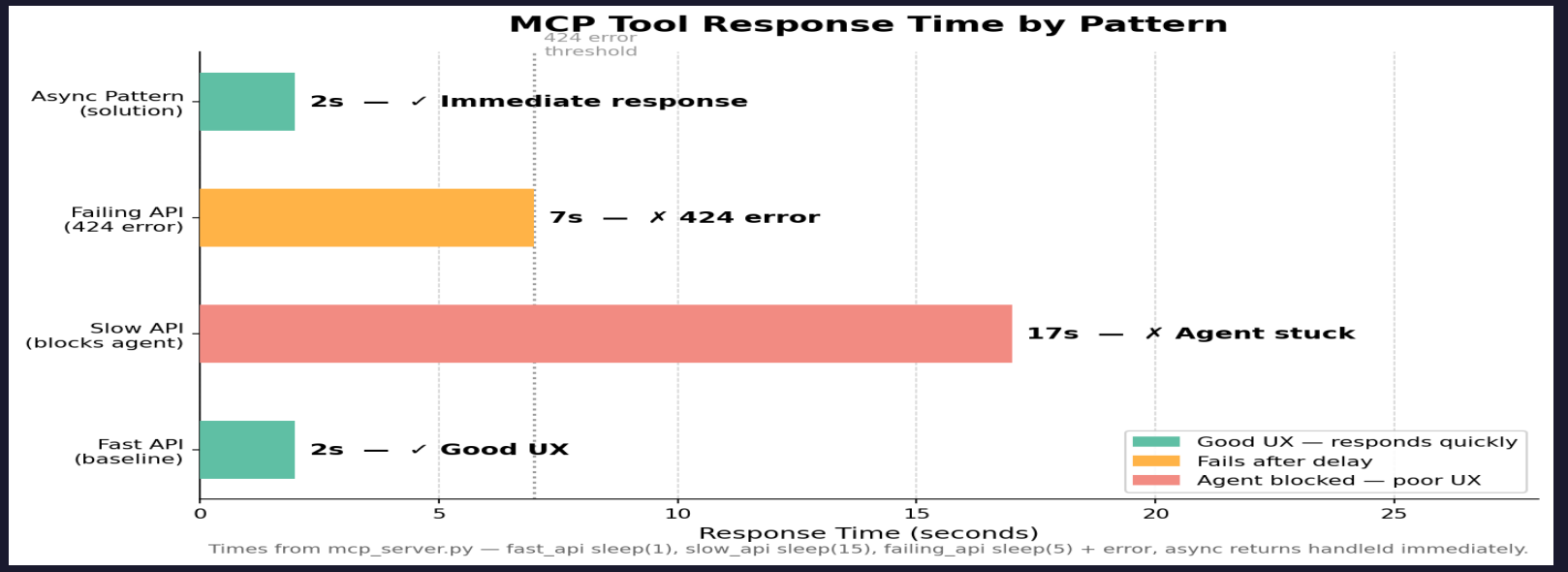
- MCP tool calls slow API (15s)

Agent waits 17.2 seconds
424 error: workflow frozen
No way to recover

Solution

- start_long_job returns handle

check_status polls for result
Agent continues working
17.2s → 1.7s response time



Demo Time

MCP Async: 4 scenarios, 17.2s → 1.7s

Fix 3: DebounceHook + Clear States

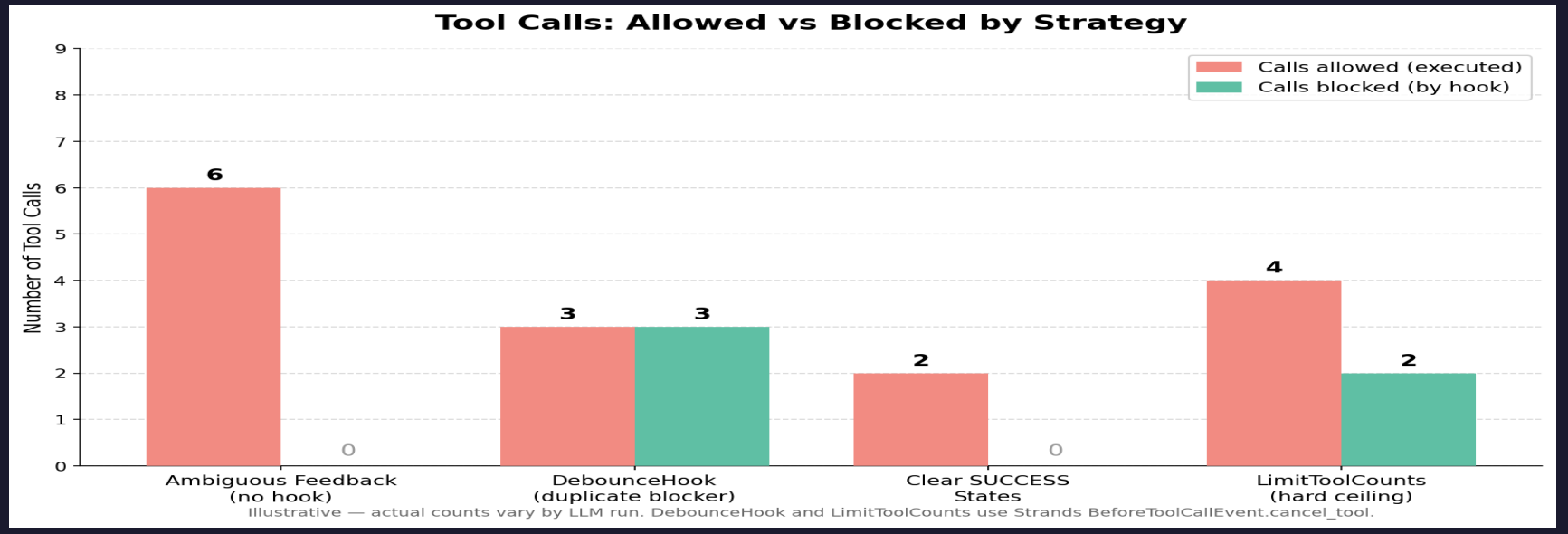
The Decoder: Language models can overthink

Problem

- Ambiguous feedback:
'more results available'
Agent calls 14 times
21 seconds wasted

Solution

- DebounceHook blocks duplicates
SUCCESS/FAILED states clear
Agent knows when to stop
14 calls → 2 calls



Demo Time

DebounceHook: 14 calls → 2 calls, 21s → 4s

Key Takeaways

- Implement the Memory Pointer Pattern for large tool outputs

Apply async handleId pattern to all MCP tools that call external APIs

Design tool responses with clear SUCCESS/FAILED states

Evaluate which failure mode is causing your agent's token waste

Deploy production-ready code from the open-source repository

Resources

GitHub Repository

github.com/aws-samples/sample-why-agents-fail

Blog Post

dev.to/aws/why-ai-agents-fail-3-failure-modes

Research Papers

IBM, Octopus, The Decoder (links in repo README)

Thank You!

Elizabeth Fuentes Leone

elifuentes.tech

Resources

bit.ly/4cRzkvJ

Blog & Socials

elifuentes.tech