IMPROVING CRAWLING WITH AN LLM

DANIEL GOLDBERG

TENZAI

whoami - Daniel Goldberg

- Tenzai - Security Researcher
- BlueHat IL - Content chair
- Guardicore - Infection Monkey (Pre AI BAS)
- Vuln Research and Exploitation for nearly 20 years
    - Windows
    - Browsers

- @ Tenzai - Evaluations, prototyping and hacking stuff

WHERE CRAWLERS BREAK

1

LOGIN

Session state and gated routes.

2

SAFETY

Blocked or destructive paths.

3

WIZARDS

Stateful, branching user journeys.

4

MULTI-STEP FLOWS

Interactions spread across pages.

REPLACE EVERYTHING WITH AI?

1

AI EVERYWHERE

Slow. Expensive. Hard to trust when the task is deterministic.

2

DETERMINISTIC CODE

Fast, predictable, and easier to trust for well-defined work.

USE AI FOR FUZZY PROBLEMS

1

AUTHENTICATION

Navigate login and recover session state.

2

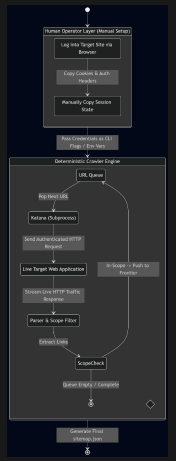
SAFETY

Recognize blocked and destructive paths.

THREE CONCEPTS FOR ADDING LLMS

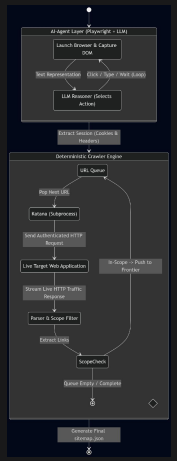
1

MANUAL SETUP



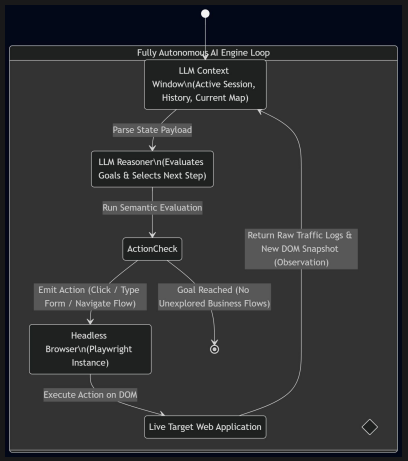
2

AI-ASSISTED



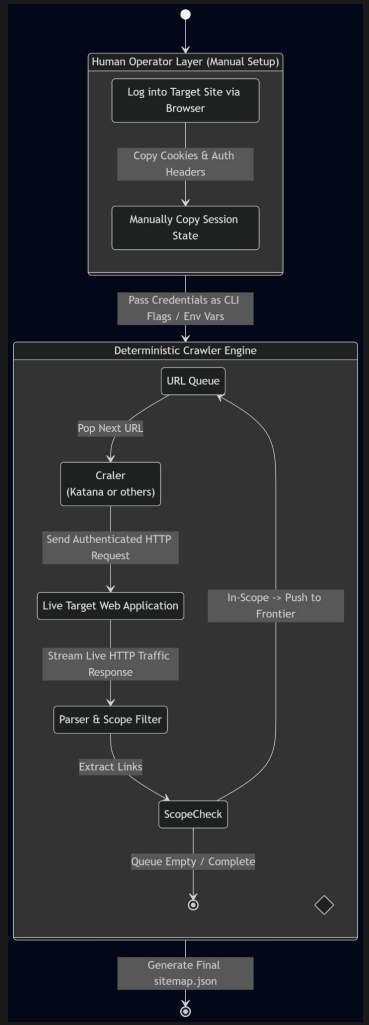
3

FULLY AUTONOMOUS



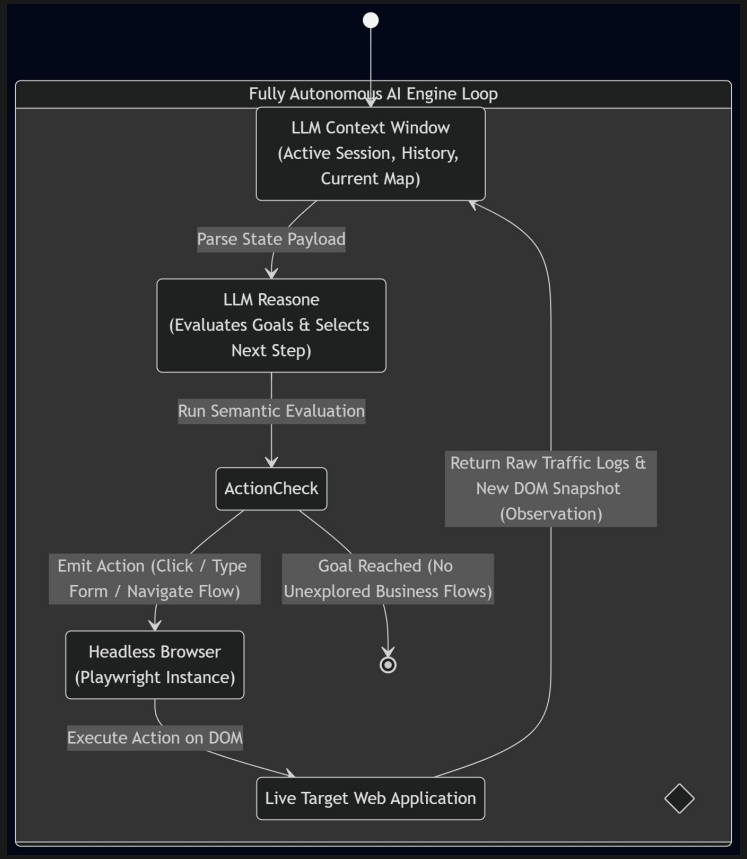
Fully Deterministic

- Requires manual work
- Hard to handle auth, wizards, forms, etc.
- Deterministic
- Cheap!
- Guardrails
  - Defaults work…sometimes
  - Manual work to add guardrails



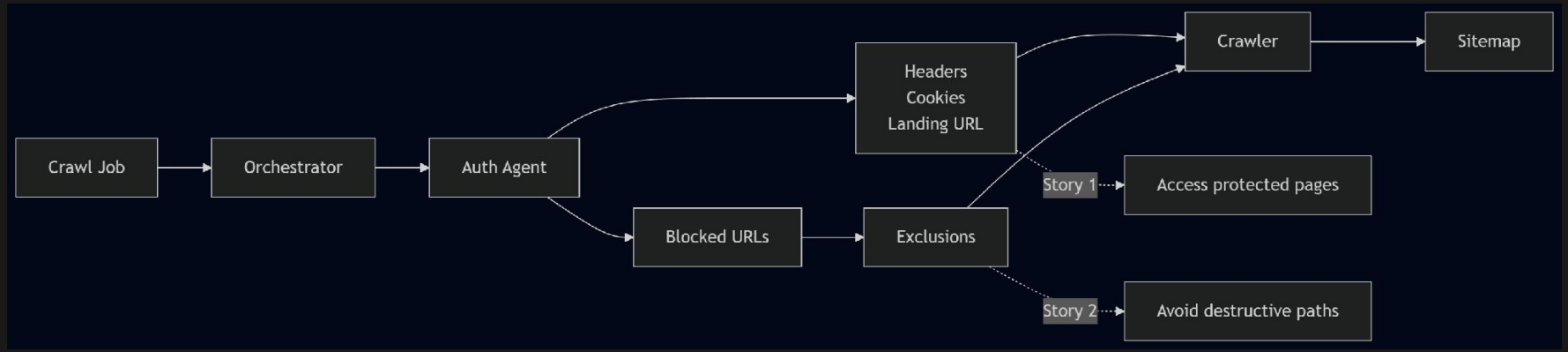
Fully LLM

- Handles “everything”
- Non deterministic
- Guardrails?
  - We don’t need no guardrails
  - Proxy as guardrails
- Expensive



MY SOLUTION: ONE LLM AGENT

A simple, non-SOTA model produces the minimum inputs the crawler needs.



Authentication headers + exclusion ruleset

WHAT CAN WE MEASURE?

1 **COVERAGE**

How much of the application did we reach?

2 **SAFETY**

Did we avoid blocked or destructive URLs?

3 **SPEED**

How long did it take?

4 **RELIABILITY**

Can we trust it to work?

RESULTS

1 COVERAGE

3× more coverage on test sites.

2 SAFETY

Avoided destructive URLs revealed after login.

3 SPEED

≈50 seconds— 10× slower than non-LLM crawling.

4 RELIABILITY

Depends on the underlying LLM.

DONE? NOT YET.

1

WIZARDS + WORKFLOWS

Still haven’t solved stateful flows

2

ITERATIVE SAFETY

Safety needs feedback across more pages.

3

BURP + CAIDO

Match the formats already in use.

PRACTICE

github.com/TenzaiLabs/crawlee

1

TEST #1

https://quotes.toscrape.com/js/
https://quotes.toscrape.com/login

2

TEST #2

https://qscrape.dev/

https://www.scrapethissite.com/pages/

3

SITE #3??

Time to test it yourself!

AI AUGMENTS.
IT DOESN’T REPLACE.

github.com/TenzaiLabs/crawlee