

From Entertaining to Useful: LLMs for Test Tooling

Rikard Edgren

www.inera.se

 **inera**
Ett företag inom SKR

BankID med QR-kod



Latest examples

QR Code Duplicate Detector

420
CODES STORED

✓
ALL UNIQUE

Export Codes

Import Codes

Clear All

Debug Mode

v1.3.0

SAML Collector

Export Filtered Import Clear

TOTAL MESSAGES:	REQUESTS:	RESPONSES:	LOGIN:	LOGOUT:
4	2	2	2	2

Validation Mode *Experimental SAML validation and best practices checks, use with care* Debug Mode

All Requests Responses Login Logout

SAMLRESPONSE Logout 2025-10-16 11:12:31

Type: LogoutResponse Status: Success

Issuer: [REDACTED]

SAMLREQUEST Logout

Type: LogoutRequest

Issuer: [REDACTED]

SAMLRESPONSE Login

SAML XML Analyzer

Analyze and compare SAML requests and responses for consistency and robustness testing

Choose Folder with SAML XML Files

Login Request 2 Login Response 2 Logout Request 2 Logout Response 2

<ds:CanonicalizationMethod>

Present in 1 file | Missing in 1 file

Attributes:

@Algorithm: (1 unique value)

<http://www.w3.org/2001/10/xml-exc-c14n#> 1

<ds:DigestMethod>

Present in 1 file | Missing in 1 file

Attributes:

@Algorithm: (1 unique value)

<http://www.w3.org/2001/04/xmlenc#sha256> 1

About

Rikard Edgren

Lives in Karlstad, Värmland, Sweden.

Three children. Long distance runner.

Works with testing digital infrastructure in Swedish healthcare.

Author of The Little Black Book on Test Design and
Den Lilla Svarta Om Teststrategi.

Member of thetesteye.com think-tank.

Øredev talk 11.11.11 11:10 on Binary Disease

rikard.edgren@inera.se



Introduction

- As most of us, I was really fascinated by the LLMs when they arrived, but they couldn't help me with my testing.
- Claude 3.5 changed this, because it was good enough for simpler programming, e.g. custom test tools!
- As a quite weak programmer, I have created own tools before, but now it is so much faster, and cost almost nothing to give a shot.
- Most tools I will show today have been done within an hour, and most of that time has been testing.
- The tools are usually very specific, built for me; here, and now.
- I use Cursor, which is an IDE that interacts with your chosen AI model, and your codebase.

a mega-fast ultra-junior developer that knows almost everything, but yet nothing

Agenda

- Intro
 - Examples
 - Learnings
 - LLM Syndromes
 - Questions
-
- My goal is that some of you soon will have a new tool to help with your testing.

When?

When do I need a tool?

- things that are difficult
- things that take time
- things that are boring

- but not if solved better in other ways (existing tool, automated test)

My Tooling Triangle

boring

automation

I don't understand enough

start wherever

takes time

custom tool

difficult

What do you need?

- Ability to recognize needs and opportunities
 - Technical understanding how it can be done
 - Some programming knowledge to notice when it's heading in the wrong direction
 - Testing skills to evaluate the tool and pinpoint problems
-
- Try!
 - Test
 - Redo
 - (Throw)

Prompt examples

- A clear idea
- Concrete details
- Examples work good
- Technical choices are welcome
- First try is most often useful

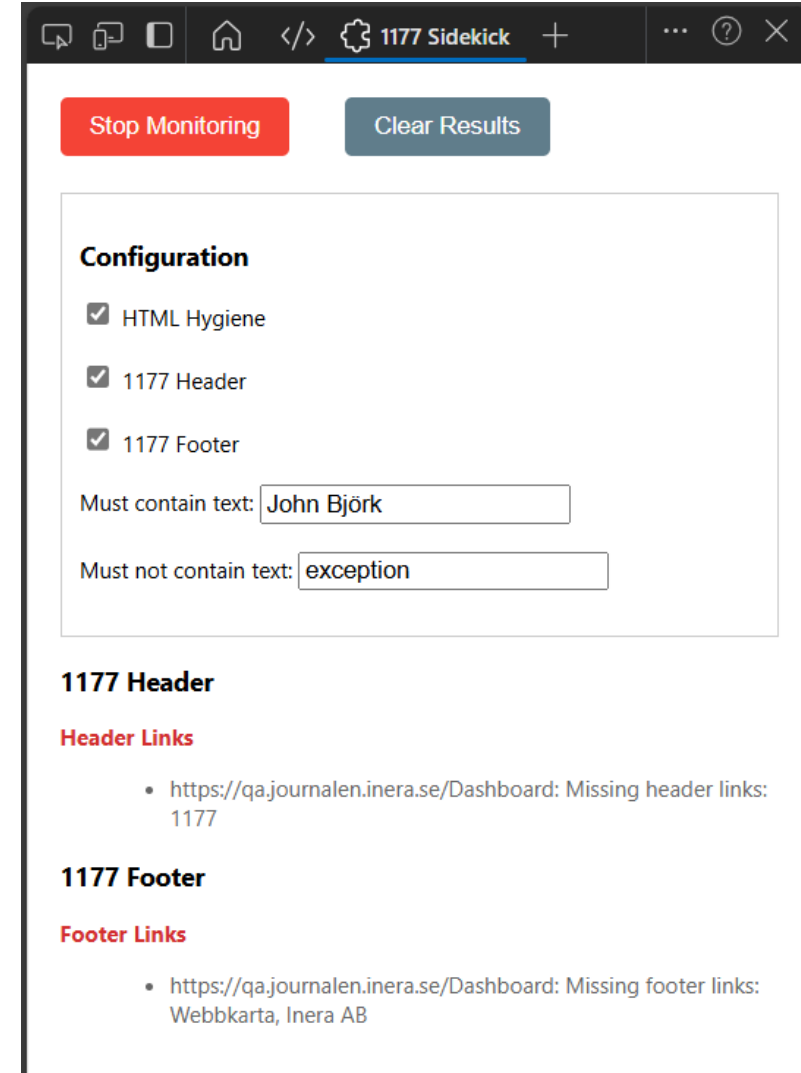
```
I want you to build a JMeter plug-in according to best practices. It should be a Listener that captures SAMLResponses in the traffic. The SAML respons should be found in the Body of the response. The plugin should then display a list of all SAMLResponses it found during the execution of the performance test. When clicking on one of the SAMLResponses captured it should decode the SAML and display as XML for the user.
```

```
I want you to build a Chrome extension according to best practices. It should be easy for me to setup and test. The extension should be a tool for testers to help them with session management in web services. I want the ability to refresh the current page every X seconds (configurable), and check if the user session is still valid. This check is verified against a text pattern that the user configures, with one edit box for "Page must include" and one for "Page must not include". I also would like to set a delay on X seconds for the verification, because some pages are not fully loaded directly after refresh. Each refresh and its results should be logged. Make a nice GUI with a popup interface. Don't bother with icons, I can create them myself. Make sure you specify the right permissions in manifest.json, I guess we might need "storage", "activeTab", "scripting", "alarms", "tabs"
```

```
I want you to build a Python script that extracts strings from a JavaScript file. The file we are working with is @main.js Strings that are interesting can be found on lines that includes "return" and after that has a quotation mark and the interesting string inside. There can be more than one string on one row. Example: return this.authenticationContextDatasource.getIsSignature() ? "V\xE4lj signeringsmetod" : "V\xE4lj legitimeringsmetod"; I also want you to adjust encodings \xe5, \xe4, \xf6 etc, so the example output is "Välj signeringsmetod", "Välj legitimeringsmetod" Put each string on a separate line in a new file. Arguments to this command-line tool is input file name and output file name. Include this prompt as a comment in the script. Add documentation of the script as comments in the end of the script.
```

Tool category – Passive scanning

- Testing for free in the background while you're at it
- Browser plugin 1177 Sidekick that verifies things that all pages should have
- OWASP ZAP plug-in that identifies swedish personal numbers



The screenshot shows the 1177 Sidekick browser extension interface. At the top, there are two buttons: "Stop Monitoring" (red) and "Clear Results" (blue). Below these is a "Configuration" section with three checked options: "HTML Hygiene", "1177 Header", and "1177 Footer". There are also two text input fields: "Must contain text:" with the value "John Björk" and "Must not contain text:" with the value "exception".

1177 Header

Header Links

- <https://qa.journalen.inera.se/Dashboard>: Missing header links: 1177

1177 Footer

Footer Links

- <https://qa.journalen.inera.se/Dashboard>: Missing footer links: Webbkartan, Inera AB

Tool category – Test improver

- Making difficult testing easier
- Session Handler keeps a user logged in for hours
- NotifyMock that checks inbox for incoming mails days later
- QR Code Duplicate Detector

Sidan måste innehålla:

Välkommen till Journalen

Sidan får inte innehålla:

Text som INTE får finnas på sidan

Intervall (sek):

300

Kontroll-delay (sek):

3

Stoppa övervakning

2025-03-14 15:35:51: Sidan uppdaterades

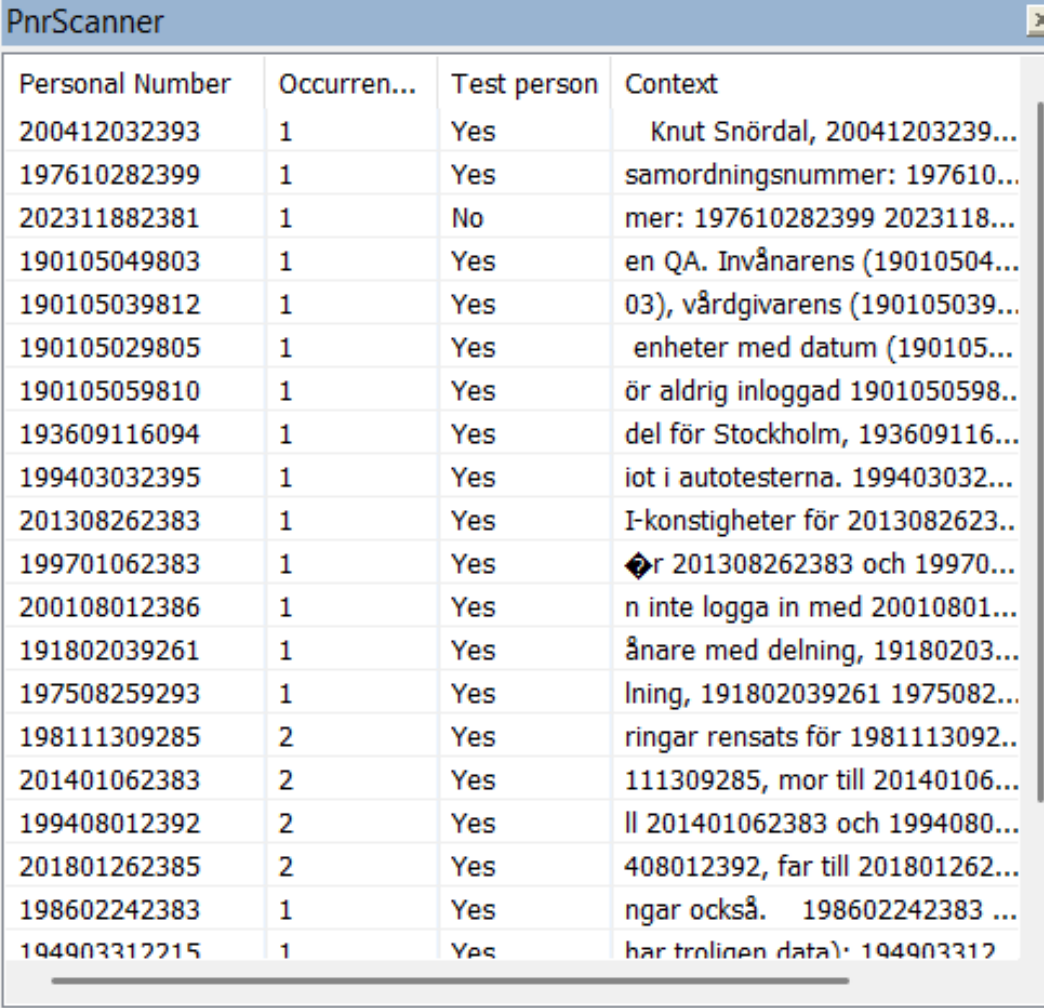
2025-03-14 15:30:51: Sidan uppdaterades

2025-03-14 15:25:51: Övervakning startad med 300 sekunders intervall

Rensa loggen

Tool category – Information Collector

- Gather information, enabling more tests (and tools)
- PnrScanner find personal numbers in Notepad++
- JMeter SAML Listener decodes SAMLResponses
- SAML Collector for further testing



The screenshot shows the PnrScanner application window. It contains a table with the following columns: Personal Number, Occurrence, Test person, and Context. Below the table are buttons for Scan, Clear, Mask, and a checked checkbox for 'Check for approved test person:'. At the bottom, it states 'Found 22 unique personal numbers' and shows system information: 342, Windows (CR LF), UTF-8, and INS.

Personal Number	Occurren...	Test person	Context
200412032393	1	Yes	Knut Snördal, 20041203239...
197610282399	1	Yes	samordningsnummer: 197610...
202311882381	1	No	mer: 197610282399 2023118...
190105049803	1	Yes	en QA. Invånarens (19010504...
190105039812	1	Yes	03), vårdgivarens (190105039...
190105029805	1	Yes	enheter med datum (190105...
190105059810	1	Yes	ör aldrig inloggad 1901050598..
193609116094	1	Yes	del för Stockholm, 193609116...
199403032395	1	Yes	iot i autotesterna. 199403032...
201308262383	1	Yes	I-konstigheter för 2013082623..
199701062383	1	Yes	ör 201308262383 och 19970...
200108012386	1	Yes	n inte logga in med 20010801...
191802039261	1	Yes	ånare med delning, 19180203...
197508259293	1	Yes	lning, 191802039261 1975082...
198111309285	2	Yes	ringar rensats för 1981113092..
201401062383	2	Yes	111309285, mor till 20140106...
199408012392	2	Yes	ll 201401062383 och 1994080...
201801262385	2	Yes	408012392, far till 201801262...
198602242383	1	Yes	ngar också. 198602242383 ...
194903312215	1	Yes	har trölinen data\ 194903312

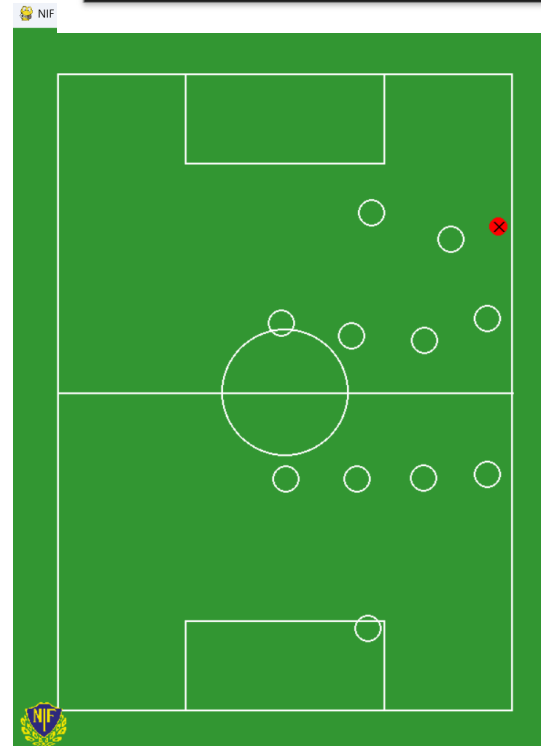
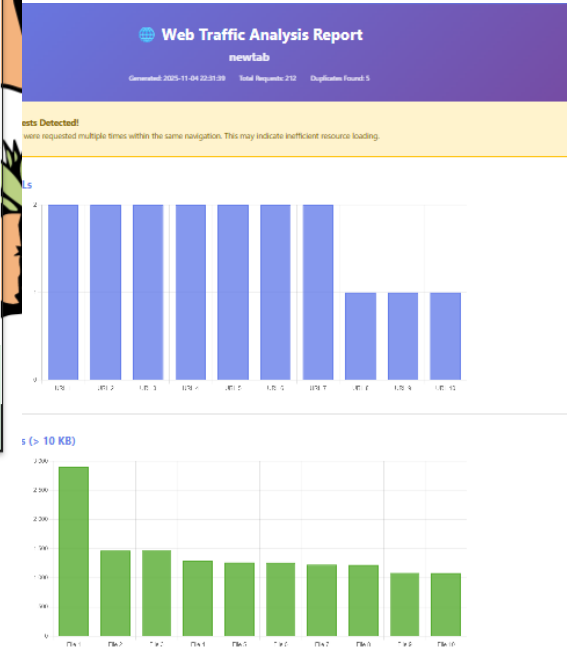
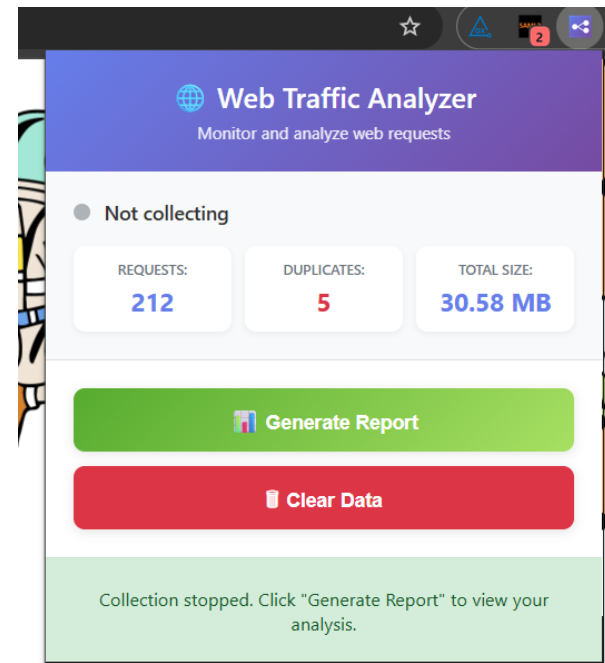
Scan Clear Mask Check for approved test person:

Found 22 unique personal numbers

342 Windows (CR LF) UTF-8 INS

Tool category – Visualizer

- Easier interpretation of information
- Web Traffic Analyzer to see outliers and duplicates
- Soccer Positional Defence



File Types Distribution

File Type	Count	Total Size	Avg Size
.img	82	20.07 MB	250.57 KB
.js	61	8.66 MB	141.97 KB
.css	20	8.6 KB	430.00 B
(no extension)	15	183.62 KB	12.24 KB
.woff	12	279.16 KB	23.26 KB
.js	9	550.69 KB	61.19 KB
.gif	5	5.72 MB	1.14 MB
.css	2	36.18 KB	18.09 KB
.css	1	892 B	892 B
unknownType	1	261 B	261 B
.css	1	605 B	605 B
.js	1	118 KB	118 KB

Duplicate Requests

URL	Request Count
https://font.googleapis.com/robots.txt/robots.txt?m=1&v=1	5
https://font.googleapis.com/robots.txt/robots.txt?m=1&v=1	5
https://www.gstatic.com/robotstxt/latest/robots.txt?m=1&v=1	5

I want to build a tool that help me understand the web traffic when I am testing web services. Quite often I see things downloaded twice, and requests that aren't used, or downloads that are way too big. This tool will make this kind of testing a lot faster (some testers don't even care)

So this should be a browser add-on according to best practices.

There should be a button "Collect web traffic", and when it is clicked, our tool should collect all web requests that are made. We need method, URL, response size, response code, and wheather the URL needed to be fetched or was available in cache.

I want you to notice when user navigates web site, and if a request is made to the same URL twice (or more) within a navigation, I want the tool to log this as duplicate requests (a possible bug)

The user should be able to stop collecting web traffic, and when that is done we should have a new button available: Generate Report

This button should create a HTML report that is downloaded for the user.

The report should contain:

A heading with the domain (base URL when session started)

A bar chart with the URL with most hits (URL as a tooltip)

A bar chart with the biggest size for downloads (ignore downloads below 10 kB)

A table with the types of files and how many such (go by the ending, e.g. .png, .js etc.)

A table with all URLs with columns and count for total number, all different response codes, number of Not Modified. This table should be possbile to sort.

Make sure you use the right permissions in manifest,json, that is often problematic.

And put some effort on the graphics, especially in the generated report, that will impress my colleagues!

Learnings

- The easier and more specific, the better results
- When code size increases, review each change (“vibe coding” in the beginning)
- If I know what I want I use Claude 3.7. If I want more creativity, I use Claude 4.5
- Better results with Python and JavaScript compared to C++ and Java
- Very strong with plug-ins and regular expressions, weaker with GUI
- Writes decent requirements, and excessive documentation
- Except the first prompt, do one thing at a time!
- Sometimes extraordinary capable
- Sometimes extraordinary stupid
- Unit tests get good with strong guidance
- Exploratory testing is important, any kind of problem can appear
- For simpler bugs, just the error details is enough
- For trickier bugs, pinpointing as much as possible
- Have not seen improvements with “AI/User rules” or .cursorrules, but GPT-5 performs remarkably better in Cursor compared to ChatGPT

Failed examples

- Documenting testing with voice recognition

```
[14:15:10] Ja, nu kör vi igång. Ska testa innan...  
[14:15:18] Yay!  
[14:15:28] Det är en kärn. Det är en kärn. Det är en kärn. Det är en kärn. Det  
är en kärn. Det är en kärn. Det är en kärn. Det är en kärn. Det är en kärn. Det  
är en kärn. Det är en kärn. Det är en kärn. Det är en kärn. Det är en kärn.  
[14:15:39] Nu ska man kolla i till hjarna Fell i prolongs ett stort fära Det  
här s can
```

- Wordpress Theme
- Auto Open DevTools



LLM Syndromes

Incuriosity	Avoids asking questions; does not seek clarification.
Placation	Immediately changes its answer whenever any concern is shown about that answer.
Hallucination	Invents facts; makes reckless assumptions.
Arrogance	Confident assertion of an untrue statement, especially in the face of user skepticism.
Manic	Rushes conversations, tends to overwhelm the user, and fails to track the state of cooperative tasks.
Indiscretion	Discloses information that it was explicitly forbidden to share.
Misalignment	Seems to express or demonstrate intentions contrary to those of its designers.
Offensiveness	Provides answers that are abusive, upsetting, or repugnant.
Incorrectness	Provides answers that are demonstrably wrong in some way (e.g. counter to known facts, math errors, based on obsolete training data)
Capriciousness	Cannot reliably give a consistent answer to a similar question in similar circumstances.
Forgetfulness	Appears not to remember its earlier output. Rarely refers to its earlier output. Limited to data within token window.
Redundancy	Needlessly repeats the same information within the same response or across responses in the same conversation.
Incongruence	Does not apply its own stated processes and advice to its own actual process. For instance, it may declare that it made a mistake, state a different process for fixing the problem, then fail to perform that process and make the same mistake again or commit a new mistake.
Negligence/Laziness	Gives answers that have important omissions; fails to warn about nuances and critical ambiguities.
Opacity	Gives little guidance about the reasoning behind its answers; unable to elaborate when challenged.
Unteachability	Cannot be improved through discussion or debate. Model must be retrained.
Non-responsiveness	Provides answers that may not answer the question posed in the prompt.
Vacuousness	Provides text that communicates no useful information.
Voldermort Syndrome	An unaccountable aversion to or obsession with certain strings of text.

Tool category – Other

- Endless possibilities!
- The Little Orange Tool for Test Design
- Selenium Video Overview
- Guess The Number porting (php Wordpress plug-in)

**Obviously you have time to spare, so
go through all web requests and look
at what is really there**

Drawbacks

- Can get too eager and look for tools when they aren't needed
- Easy to get overwhelmed by omnipotence
- You might believe you have created a generic tool for anyone, but you haven't

- Energy consumption is higher, compensated by train travelling

- Maybe I am contributing to the over-blown hype?

Ending

- The more specific, the better chance of help
- Whatever you test, there are tooling opportunities

- Does it change my job dramatically? No
- Is my testing faster? Probably not, but it is better
- Is the work more fun? Yes

- Where will all of this end?
 - It's a long, long way to programming of important, complex things
 - Even longer to testing of important, complex things

Tack!

www.inera.se

 **inera**
Ett företag inom SKR