Beyond Filter Lists: Rethinking Ad Blocking with LLMs

Maxim Topciu
Browser Extensions Team Lead at AdGuard
mtopciu@gmail.com
@maximtop

Introduction

What is this talk about?

Who I am

- Software Engineer at AdGuard for 7 years.
- I lead the team that builds our browser extensions (Ad Blocker, VPN, Assistant) for all major browsers.

What is this talk about?

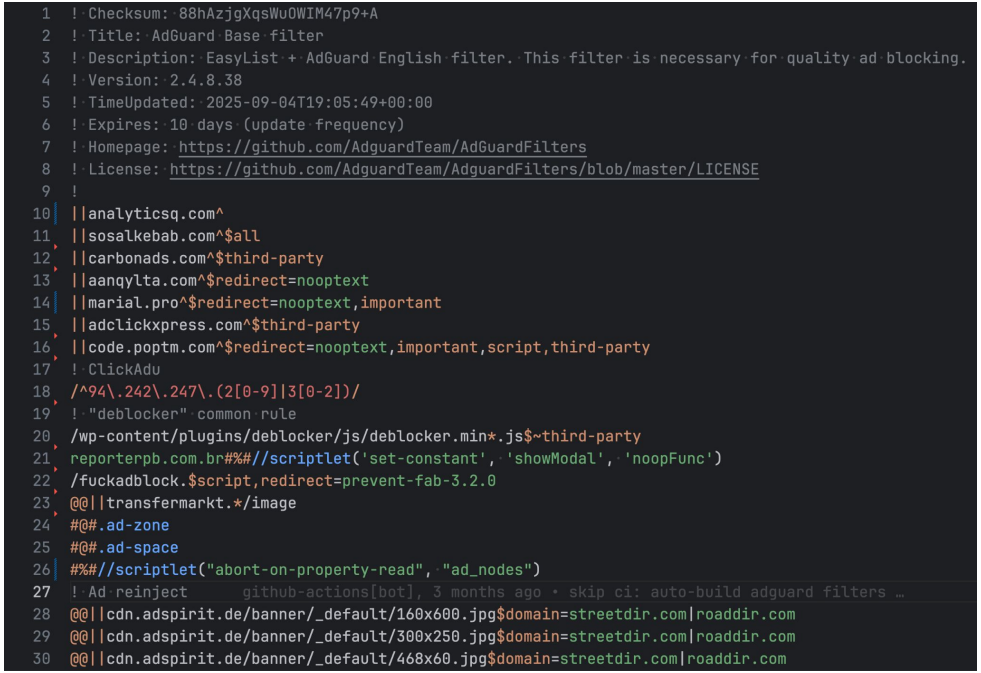
- **Filtering today:** How it works and where it fails.
- **ML's history:** A look at past attempts.
- **My experiments:** Rethinking blocking with LLMs.
- **The future:** Where this approach is headed.

Filtering Today

How it works and where it fails.

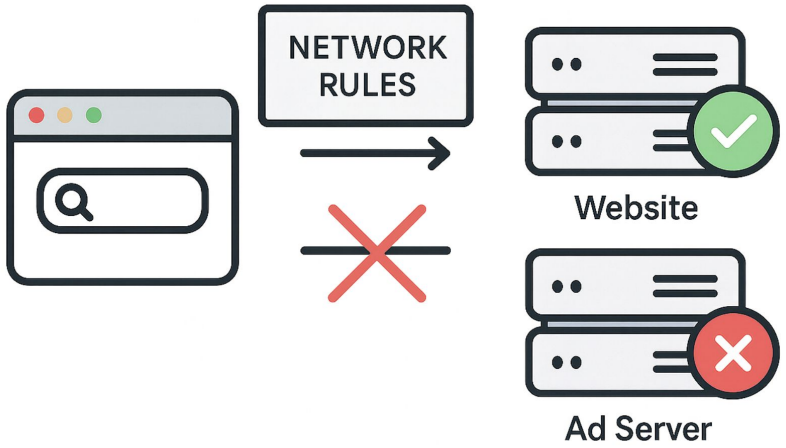
The Foundation: Filter Lists

- Ad blockers use community-maintained **filter lists**.
- Thousands of rules in two main categories:
  - Network rules
  - Cosmetic rules



Network Rules: The First Line of Defense

- **Purpose:** Block ads and trackers at the network level.
- **How it works:** Stops connections to ad servers before content is downloaded.



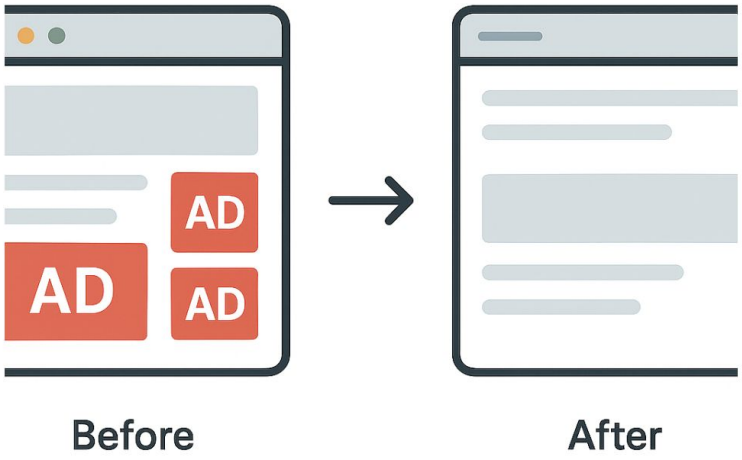
Network Rules: The First Line of Defense

- **Actions:** Block, redirect, or modify requests.
- **Example:**

```
||evil-ads.com^
```

Blocks connections to evil-ads.com and its subdomains.

Cosmetic Rules: Cleaning Up The Page



- **Purpose:** To clean up leftover ad elements that network rules can't block.
- **Actions:** Uses CSS selectors to hide unwanted elements or apply custom styles.

Cosmetic Rules: Cleaning Up The Page

- **Actions:** Hide elements or inject custom styles.
- **Example:**

```
example.com##.ad-banner
```

Hides any element with the class `ad-banner` on example.com.

Beyond CSS: Scriptlet Rules

- Handle complex scripts that CSS can't fix.
- Use JavaScript to counteract unwanted behavior.

Beyond CSS: Scriptlet Rules

- **Actions:** Modify or disable specific script functionalities on the page.
- **Example:**

```
example.com#%#//scriptlet('abort-on-property-read', 'alert')
```

Stops a script on example.com if it tries to access a specific browser feature (like `alert`)

The Power and the Limits

- **Power:** Extremely efficient and precise for known ad patterns.
- **Limits:**
  - Struggles with **native advertising**.
  - Requires constant filter list updates.
  - Manifest V3 makes updates much harder.

The Dream: No Filter Lists

- What if a blocker could decide on its own?
- No manual updates. No cat-and-mouse game.

A Brief History of ML in Ad Blocking

Why didn't ML replace filter lists?

eyeo: Project Moonshot

https://www.youtube.com/watch?v=1nJfvtvOOs0

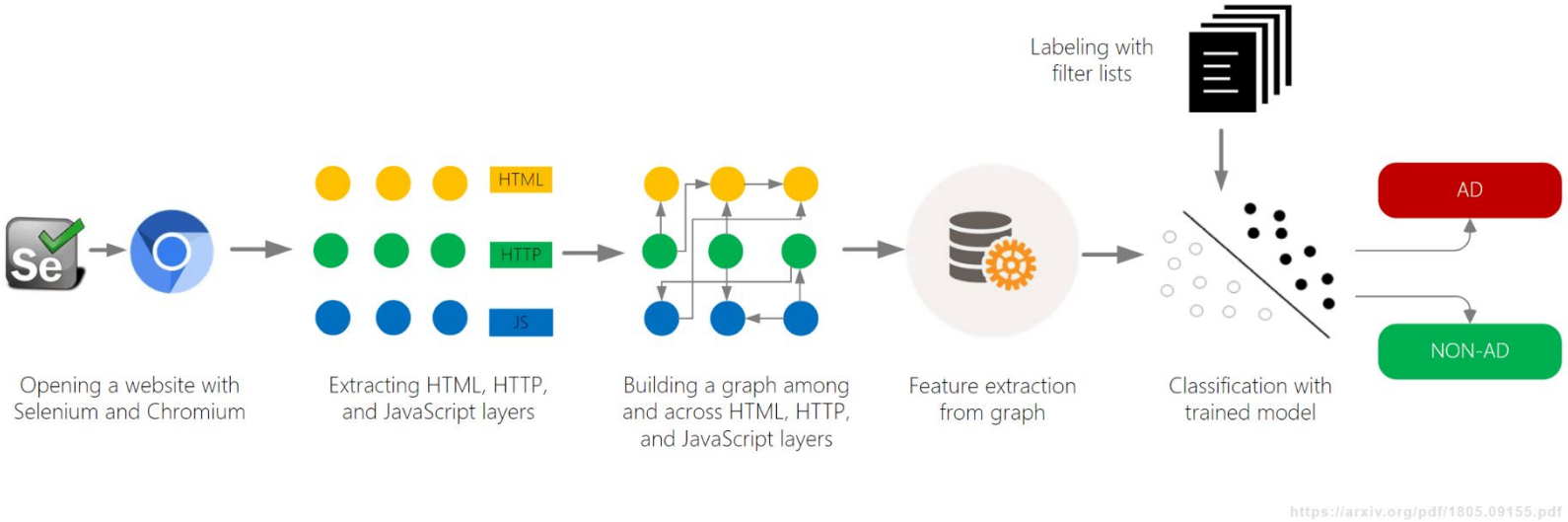
eyeo: Project Moonshot

- **Goal:** Automate cosmetic filtering at scale.
- **Method:**
  - Trained an ML model on page structure (DOM, HTML, CSS).
  - Used existing filter lists for labeling.
  - Analyzed pages directly inside the browser extension.

eyeo: Project Moonshot - Outcome

- **Result:** Predicted and hid ad elements, complementing network blocking.
- **Key Idea:** Decisions based on page structure, not images.
- **Challenges:** Data imbalance, deployment friction, and constant retraining.

Brave: AdGraph



https://arxiv.org/pdf/1805.09155

AFDS 2019

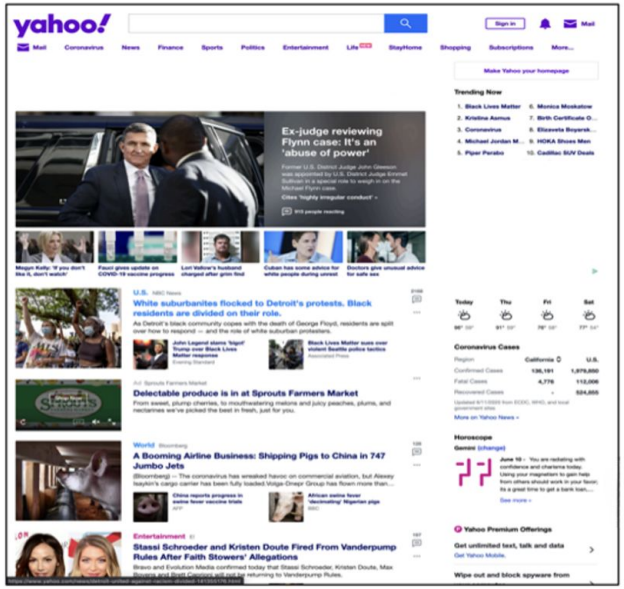
Brave: AdGraph

- **Goal:** Block ads and trackers in real-time.
- **Method:**
  - Built a graph connecting all page activities (DOM, network, JS).
  - Classified content based on its context within the graph.

Brave: AdGraph - Outcome

- **Result:** Very high accuracy (~95–98%) and robust against obfuscation.
- **Key Idea:** Decisions based on causality, not just static URL patterns.
- **Challenges:** Required deep browser integration and constant maintenance.

Brave: PERCIVAL



https://brave.com/research/files/atc20-din.pdf

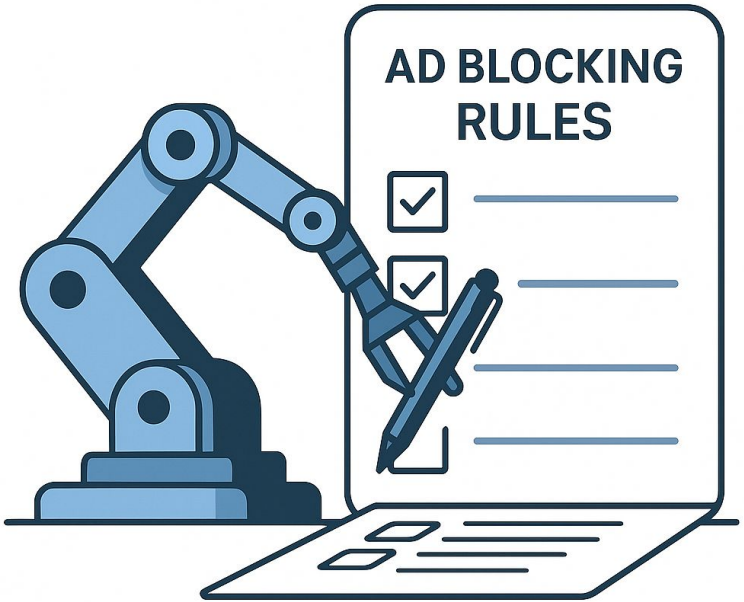
Brave: PERCIVAL

- **Goal:** Block ad images in real time.
- **Method:**
  - Used a compact neural network (CNN) to classify images.
  - Embedded it directly in the browser's image rendering pipeline.

Brave: PERCIVAL - Outcome

- **Result:** ~97% accuracy with low rendering overhead.
- **Key Idea:** Analyze an image's visual content, not just its URL or metadata.
- **Challenges:** Vulnerable to adversarial images; limited to image-based ads only.

Academic Research: AutoFR



https://www.youtube.com/watch?v=kwuIDFGzAW0

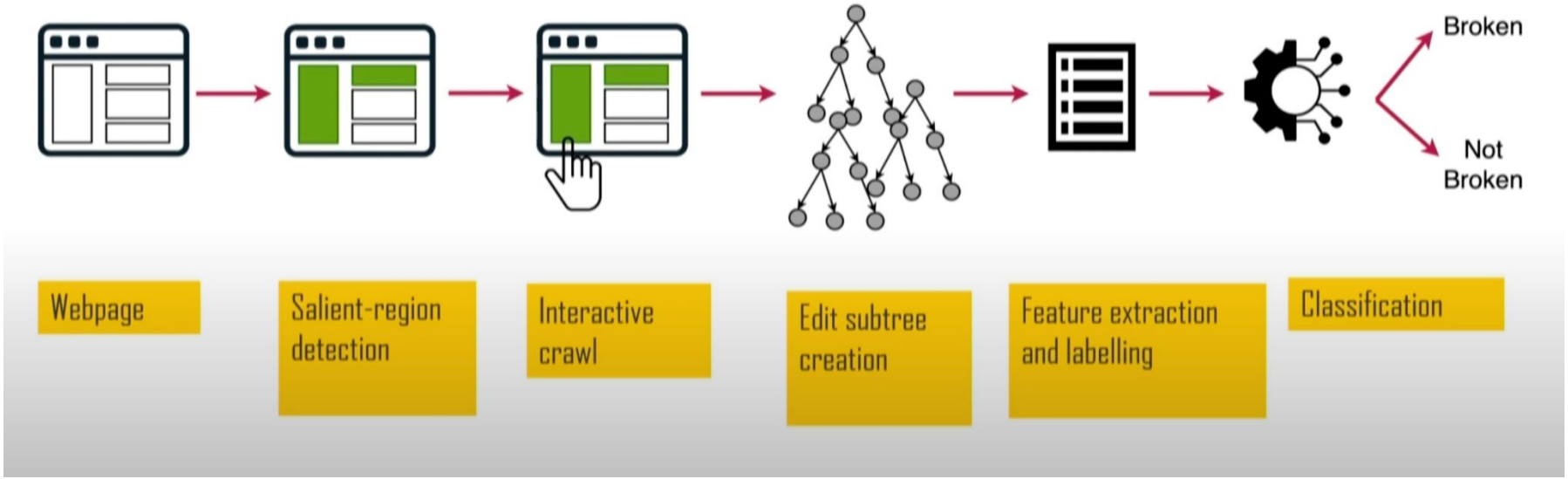
Academic Research: AutoFR

- **Goal:** Automatically generate filter rules from scratch.
- **Method:**
  - Used reinforcement learning (a trial-and-error system) to test rules.
  - Analyzed page content to avoid breaking the site.

Academic Research: AutoFR - Outcome

- **Result:** ~86% blocking effectiveness; rules generated in minutes.
- **Key Idea:** Automated rule generation with site breakage awareness.

Academic Research: SINBAD



https://www.youtube.com/watch?v=70U5BsdDIko

Academic Research: SINBAD

- **Goal:** Detect and pinpoint site breakage caused by ad blocking.
- **Method:**
  - Used "web saliency" to identify important visual elements.
  - Compared page versions (with/without blocker) to find what broke.

Academic Research: SINBAD - Outcome

- **Result:** Higher accuracy in detecting breakage with specific, actionable reports.
- **Key Idea:** Focus on user-visible impact to find and fix issues faster.

Summary: Why ML Didn't Take Over

- **High Bar:** Human-curated filter lists are extremely effective and mature.
- **High Cost:** Creating and maintaining large, high-quality datasets is expensive.
- **Evasion:** Specialized models can be vulnerable to adversarial attacks.
- **Key Takeaway:** Building specialized models from scratch is slow, expensive, and inflexible.

Enter LLMs

Big, Expensive... But Different

Rethinking Blocking with LLMs

The Power of Rapid Prototyping

About LLMs

- Appeared recently
- Provide APIs
- Actively evolving
- Cloud-based and local models available
- Extremely capable
- Expensive

About LLMs

With them, you can test ideas very quickly.

Blocking by Meaning

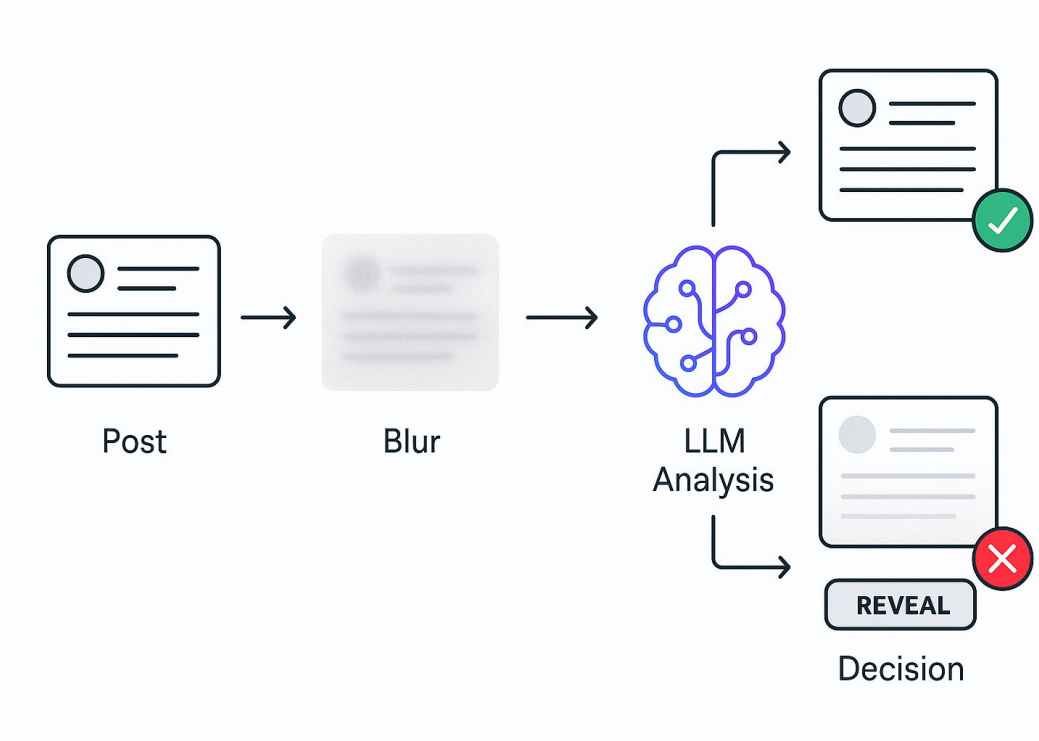
Experiment 1

Blocking by Meaning: The Motivation

- **The Question:** Can we block content based on its **meaning**, not just specific keywords?
- **The Problem:** Keyword filters are inflexible. Blocking "politics" misses thousands of related terms.

Blocking by Meaning: The Idea

1. Blur post immediately
2. LLM analyzes content
3. Unblur if safe, or keep blurred



Blocking by Meaning: The Result

- **It worked.**
- **Key Takeaway:** A new, semantic way of filtering content is possible.
- **The Proof:** This was prototyped in a few hours, a task that would have taken months with traditional ML.

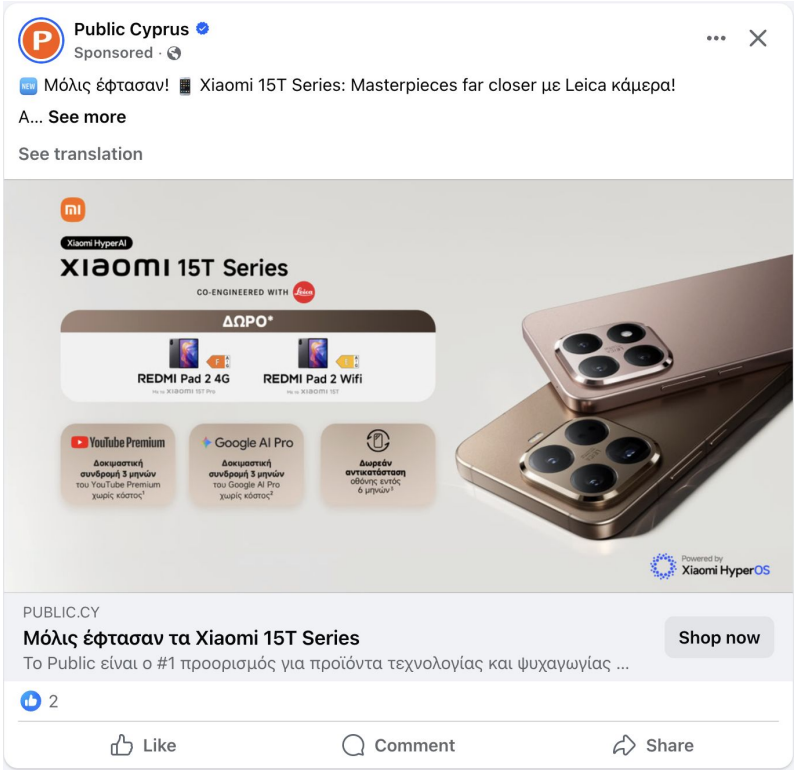
Blocking by Meaning: Demo

Blocking by Visual Meaning

Experiment 2

Blocking by Visual Meaning: The Motivation

- Posts often have minimal text



Blocking by Visual Meaning: The Motivation

- "Sponsored" labels hidden in randomized HTML

```
▸<div class="html-div xdj266r x14z9mp xat24cr x1lziwak x
  exx8yu xyri2b x18d9i69 x1c1uobl">…</div>
```

```
temp1.innerText
```

```
'Facebook\nFacebook\nFacebook\nFacebook\nFacebook\nFaceboo
k\nFacebook\nFacebook\nFacebook\nFacebook\nFacebook\nFaceb
ook\nFacebook\nFacebook\nFacebook\nFacebook\nFacebook\nFac
ebook\nFacebook\nFacebook\nFacebook\nFacebook\nAtria Musi
c. The best events in Cyprus\nd\np\nt\ns\no\nS\ne\no\nr\nn
\np\nr\nn\nu\nS\n5\ng\n5\n6\nt\nu\n8\n9\no\no\nm\n4\n1\n3
\n3\n8\n9\ne\n4\n1\n6\nc\n1\na\n0\ng\n3\nt\n5\n5\nh\n0\n4
\n4\n4\n6\ni\ns\nc\nd\n2\n5\n8\nu\nf\n \n·\nAfter all this
time? – Always!\nThe renowned orchestra Lords of the Sound
presents the symphonic show Hogwarts Magic Symphony in Cyp
rus! Ready to step into a world where… See more\nTICKETS.A
TRIAMUSIC.COM\nBuy tickets\nHogwarts Magic Symphony\nd\np
\nt\ns\no\nS\ne\no\nr\nn\na\no\nn\nu\nL\n5\ng\n5\n6\nt\nu
\n8\n9\nr\nm\nm\n4\n1\n3\n3\n8\n9\nr\n4\n1\n6\nc\n1\na\n0
\ne\n3\nt\n5\n5\nh\n0\n4\n4\n4\n6\ni\n \nc\ne\n2\n5\n8\nu
\nf\nAll reactions:\n8\n8\nLike\nComment\nShare'
```

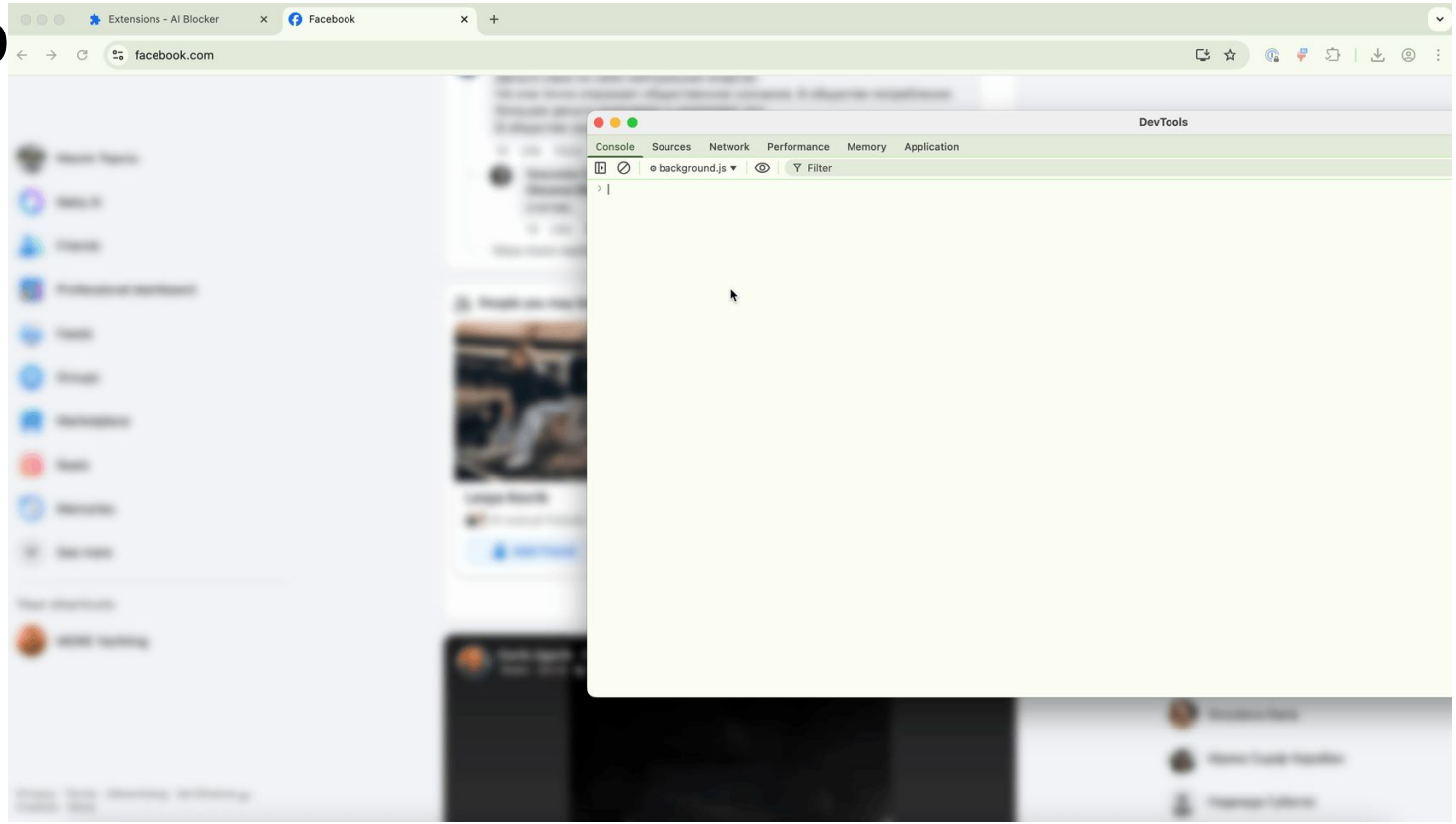
Blocking by Visual Meaning: The Idea

1. Blur post immediately
2. Vision LLM analyzes screenshot
3. Unblur if safe, or keep blurred

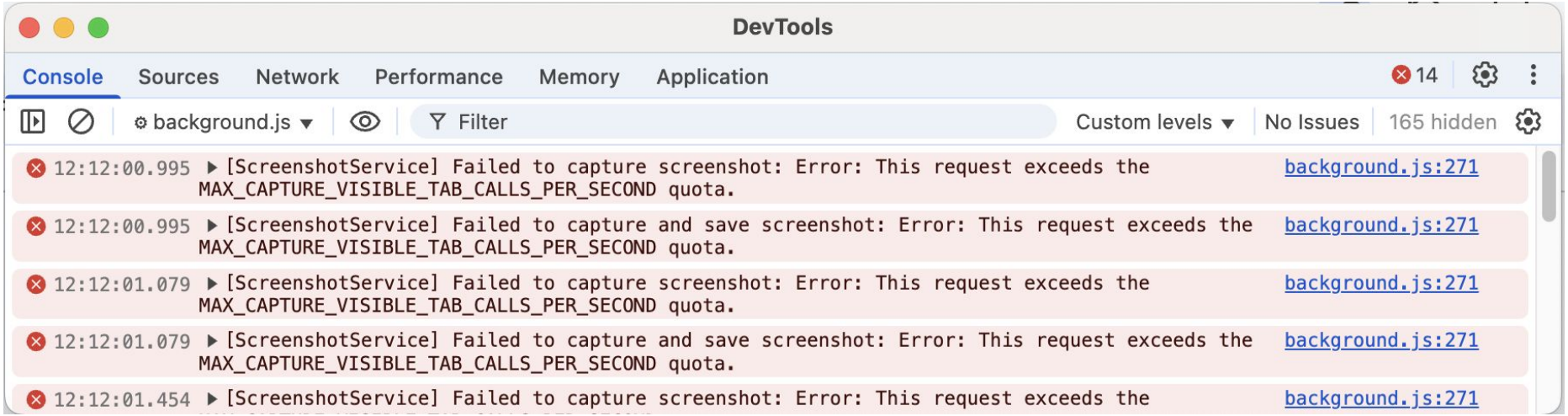
Blocking by Visual Meaning: The Result

- **It worked,** but revealed a major challenge.
- **The challenge:** Taking screenshots in a browser extension is hard.
- **The speed:** The core idea was prototyped in about 3 hours.

Blocking by Visual Meaning: Debug API App



Blocking by Visual Meaning: Built-in API Limitations



Extending Filter Lists: A New Primitive

Experiment 3

Extending Filter Lists: The Motivation

- **The goal:** Generalize the power of LLMs into a reusable tool for filter list authors.
- **The problem:** Writing a custom extension for every semantic task is not scalable.

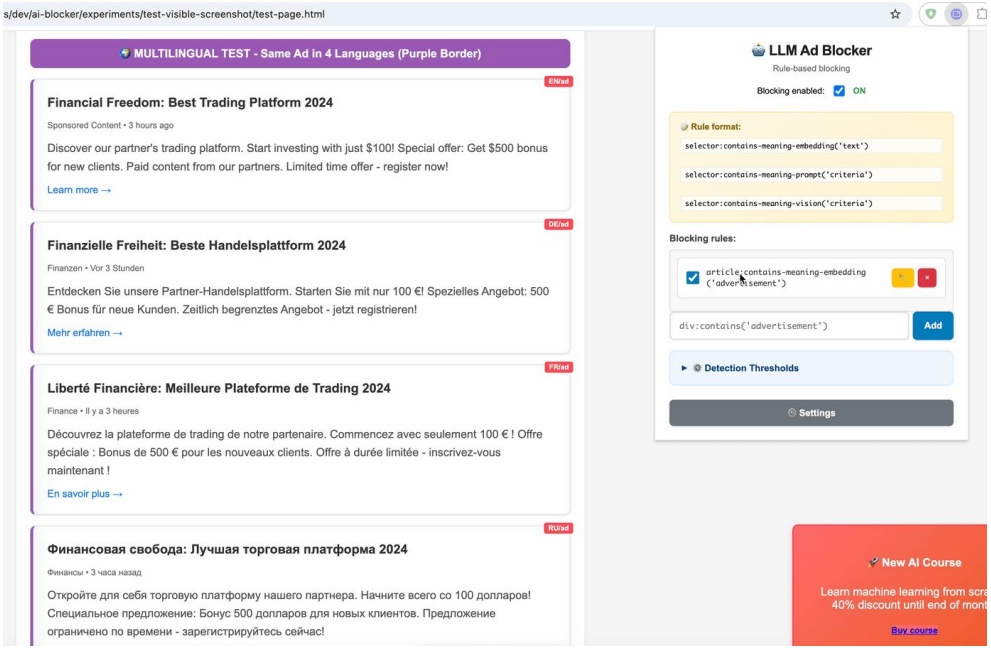
Extending Filter Lists: The Idea

- **The inspiration:** Extended CSS pseudo-class, `:contains`.
- **The question:** What if we could check for meaning, not just text?
- **The result:** Three new experimental pseudo-classes:

```
selector:contains-meaning-embedding('criteria')
selector:contains-meaning-prompt('criteria')
selector:contains-meaning-vision('criteria')
```

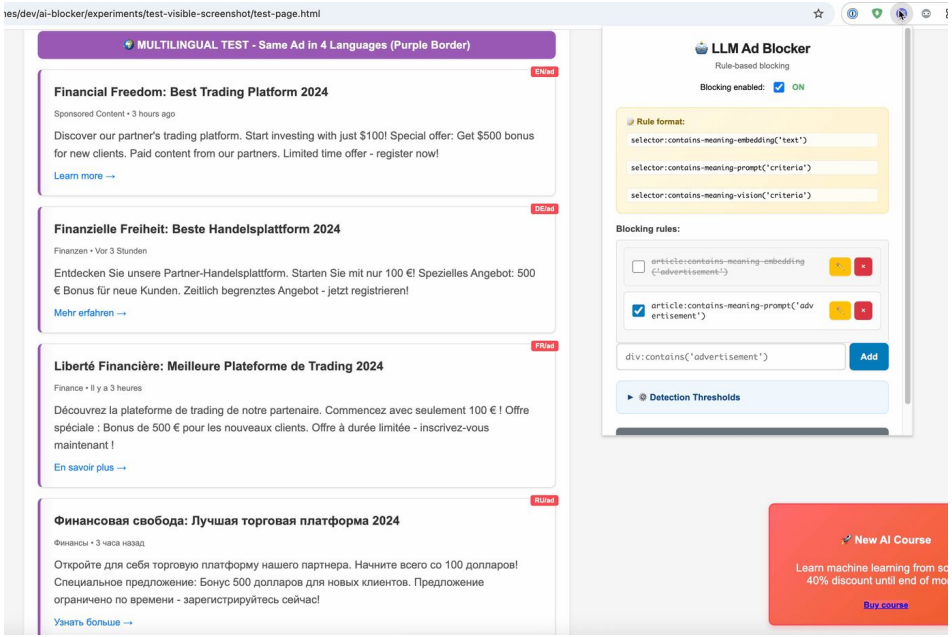
:contains-meaning-embedding

- **How it works:** compare similarities between text and criteria
- **Pros:** Very fast and cheap.
- **Cons:** Requires setting thresholds and struggles with multiple languages.



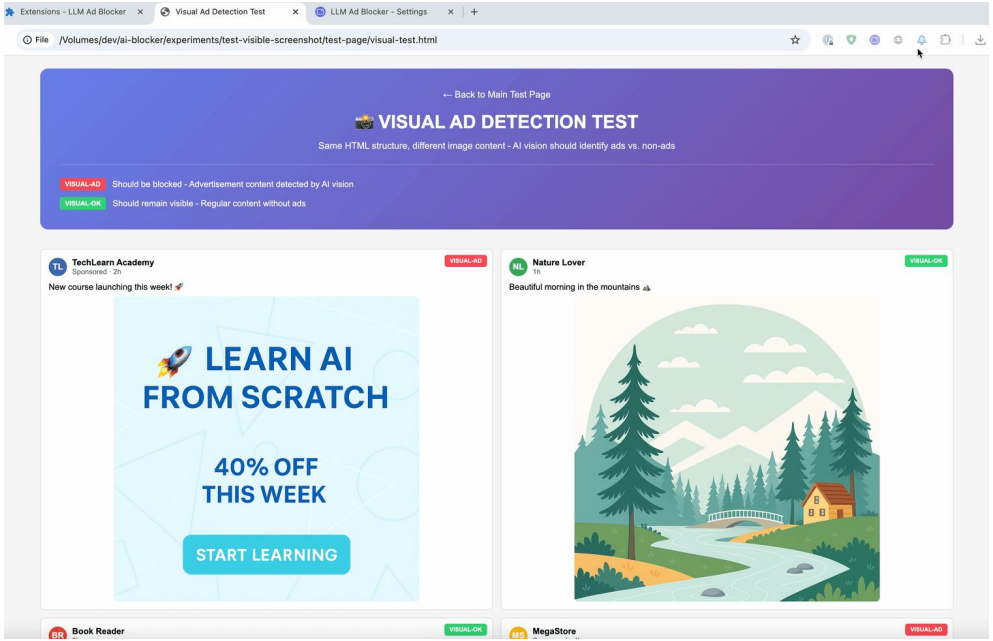
:contains-meaning-prompt

- **How it works:** ask LLM if content matches criteria
- **Pros:** More accurate, no thresholds, language-agnostic.
- **Cons:** Slower and more expensive.



:contains-meaning-vision

- **How it works:** ask LLM if screenshot matches criteria
- **Pros:** Catches things text and embeddings miss.
- **Cons:** Complex UX



Extending Filter Lists: The User Experience

- **The result:** A flexible new tool for filter authors.
- **The UX:** To handle the analysis delay, elements are blurred first, then either un-blurred or kept hidden.

Performance & Cost Analysis

Embeddings

Model	Latency	Accuracy*	FP**	FN**	Cost
text-embedding-3-large (OpenAI)	548ms	60.0%	0%	85.7%	$0.0006
qwen3-embedding-0.6b (local)	62ms	100%	0%	0%	FREE

* Accuracy = (TP + TN) / (TP + TN + FP + FN)

** TP: Correctly blocked ads | TN: Correctly kept non-ads | FP: Incorrectly blocked non-ads | FN: Incorrectly kept ads

Prompts

Model	Latency	Accuracy*	FP**	FN**	Cost/Request
OpenAI GPT-5 Nano	4,146ms	100%	0%	0%	$0.000178
Gemini 2.5 Flash Lite	829ms	100%	0%	0%	$0.000060
Gemini 2.5 Flash	982ms	90.5%	19.0%	0%	$0.000224
Claude 3 Haiku	955ms	69.0%	61.9%	0%	$0.000169
Chrome Gemini Nano (local)	1,484ms	66.7%	66.7%	0%	FREE
Gemma 3N (local)	1,588ms	69.0%	61.9%	0%	FREE

* Accuracy = (TP + TN) / (TP + TN + FP + FN)

** TP: Correctly blocked ads | TN: Correctly kept non-ads | FP: Incorrectly blocked non-ads | FN: Incorrectly kept ads

Vision

Model	Latency	Accuracy*	FP**	FN**	Cost/Request
Gemini 2.5 Flash Lite	10,697ms	100%	0%	0%	$0.000064
Claude 3 Haiku	12,862ms	98.0%	4.0%	0%	$0.000512
Gemini 2.5 Flash	12,408ms	90.0%	20.0%	0%	$0.000236
OpenAI GPT-5 Mini	14,795ms	98.0%	4.2%	0%	$0.000229
Chrome Gemini Nano (local)	14,418ms	74.0%	52.0%	0%	FREE

* Accuracy = (TP + TN) / (TP + TN + FP + FN)

** TP: Correctly blocked ads | TN: Correctly kept non-ads | FP: Incorrectly blocked non-ads | FN: Incorrectly kept ads

Methods Comparison

Method	Latency Range	Accuracy Range	Cost Range
Embeddings	62 - 548 ms	60 - 100%	FREE - $0.0006
Prompts	829 - 4,146 ms	67 - 100%	FREE - $0.00022
Vision	10.7 - 14.8 sec	74 - 100%	FREE - $0.00051

The Future of This Approach

From Experiment to Product

Future Directions

- **Vision:** Too slow now, improving over time.
- **Embeddings:** Impractical in extensions. Would be ideal if built into browsers.
- **Local prompts:** Experimental, needs better accuracy.

Summary

- **Summary:** LLMs enable semantic understanding of web content, opening new possibilities for filtering.
- **Key takeaway:** LLMs let us test complex ideas in hours, not months, dramatically accelerating research.

Thank you!

Questions?