Bot Detection and the User Experience

A Survey of CAPTCHA Design, Tradeoffs, and Failure

Nate Ferrell

Division of Science and Mathematics

University of Minnesota, Morris

Outline

1. Background & Evolution
2. Design Tradeoffs & Fingerprinting
3. The Cost of Friction
4. Practical Implications
5. LLM Arms Race & Future

Why CAPTCHAs Matter

The scale of the problem

51%

of all web traffic
is now automated

37%

of all web traffic
is malicious bots

1st time

bot traffic surpassed
human traffic (2024)

Source: Imperva Bad Bot Report, 2025

CAPTCHAs have been the primary defense since 2003. The premise: find a task easy for humans but hard for machines. That gap is closing fast.

What Are Bots and Why Do They Matter?

Not all bots are bad

Search engine crawlers, price comparison tools, uptime monitors — automated programs doing legitimate work.

Malicious bots are the problem

Automated programs mimicking human users to exploit websites at scale.

What malicious bots do:

- Credential stuffing — try leaked passwords at thousands of attempts/sec
- Account takeover — break into real accounts and drain them
- Scalping — buy out product releases in seconds
- Ad fraud — click ads to drain advertising budgets
- Web scraping — steal pricing data and content at scale

One person with the right tools can generate more traffic than tens of thousands of real users.

Background

History and evolution of CAPTCHA systems

Text-Based CAPTCHAs

The original approach

- Distorted text that OCR could not read
- OCR = Optical Character Recognition
- (technology that reads text from images)
- Held up for about a decade
- By 2014: deep learning reads distorted
- text at 99%+ accuracy — better than humans

The Cat-and-Mouse Timeline

1997 — AltaVista deploys first text-based CAPTCHA

2003 — Von Ahn et al. formalize CAPTCHAs

2009 — Google acquires reCAPTCHA

2014 — reCAPTCHA v2 (image classification)

2018 — reCAPTCHA v3 (behavioral scoring)

2020 — Cloudflare drops Google for hCaptcha

Each generation was built because the previous one stopped working.

By 2014, Google's own neural networks read distorted text more accurately than humans. The technology reCAPTCHA depended on made its own challenge obsolete.

Design Tradeoffs

How these systems actually affect users

The CAPTCHA Spectrum

More friction for users — *Less friction, more opacity*

Challenge-Based

reCAPTCHA v2, hCaptcha

Visible puzzles. Everyone gets friction. Privacy users flagged harder. We'll go deeper on this next.

Behavioral Scoring

reCAPTCHA v3

Invisible scoring. Most users never notice. Bad scores = silent denial with no explanation.

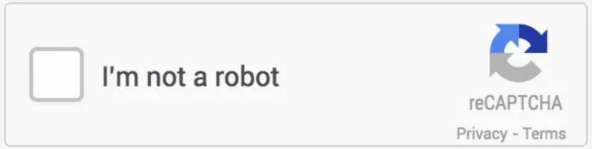
Finger-Printing Based

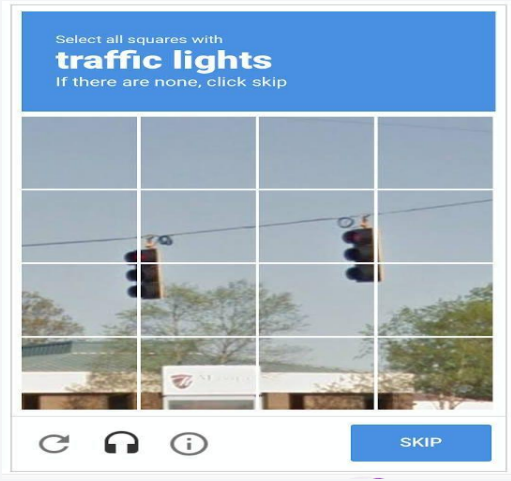
PerimeterX, Shape Security

No visible challenge. Fingerprinting + JS analysis. We'll explain how this works shortly.

Every step toward less friction shifts the cost onto a different group of users.

reCAPTCHA v2: Under the Hood





What v2 actually checks:

- Google login status & cookies
- Browser plugins & browsing history
- Mouse movement patterns
- Page navigation behavior
- The image puzzle is the last resort, not the first check

Chrome + Google account + mouse = pass every time.
Firefox + VPN + privacy settings = flagged constantly.

hCaptcha looks identical but offers a bypass cookie for users with disabilities.

reCAPTCHA v3 and the Accessibility Problem

A backend scoring layer, not a v2 replacement • AVFPU: Accessibility Value for People with Disability and Usability for all

How v3 works

Runs invisibly. Collects mouse movements, scroll behavior, typing patterns, browser data. Outputs a score 0.0–1.0 to the site owner.

Silent false positives

Bad score = denied with no explanation, no puzzle to retry, no audio fallback. Totally opaque.

Developer burden

v2 was pass/fail. v3 gives a score — many sites just pick a cutoff and block.

Accessibility (Gaggi 2022, 479 participants)

v2: **0.31**

Image puzzles nearly impossible for visually impaired. Audio fallback not reliable in many environments.

v3: **0.12**

Removes visible friction but silent denial = no fallback path at all.

Lower = worse. Less visible friction does not mean less friction.

The Intended Model: v3 + v2 Together

How it should work:

v3 scores silently

Background behavioral analysis on all traffic

→

Score-based routing

High = pass through
Low = v2 fallback

→

User has a path

Visible challenge instead of silent denial

Adoption gap (BuiltWith)

~1.2M
sites on v3

vs

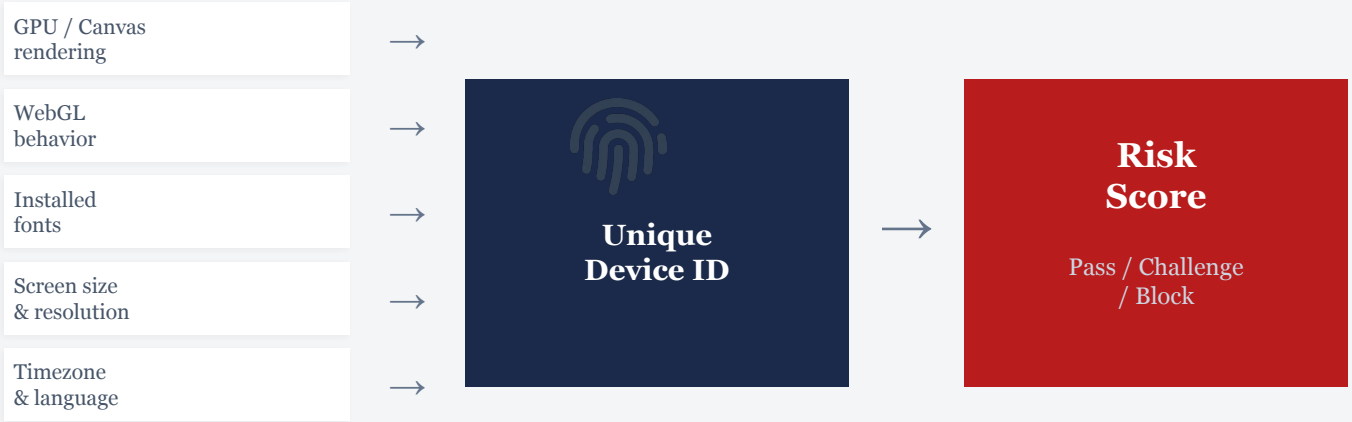
10M+
sites still on v2 alone

Most sites haven't transitioned — using v2 with all its friction or v3 with all its silent denial.

A smart setup is also dynamic — low threshold for routine traffic, tighten it for high-demand events.

What is Browser Fingerprinting?

How your device becomes an ID — no cookies needed



1.6% → 12.7%

of top sites
2016 → 2025

(Luo et al.)

"Flagging" = your fingerprint looks suspicious → higher risk score → more friction or denial

How CAPTCHAs Use Fingerprinting

reCAPTCHA

Fingerprint = one input among many

Combines fingerprint with mouse movements, cookies, login status, browsing behavior. Ambiguous score → visual challenge. You get something to attempt.

Fingerprints you to decide IF you need a puzzle.

PerimeterX / Shape

Fingerprint = the primary test

Minimal or no visual challenge. Your fingerprint IS the test.

Fingerprints you to decide if you get access at all.

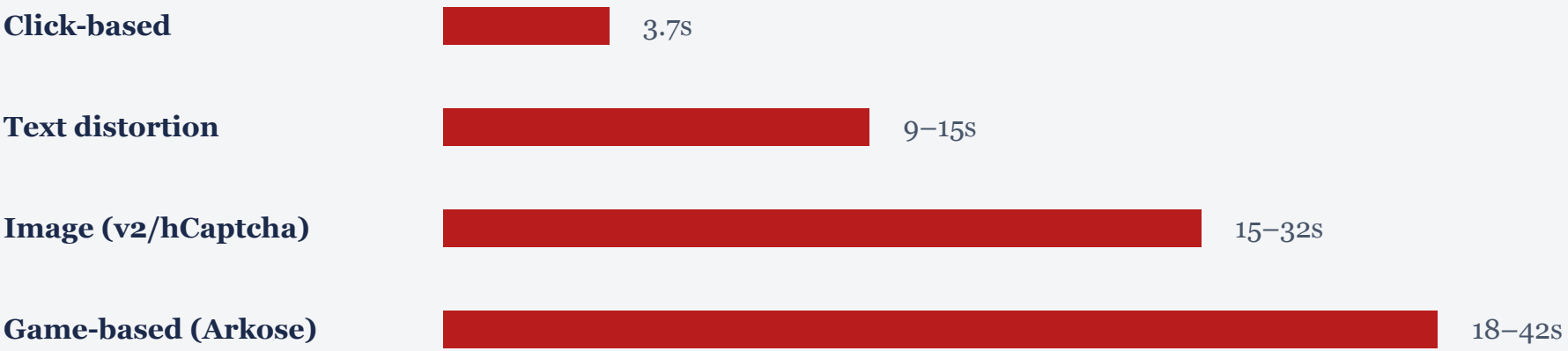
Google's Picasso system distinguishes real devices from emulators using hardware-level rendering (Laperdrix et al.)

The Cost of Friction

What happens when CAPTCHAs get in the way

Solving Times: Searles et al. 2023

1,400 participants • 14,000 tasks • Real deployed systems



Solving times varied dramatically by type. But how much worse is it in real-world conditions?

The Cost of Friction

Same CAPTCHAs. Two different settings. Very different outcomes.

Direct setting

"Here's a CAPTCHA, solve it."
No context, no goal. How almost every prior study was run.

Contextualized setting

Simulated account signup — username, email, password, then CAPTCHA.
How you actually encounter them.

30%
abandoned entirely

120%
more likely to quit
in contextualized setting

57.5%
longer solving time
in realistic context

Who Quits and When

Searles et al. secondary study • 574 participants • 174 abandoned (30%)

Direct setting

24%
low pay ($0.30)

18%
high pay ($0.60)

Contextualized setting

45%
low pay ($0.75)

28%
high pay ($1.50)

When did they quit?

Contextualized: ~50% abandoned before even attempting the first CAPTCHA

Direct: 25% abandoned before the first CAPTCHA

They are not failing and giving up. They are seeing the CAPTCHA and deciding it is not worth their time.

Practical Implications

What should sites actually do?

It Depends on What You Are Protecting

Standard Website

Blog forms, logins, sign-ups

- **v3 + v2 fallback model**
- Low threshold — most users never see a puzzle
- v2 fallback = 3.7s median (Searles)
- Avoids Gaggi's worst case (0.12 with no fallback)

High-Demand Release

Product drops, ticket sales, limited inventory

- **CAPTCHAs alone won't cut it**
- Automation adapts fast to any gimmick
- Post-processing: check orders after the fact
- Quantity limits, duplicate detection, flagging

The right answer is dynamic — adjust thresholds based on what you are protecting.

The LLM Arms Race

When the machines can solve the puzzles too

LLMs vs. CAPTCHAs: Ding et al. 2025

Zero-Shot = just ask the model to solve it • Chain-of-Thought (CoT) = tell it to reason step by step

CAPTCHA Type	Zero-Shot		Chain-of-Thought	
	GPT-4o	Gemini	GPT-4o	Gemini
Text — Simplest	77%	73%	90%	83%
Text — Overlapping	67%	60%	70%	60%
Text — Noise	70%	73%	73%	67%
Text — Noise+Overlap	37%	23%	50%	43%
Image — reCAPTCHA	40%	33%	50%	23%
Image — hCAPTCHA	40%	37%	43%	30%

Source: Ding et al. 2025, Table 1

LLMs are solving a meaningful percentage of the two most common CAPTCHA types on the web.

Game-Based CAPTCHAs

5–10x more friction than a basic CAPTCHA.

The industry's first response to LLMs

Arkose Labs, GeeTest:

Rotate objects to match angles, line up identical items, interact with 3D scenes. Interactive spatial tasks meant to stump LLMs.

18–42 seconds solving time (Searles et al.)
vs 3.7s for a basic checkbox

Reasoning CAPTCHAs

(d) 3D CAPTCHA 1 by YiDun (e) 3D CAPTCHA 2 by GeeTest

Tasks that require judgment

GeeTest, YiDun:

Require understanding context, spatial relationships, and subjective interpretation. Not just clicking — thinking.

21.7% first-try pass rate (Ding et al.)
13%+ needed three or more attempts

Harder for bots, but also harder for the people you're trying to let through.

IllusionCAPTCHA: A Different Approach

How it works (Ding et al. 2025)

A diffusion model blends a hidden image (apple) into a base image (cityscape). Humans recognize the hidden content naturally — like seeing a face in clouds. LLMs read images too literally and miss it.

0%

LLM success rate
(all strategies)

87%

Human
first-try pass rate

Beyond Puzzles: Sensor-Based Detection

Gangwal et al. 2025 — "Swiss Cheese CAPTCHA" — Accelerometer-based

Users tilt their phone to guide a ball to a target while avoiding obstacles. The system analyzes trajectory curvature and path smoothness to distinguish human movements from bots.

<7s	83%	<5%
Average solving time	Human success rate	Bot success rate

Limitations: mobile-only, Generic Sensor API not universal, motor impairment accessibility gap

Compare: 15–42s for image/game CAPTCHAs (Searles et al.) vs <7s here

Future Directions

Proof-of-Work

mCaptcha

SHA-256 computation scales with traffic. No cognitive barrier. Works with screen readers. A few seconds of background compute vs 42s clicking traffic lights.

Visual Illusions

IllusionCAPTCHA

Targets how AI processes images. 0% LLM / 87% human. Durability depends on future models — the open question.

Layered Systems

Context-aware

Different tools for different threat levels. Each fallback must test something genuinely different — not just harder.

Conclusion

People treat CAPTCHAs like they are solving a security problem, but what they are really doing is choosing which users they are okay with losing.

The traditional premise is failing.

LLMs solve 37–77% of text CAPTCHAs. The tasks hard enough to stop bots also stop real users.

The friction is real.

30% abandonment. 57.5% longer solving times in realistic contexts. Accessibility getting worse, not better.

The answer is not one system.

It is dynamic, layered, and context-aware — adjusted based on what you are protecting and who you are willing to lose.

The field has the data now. Whether it uses it is a different question.

Questions?

Bot Detection and the User Experience

Nate Ferrell • University of Minnesota, Morris

Key numbers for reference

51% bot traffic (Imperva 2025)

37–77% LLM solve rate on text CAPTCHAs (Ding et al.)

30% abandonment / 57.5% longer in context (Searles et al.)

AVFPU: v2 = 0.31, v3 = 0.12 (Gaggi 2022)

0% LLM / 87% human on IllusionCAPTCHA (Ding et al.)