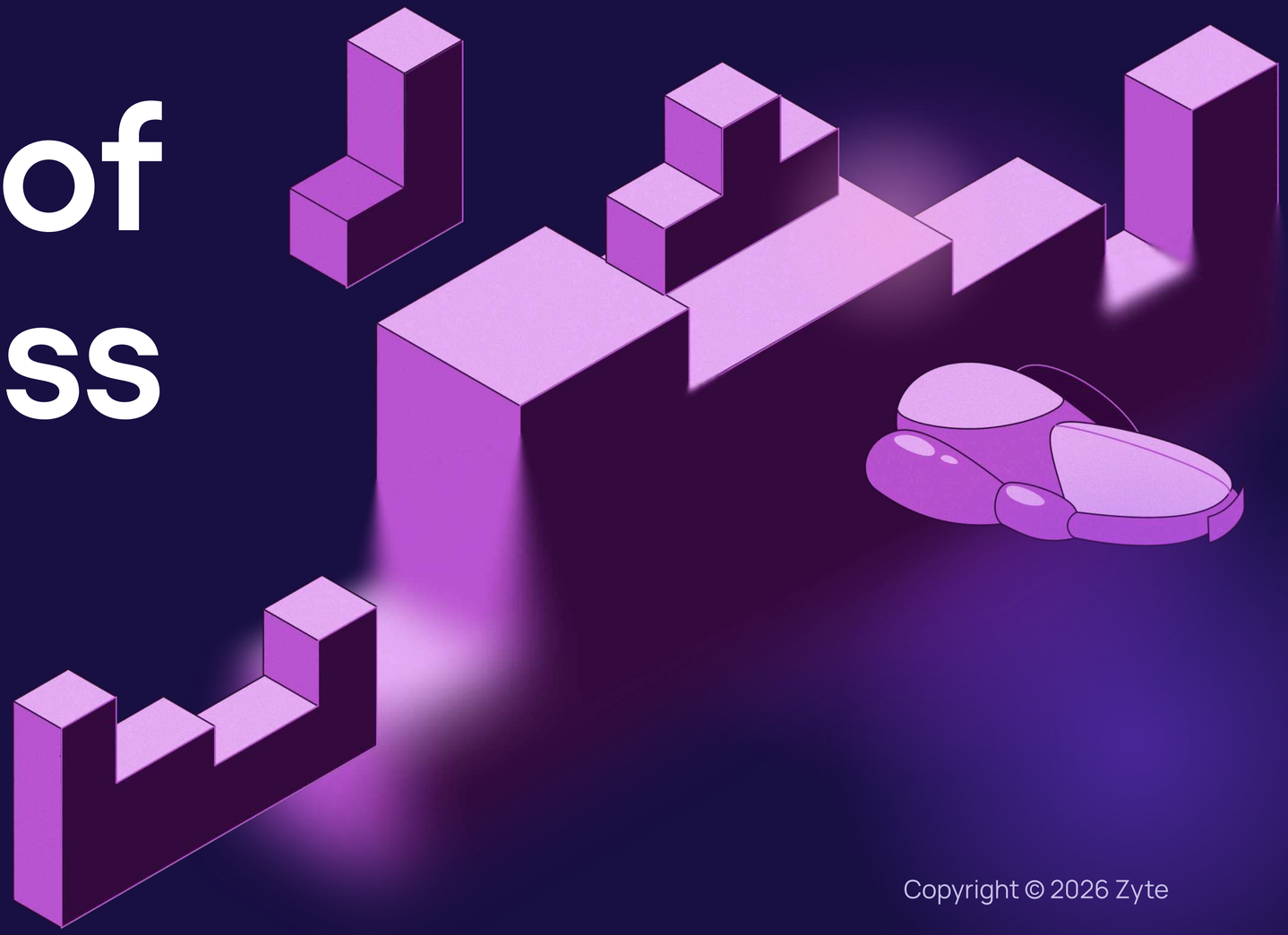
zyte

zyte

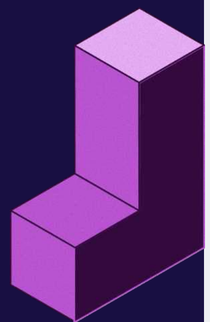
The State of Web Access

Measuring the open web

Explore the data



Contents



How open is the open web?

The internet has never been richer in data, but it is also shifting from a public commons into an ecosystem gated by new rules and new infrastructure.

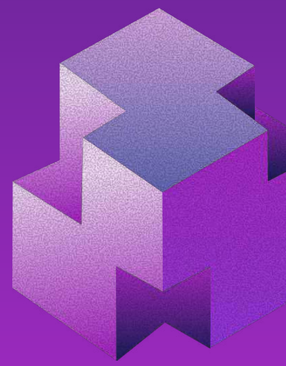
As agents and bots begin roaming the web at unprecedented scale, many websites are hardening against unwanted visitors. That is altering the practice and the economics of data access.

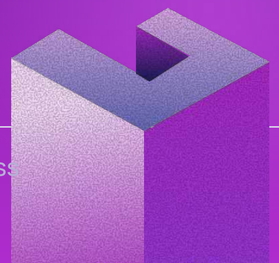
To understand this shift, we audited the 11,100 most popular landing pages across the globe and thousands more in countries around the world, to understand the specific mechanisms websites use that introduce friction at the point of access.

In total, we examined 24,898 unique websites.

We developed our own analysis algorithms, borrowed off-the-shelf tools and harnessed Zyte's own platform to analyze the entire stack: network-layer controls like web application firewalls (WAF), client-side complexity through JavaScript-heavy rendering, interactive challenges like CAPTCHA, throughput constraints via rate limiting, and specialized controls such as TLS fingerprinting and dedicated antibot services.

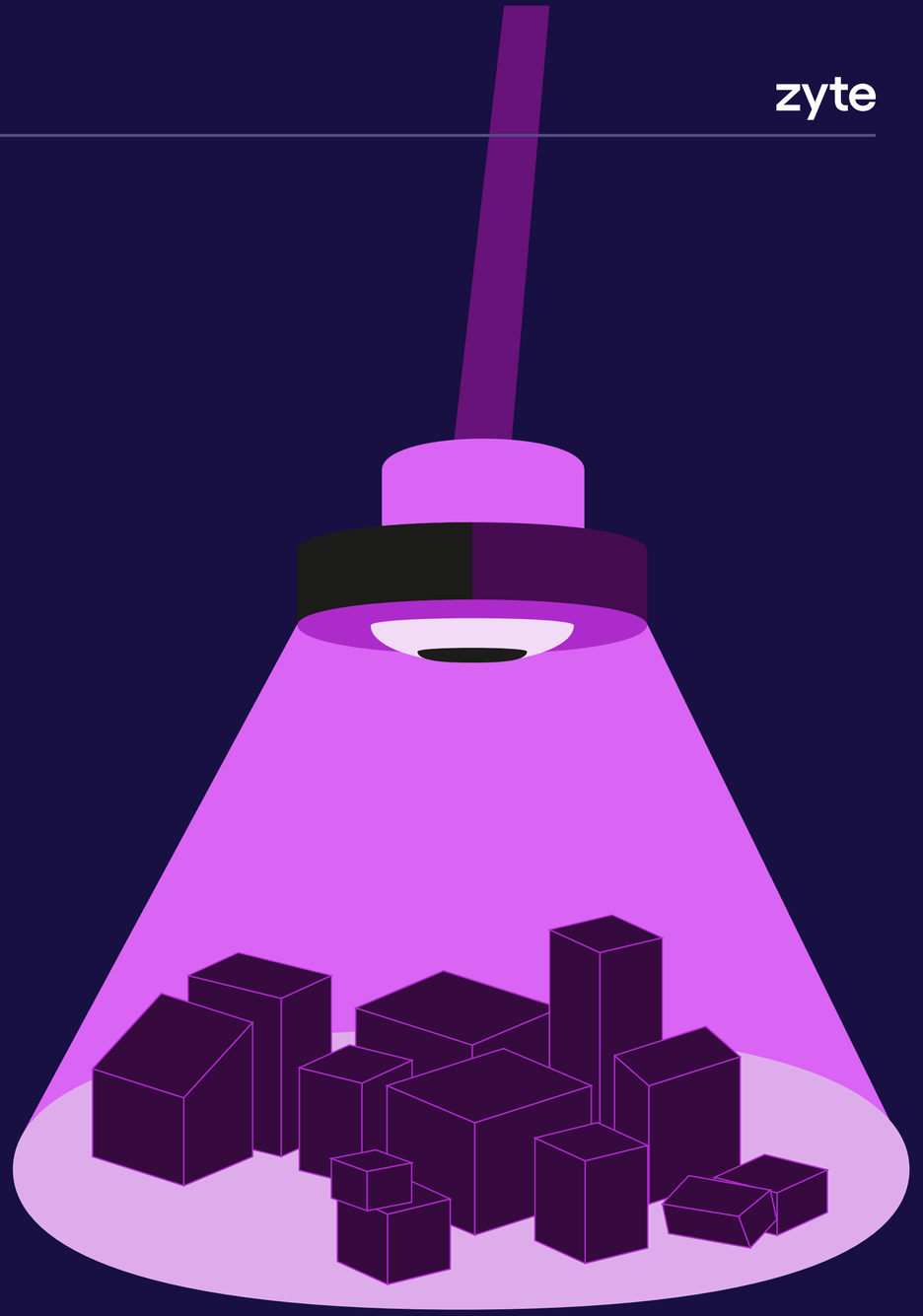
The State of Web Access is the largest audit of web access controls ever published.





zyte

Our top takeaways



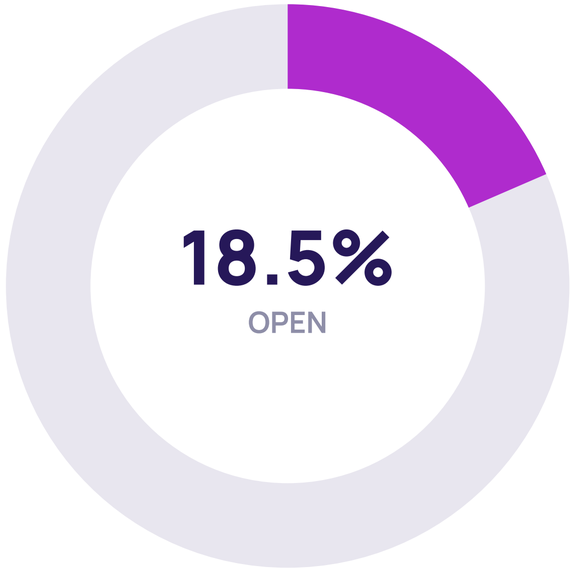
1 Open access is now in the minority

BARRIERS ARE COMMON

Today, only 18.5% of popular landing pages globally operates without any barrier-like mechanisms.

The rest, 81.5%, deploy at least one technical barrier to automated access.

OPEN ACCESS · 11,100 LANDING PAGES · 2026



2 A worldwide firewall guards the web

DIVERSE DETERRENTS

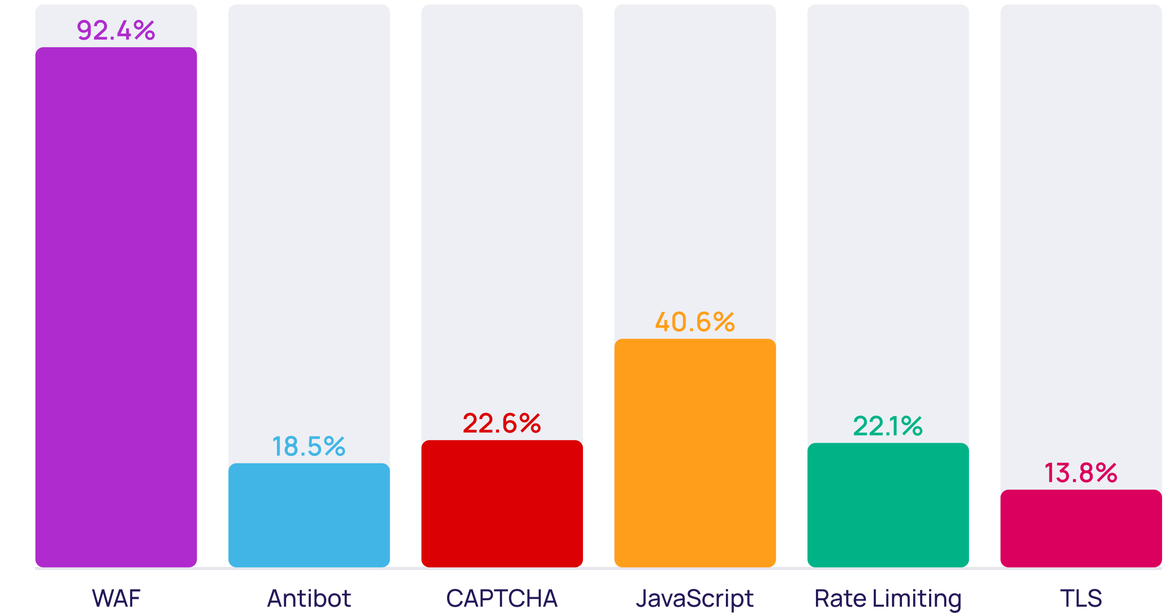
Web Application Firewalls are websites' most used access control technique.

Around a fifth deploy antibot systems, rate limiting and CAPTCHA verification.

On-page JavaScript poses a challenge to programmatic access, but this isn't always intended as a deterrent.

ACCESS BARRIER ADOPTION · 11,100 LANDING PAGES · 2026

Access control implementation by barrier type



State of Web Access · 11,100 landing pages

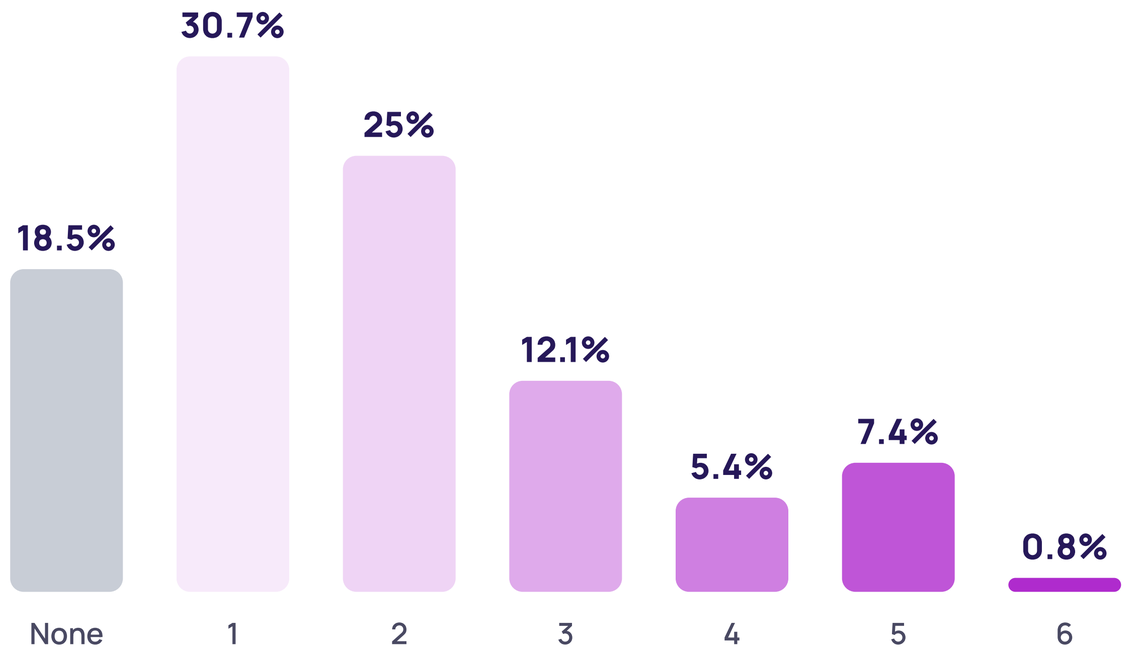
3 Access controls are multiplying

COMBINED BARRIERS

Half of all pages employ two or more access control methods.

Access difficulty depends on how multiple controls add up - and how bots respond to the combination.

BARRIER COUNT DISTRIBUTION · % OF ALL SITES

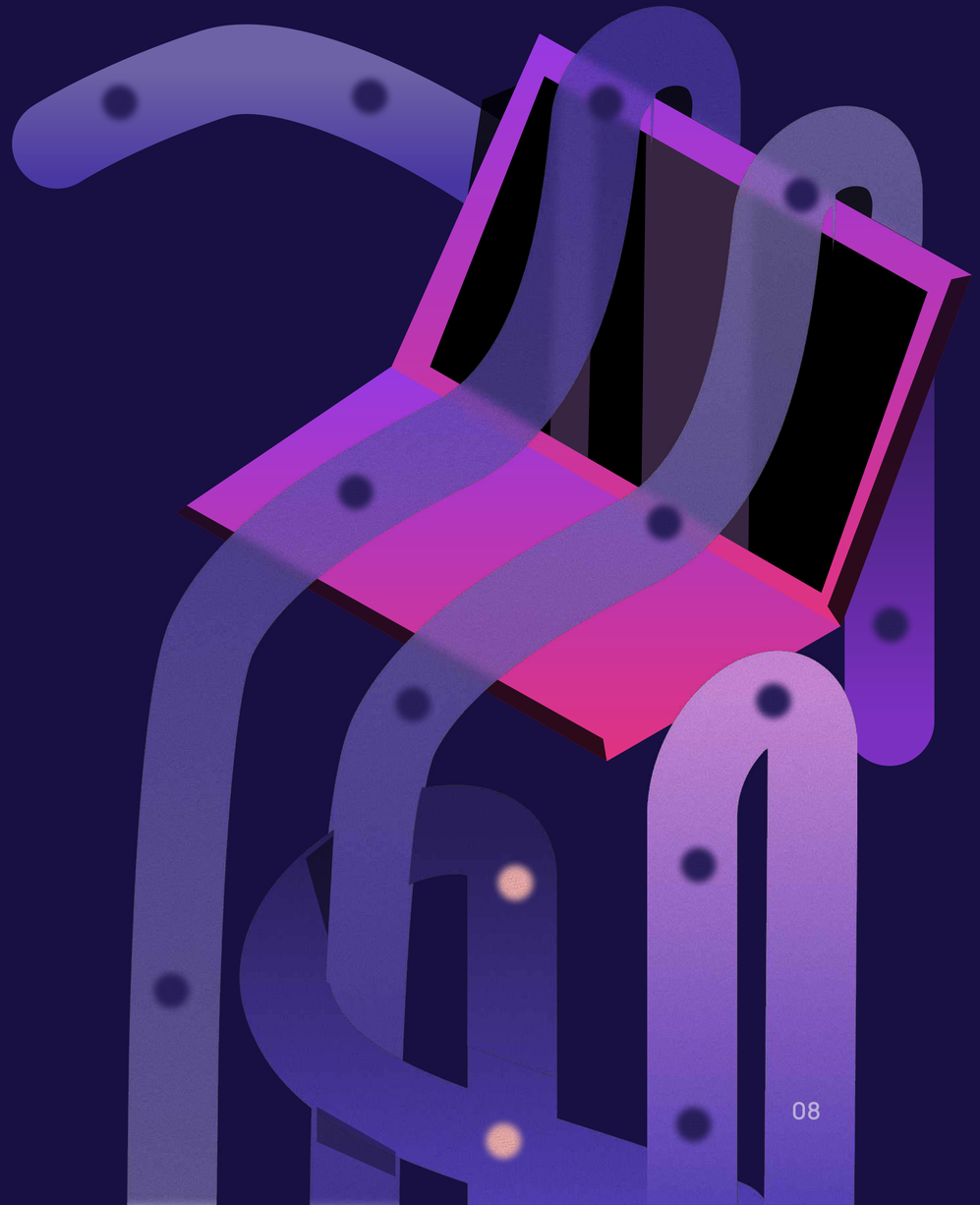


zyte

Behind the barriers

Mechanisms for controlling web access range from simple text files to sophisticated behavioral analysis engines.

Understanding these barriers requires examining their adoption rates, how they interact, and the vendor ecosystems that support them.



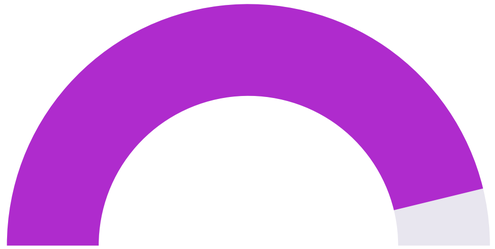
Web Application Firewall: The baseline barrier

A Web Application Firewall (WAF) is the closest thing the web has to a default security setting.

Almost two thirds (63.5%) of the top landing pages run an identifiable WAF - more than any other barrier we measured, and nearly three times the rate of dedicated antibot solutions. A further 29% sit behind generic WAF signatures that cannot fingerprint to a named vendor.

Most of these deployments, however, are not sophisticated - 62% of WAF-protected sites run a single vendor in standard mode.

A WAF is not necessarily an active strategic choice; the capability tends to come bundled with a content delivery network (CDN), and most operators leave it at the defaults.



92.4%

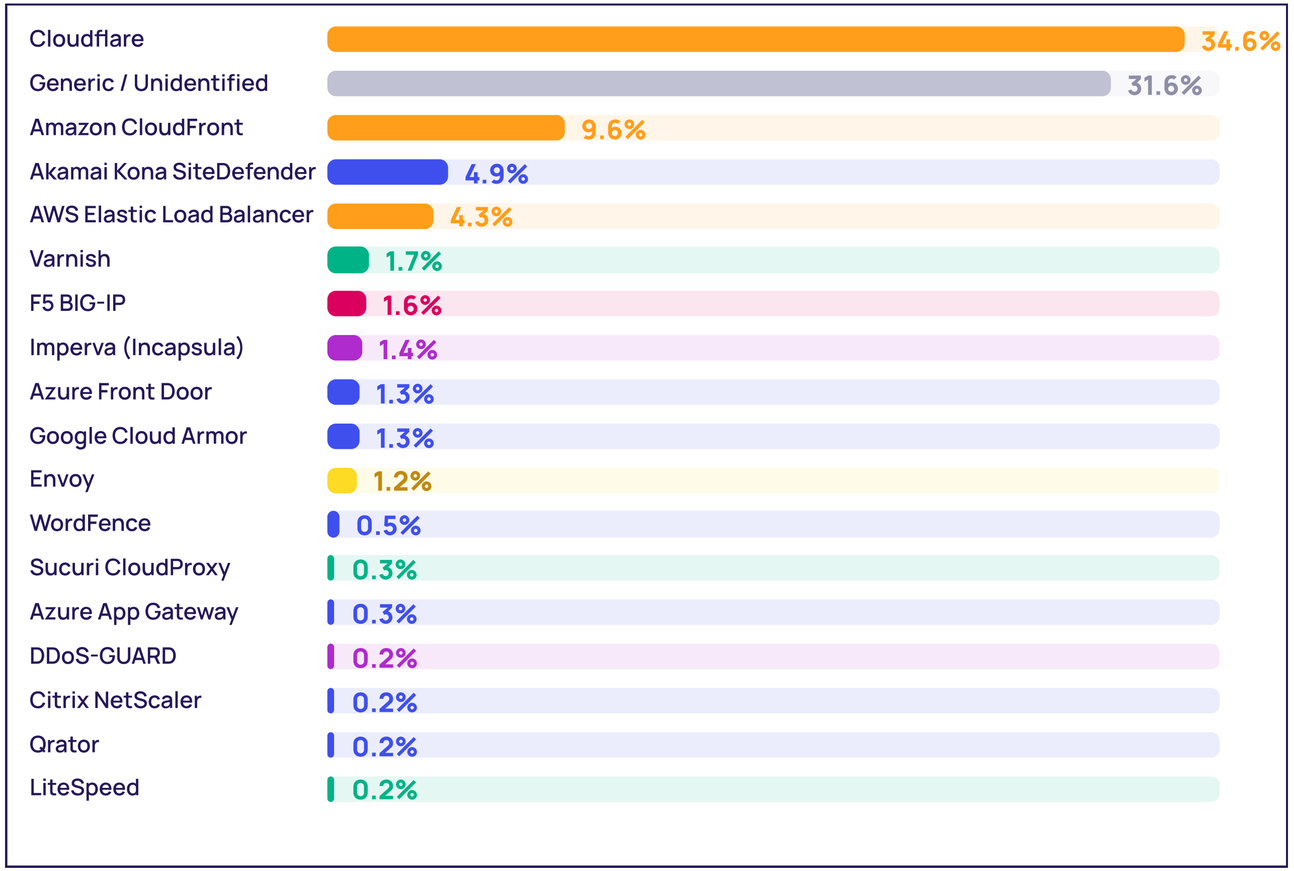
of all sites run a WAF

18 vendors provide the web's firewalls

The WAF market is more competitive than the antibot and CAPTCHA provider markets.

But it is more concentrated than the broader CDN market. The top three vendors control 71.7% of the market. This reflects bundled WAF/CDN offerings as part of the migration from on-premise to cloud infrastructure.

The WAF landscape is dominated by single-vendor deployments. 61.9% of all sites use a WAF configuration with a single vendor in standard mode. Only 1.4% of sites use more sophisticated configurations with multiple vendors or behavioral analysis.



The WAF is the web's default security setting

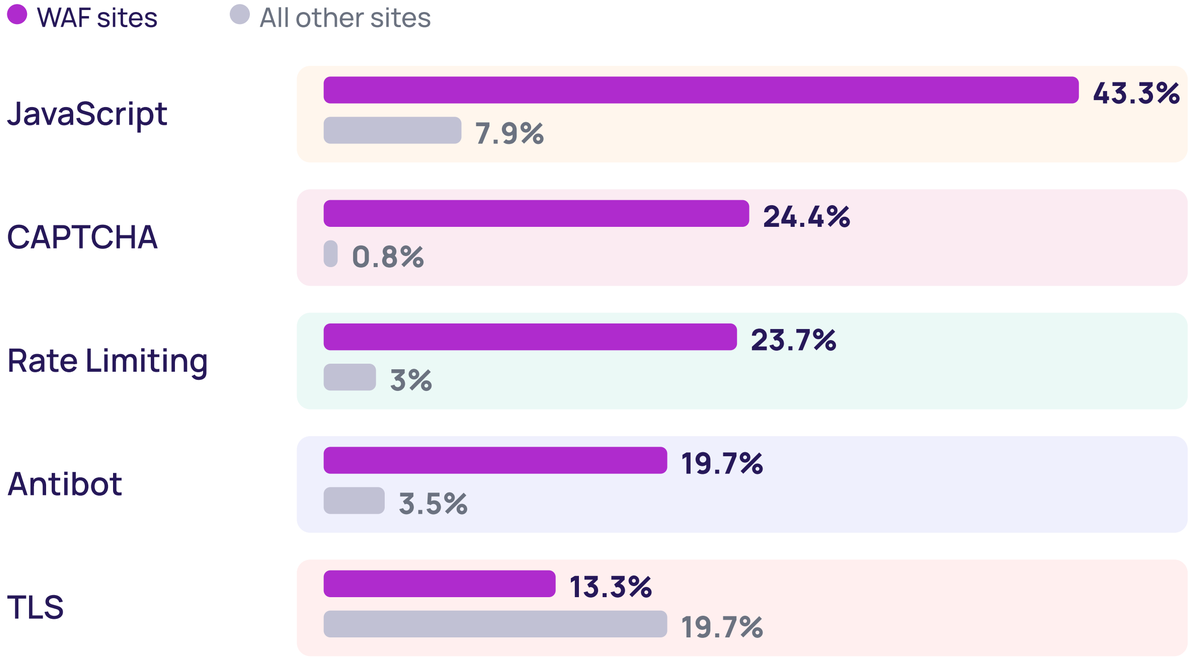
WAF is the most common access technology to be deployed on its own.

We see 18% of the URLs in our analysis relying on WAF alone (no JavaScript, no CAPTCHA, no antibot, no TLS, no rate limiting).

WAF detection operates at the network layer by analyzing HTTP responses and server headers, while antibot detection operates at the browser layer by using client-side scripts and assets to collect fingerprinting data and behavioral signals.

CO-OCCURRENCE · WAF SITES VS ALL OTHER SITES

WAF sites deploy every other barrier far more often



Antibots: The behavioral shield

Nearly one in five landing pages (18.5%) now deploy dedicated antibot solutions.

These systems represent the most sophisticated tier of access control.

Rate limiting only slows down request volume. But bot mitigation services use behavioral analysis to detect the pattern of access itself.

They monitor mouse movements, keyboard events, canvas rendering, and other browser attributes to identify rapid-fire purchase attempts and automated checkout flows.

18.5%

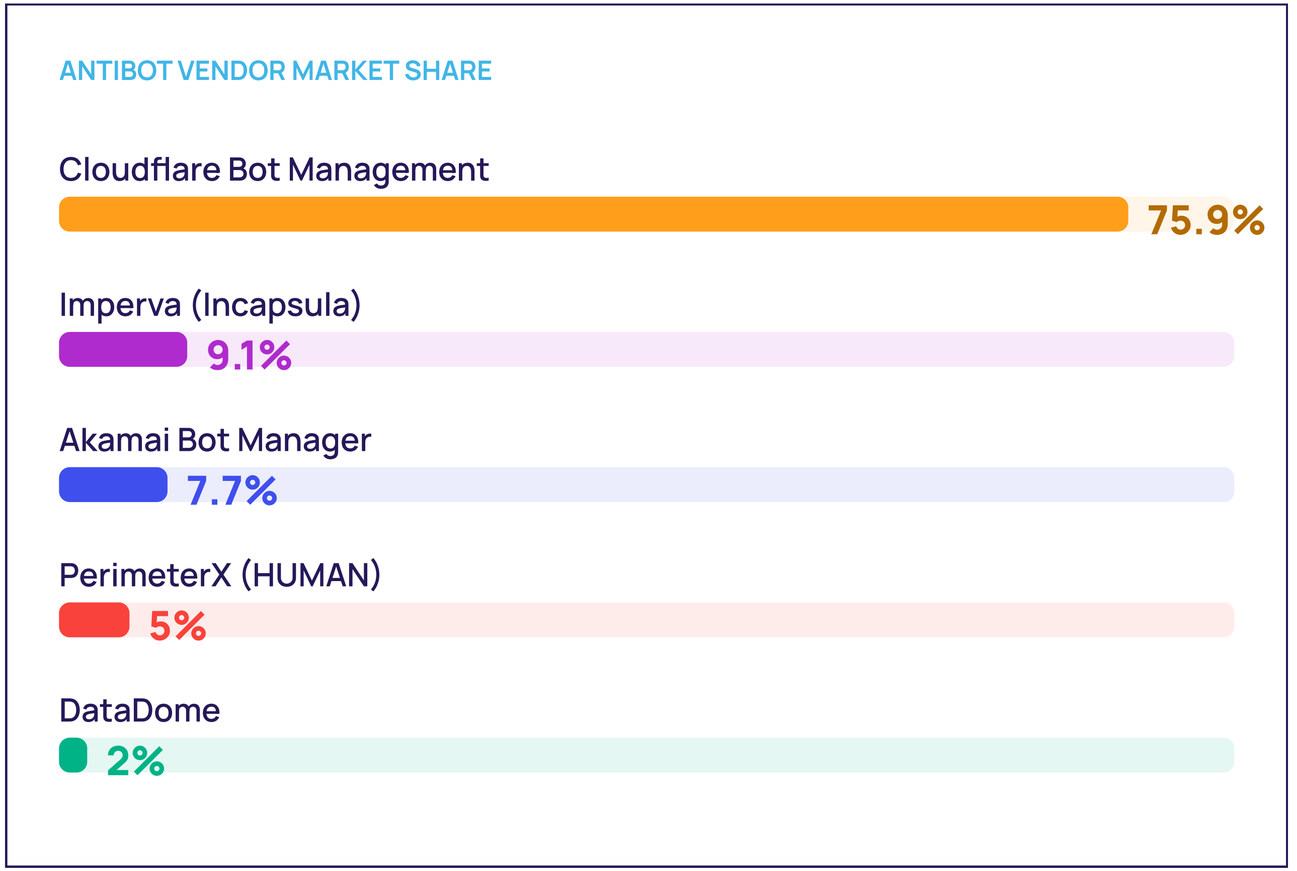
of sites run dedicated antibot protection

The antibot market is a one-horse race

The antibot market is effectively a one-horse race.

Three-quarters of all antibot deployments in this global dataset are handled by a single vendor.

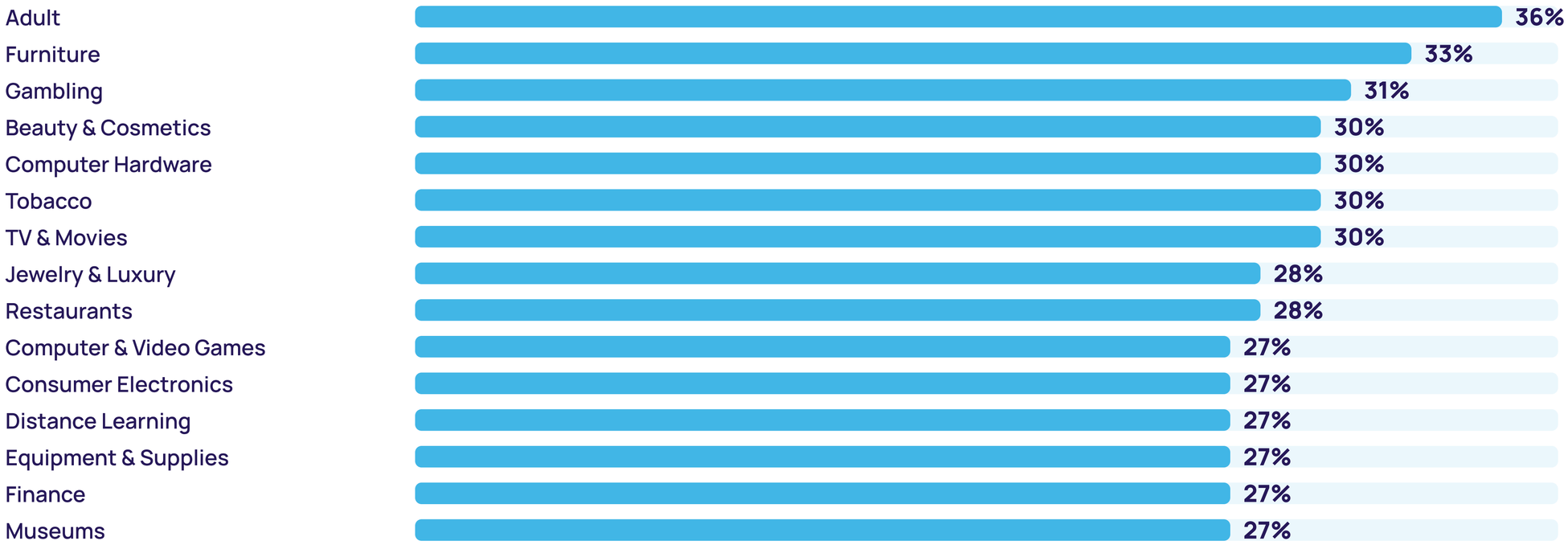
This concentration has a clear implication for data gatherers: navigating one system unlocks access to a substantial portion of the protected web.



Transactional sectors lead antibot adoption

ANTIBOT ADOPTION · BY INDUSTRY · TOP 15

Transactional and inventory-sensitive industries lead on bot management



zyte

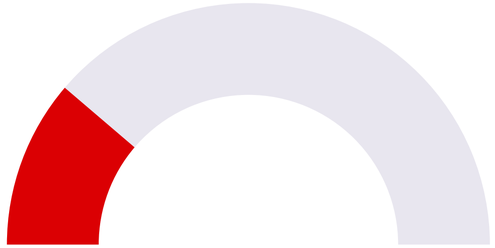
CAPTCHA: The visible friction

More than a fifth of landing pages deploy CAPTCHAs.

That makes these human puzzles the most visible defense mechanism on the web.

Unlike access restriction mechanisms such as WAFs and JavaScript, most CAPTCHAs are plain to see and disrupt the end-user's experience.

However, invisible CAPTCHAs are also common, such as reCAPTCHA v3 and Cloudflare Turnstile.



22.5%

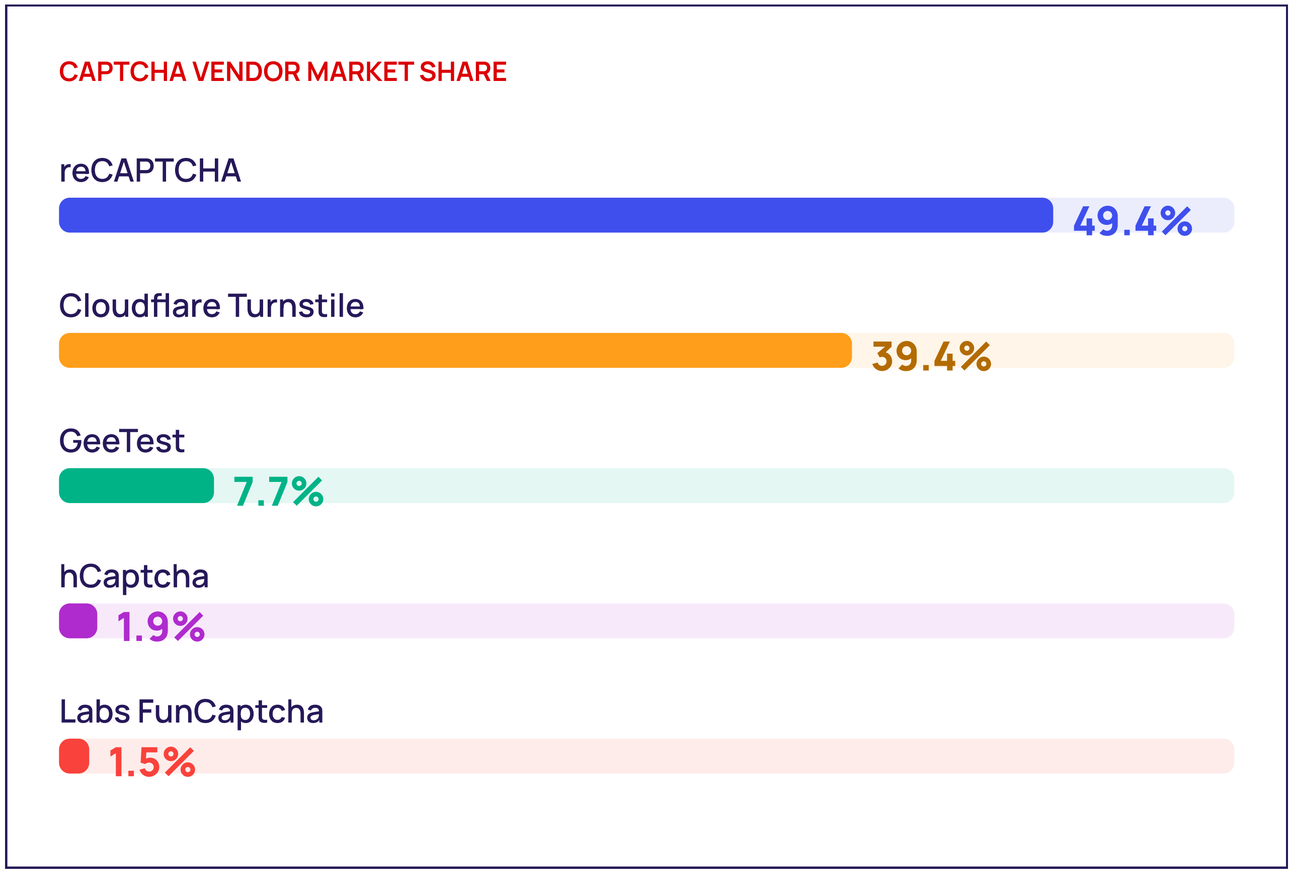
of sites deploy CAPTCHA infrastructure

Two vendors control nine in ten CAPTCHAs

Two solutions have captured the CAPTCHA market.

Users that can manage Turnstile and reCAPTCHA will gain access to 88.8% of all CAPTCHA-enabled sites.

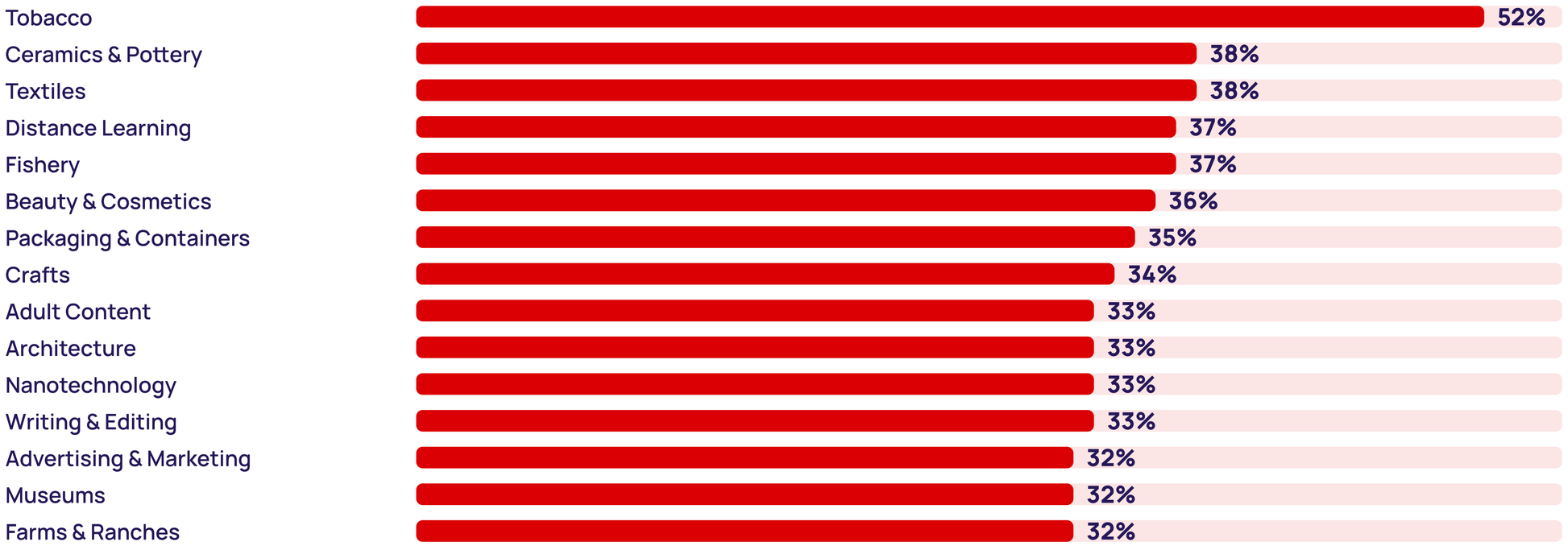
Adding GeeTest brings coverage to 96.5%.



Tobacco sites use CAPTCHA to regulate access

CAPTCHA ADOPTION · BY INDUSTRY · TOP 15

Tobacco at 52% leads a cluster of niche industries that make heavy use of CAPTCHA



State of Web Access · 11,100 landing pages

zyte

JavaScript: The architectural barrier

Almost half of the web now depends on JavaScript.

JavaScript is used by 40.3% of the 11,100 landing pages in our audit.

That is likely due, in part, simply to the growing popularity of JavaScript frameworks in recent years.

But the growing requirement that JavaScript is enabled to render web content also poses a challenge to data gatherers.

It means an expensive alternative to simple HTTP scraping.

On average, we found, JavaScript adds 140,923 bytes of content per page. This represents 75% of the final HTML payload.

40.6%

of landing pages are JavaScript-dependent

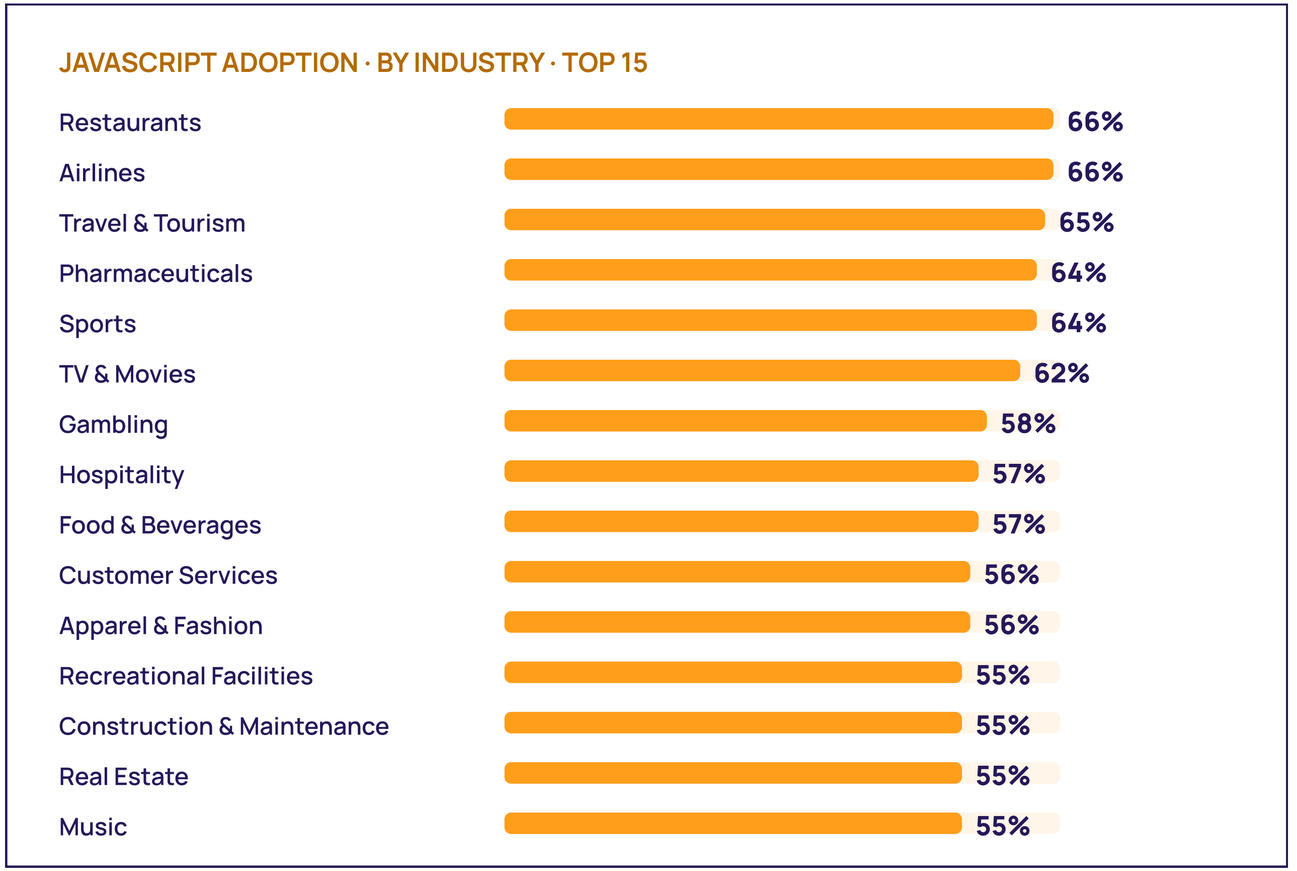
Two in five pages don’t exist without JavaScript

JavaScript forces data gatherers into difficult trade-offs.

Traditionally, scrapers had three options: use browser automation (which is slow, expensive, and fragile to detection), analyze network requests to internal APIs to retrieve data before rendering, or accept incomplete data.

As JavaScript frameworks continue to dominate web development, websites’ dependence on JavaScript will likely increase.

This makes browser automation an increasingly important capability.



Rate limiting: The blunt instrument

One in five landing pages (22%) deploy rate-limiting.

Among those that do, 94% block after just one request and 5% block after a burst of eight requests.

This indicates aggressive rate-limiting strategies.

From a scraper's point of view, this necessitates throttling, distributed requests, and rotating IPs.

22.1%

of sites throttle bursts of automated traffic

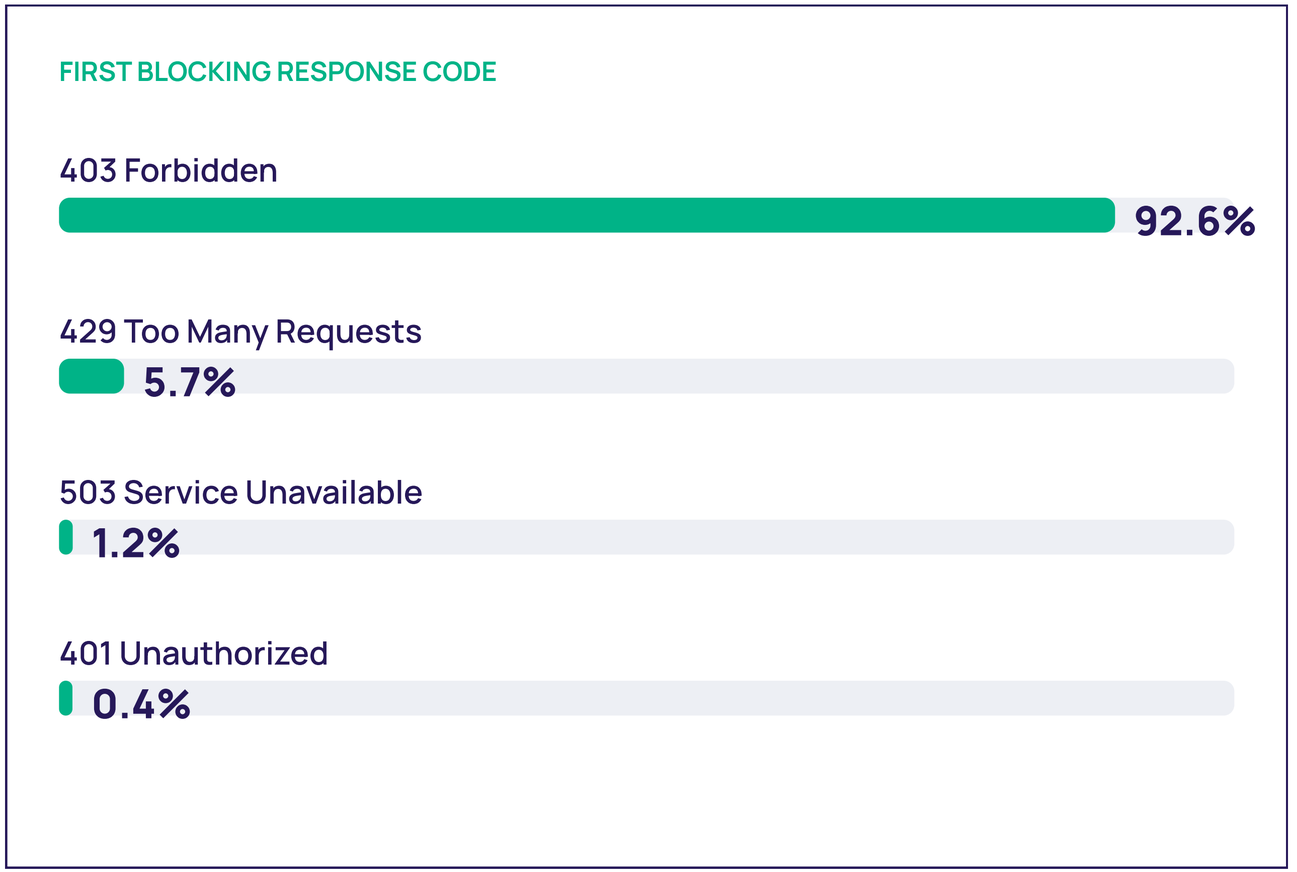
Rate limiting is actually blocking

Aggression levels are consistent across categories.

When sites implement rate-limiting, they implement it aggressively regardless of industry. Rate-limiting implementations are diverse and often crude.

Many sites re not following best practices for rate-limiting. Most return HTTP 403 (not 429, the RFC-standard status code for rate-limiting).

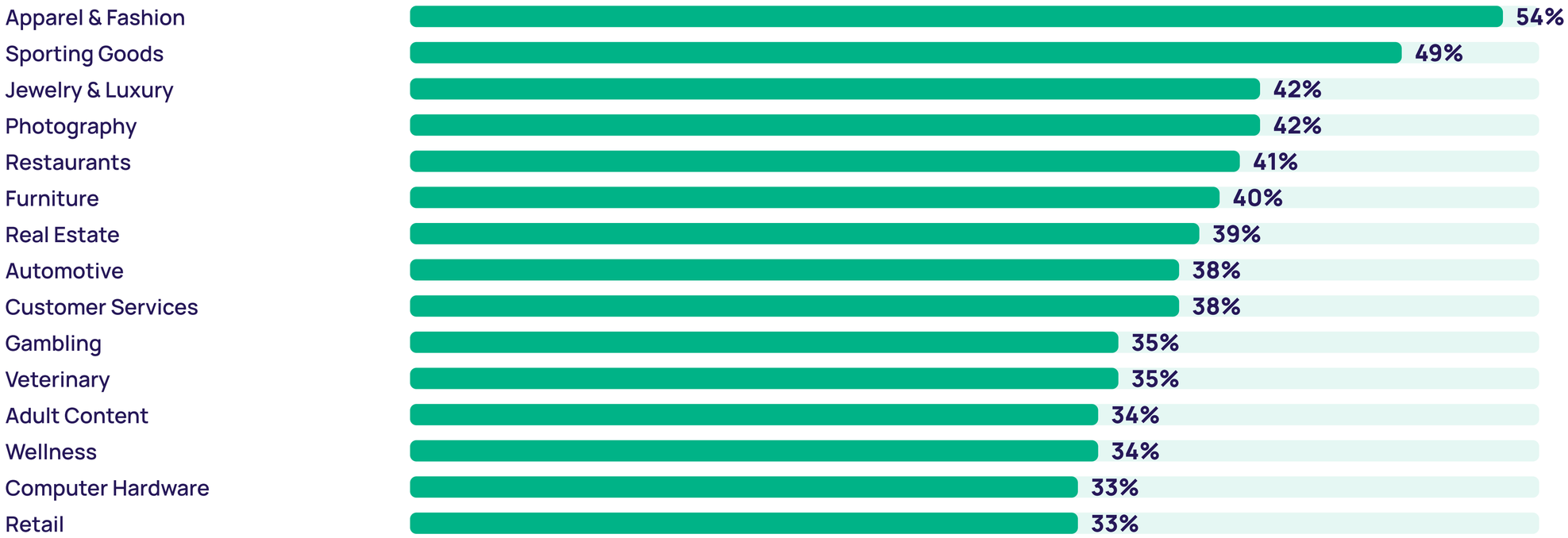
High-traffic sites are 66% more likely to use rate-limiting than low-traffic sites. Rate-limiting is a response to scale and attack volume.



Fashion and high-value inventory lead rate limiting

RATE LIMITING ADOPTION · BY INDUSTRY · TOP 15

Fashion leads at 54% — consumer-facing commerce with high-value inventory drives adoption



Transport Layer Security (TLS): The protocol check

At 13.8% adoption across the landing pages we looked at, TLS fingerprinting is meaningful but not ubiquitous.

However, for data gatherers working with high-value sites often associated with the need for secure connections, TLS fingerprinting is a notable barrier.

Unlike visible defenses like CAPTCHA, TLS fingerprinting operates at the protocol level, before any HTTP request is processed.

13.8%

of sites screen clients at the TLS layer

TLS fingerprinting marks the security-conscious web

TLS fingerprinting is a network-layer security mechanism that identifies clients based on unique patterns in their TLS handshake parameters.

Its primary purpose is security threat detection: identifying malware, detecting man-in-the-middle attacks, and blocking non-standard clients that may pose security risks.

TLS is the best indicator of security-conscious sites. Sites with TLS blocking are also more likely to employ other mechanisms.

TLS SITES ARE FAR MORE LIKELY TO RUN EVERY OTHER BARRIER

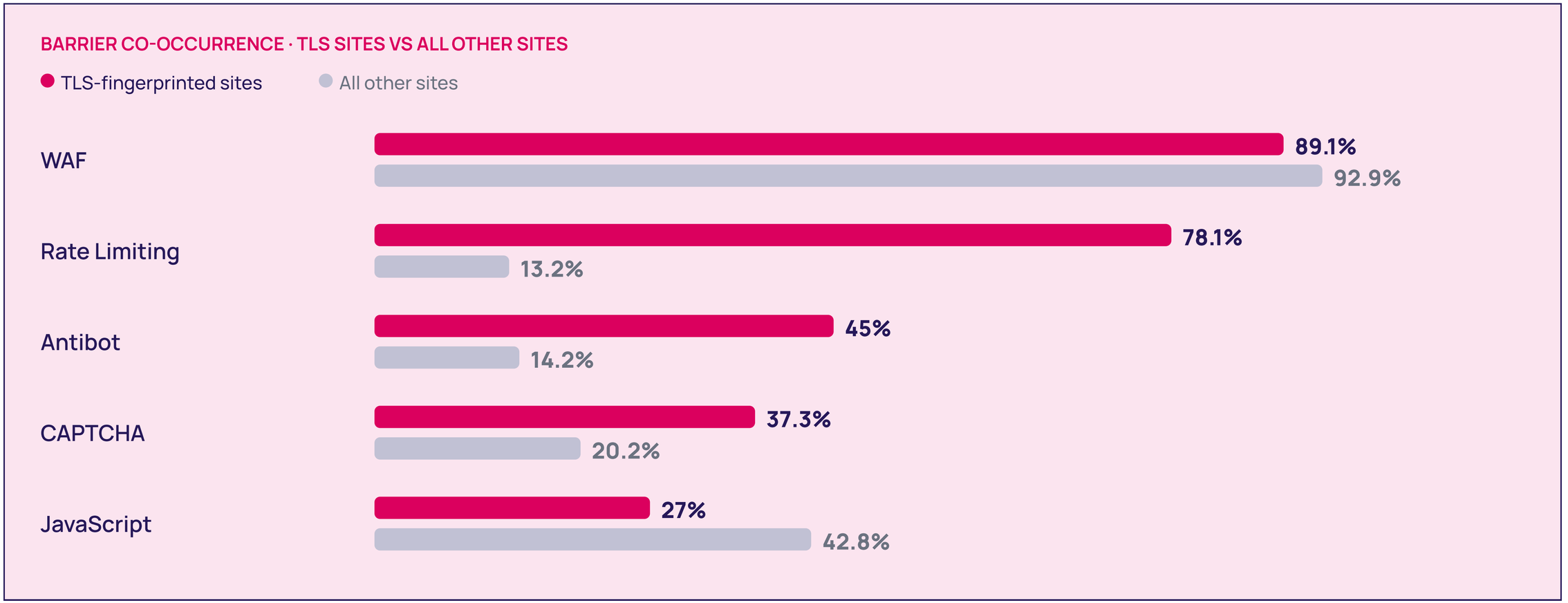
5.9×
more likely to run rate limiting

3.2×
more likely to run antibot

1.8×
more likely to run CAPTCHA

1.2×
more likely to run WAF

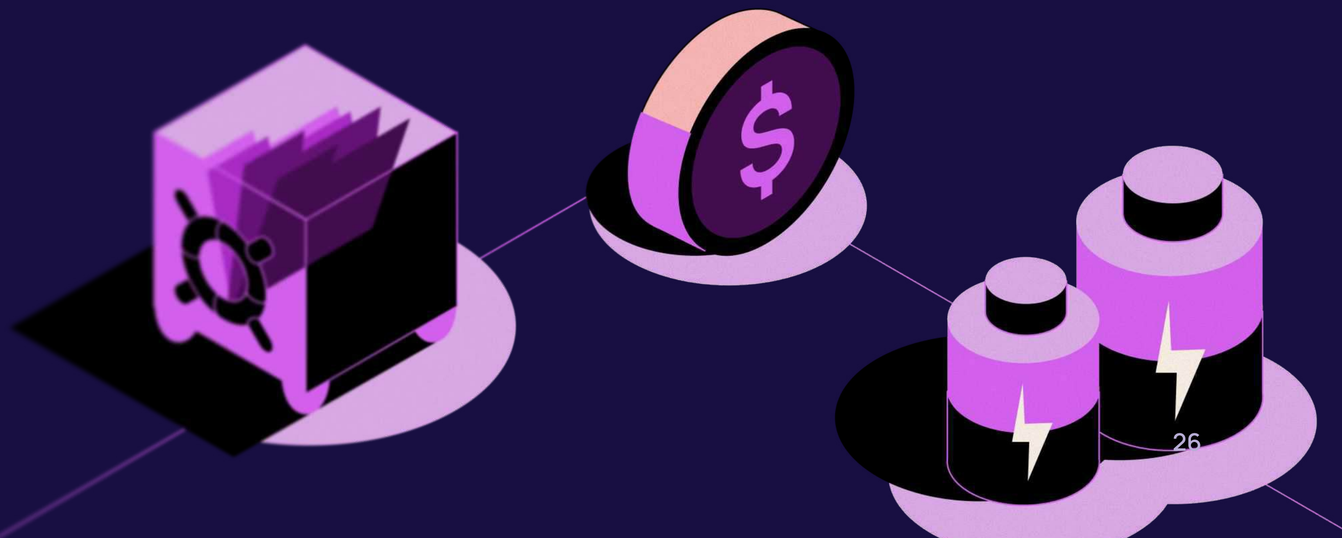
TLS-fingerprinted sites add other barriers, too



zyte

The cost of complexity

The impact of these access barriers is cumulative. Each additional barrier increases the cost of a request, as well as the number of moving parts you need to handle.



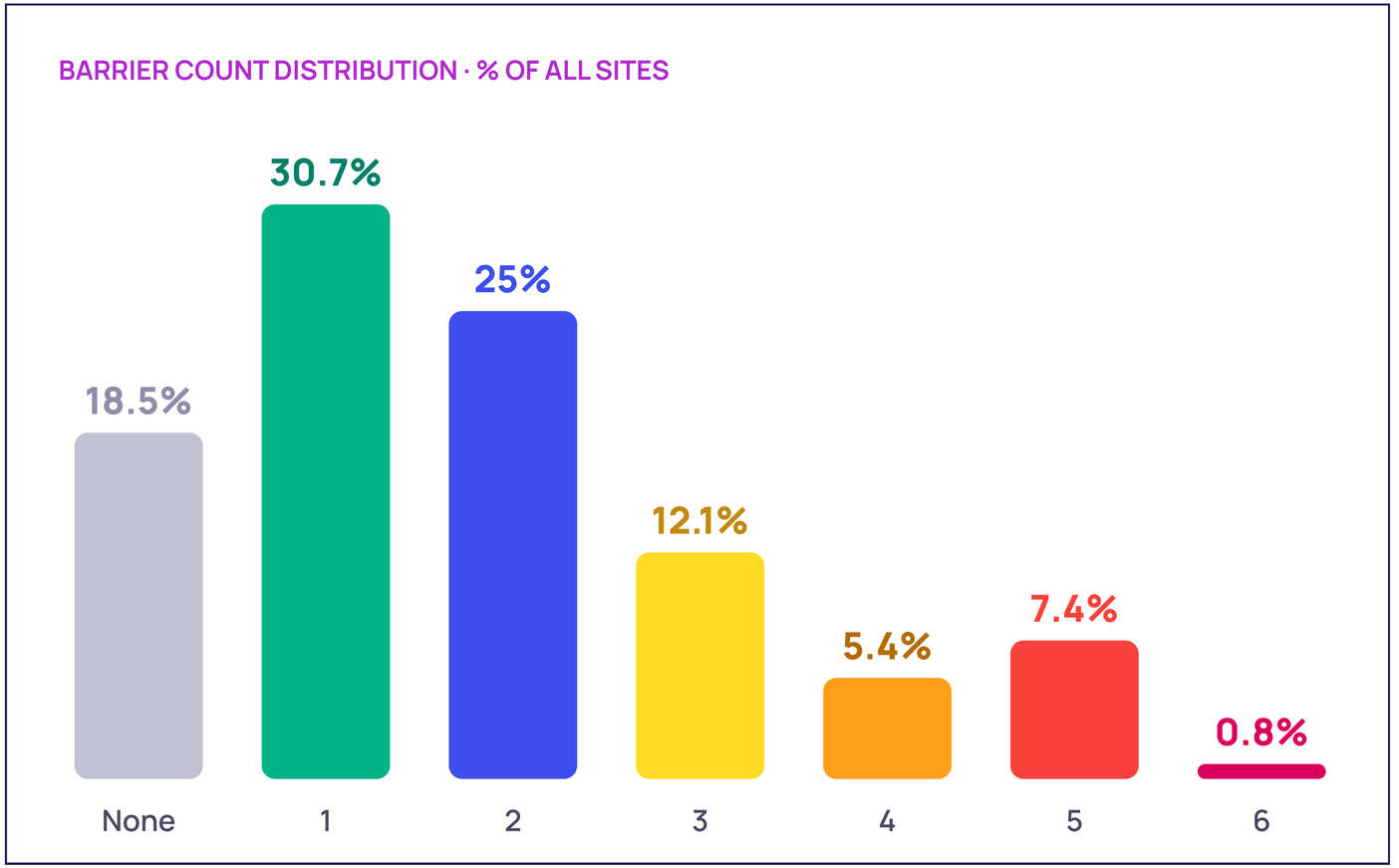
Half the web stacks its defenses

Most of the defended web is not hardened - it is layered.

Sites don't pick one control; they stack mechanisms across the network, browser, and protocol layers until the friction is high enough to deter most automated traffic.

Nearly one in five landing pages (18.5%) carry no detectable barrier. At the other extreme, 0.8% of sites run all six barriers.

But the story is in the middle: more than half of all sites stack two or more mechanisms.



Only the firewall stands alone

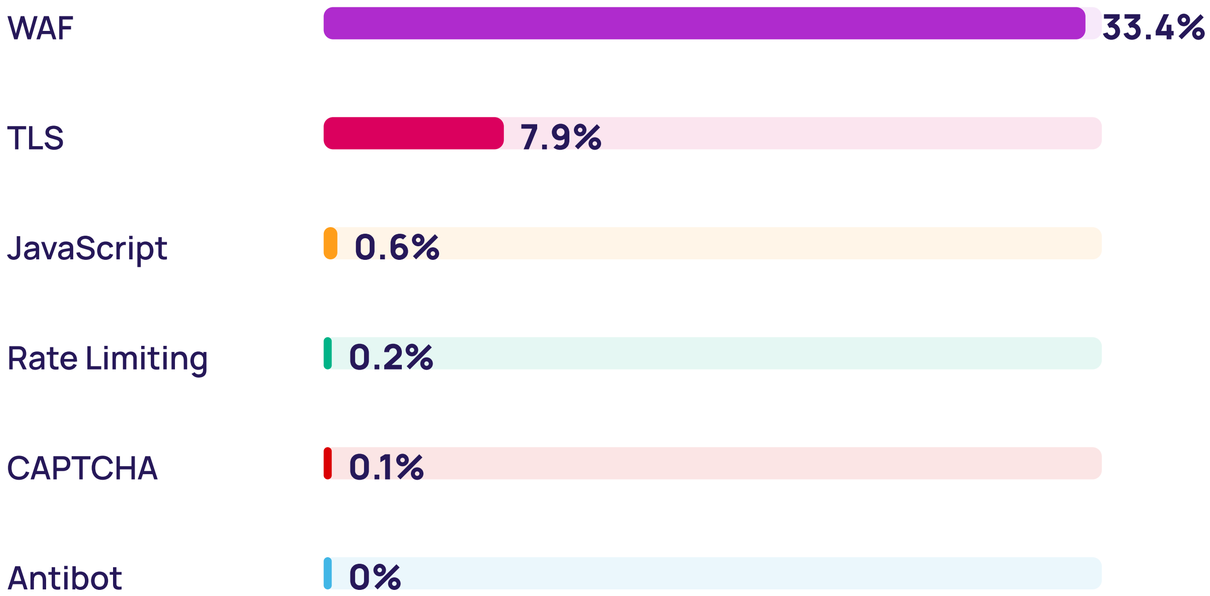
WAF is the only barrier that commonly operates alone. A third of WAF-protected sites deploy nothing else - they treat the firewall as sufficient. Every other barrier tells the opposite story.

TLS fingerprinting alone appears on fewer than 8% of TLS sites. JavaScript, Rate Limiting, CAPTCHA, and Antibot practically never operate without WAF.

Two stacks dominate the defended web. WAF-only (31%) and WAF + JavaScript (20%) together account for over half of all active-defense postures. The full-stack combination appears on just 3.3% of sites, concentrated in the highest-value commercial categories.

BARRIER STANDALONE USAGE · BY BARRIER TYPE

WAF is the only barrier that regularly operates alone — all others need a partner



Strategic barrier response

The era of one-off scrapers is over. Durable collection now looks like systems engineering. Taken together, this combination of layers means data gathering pipelines need to be built as systems:

01

Plan for multi-path retrieval

HTML fetch vs. rendered vs. underlying APIs.

02

Instrument failures by layer

TLS to WAF to JS rendering to CAPTCHA.

03

Treat sessions as an asset

Persistence, reputation, and state matter.

Site complexity can translate to cost

Barriers translate directly into infrastructure cost.

WAF requires proxy capacity, JavaScript demands a headless browser, and CAPTCHA needs a solving service. Each layer compounds the one before it. These costs compound when a site has multiple barriers in place.

Estimating access costs can be difficult.

A team running 50 active crawlers with a 10% weekly ban rate burns through roughly 30 developer-hours per week just fixing broken scrapers. Web scraping APIs cut through this.

ZYTE TIER MAPPING · ALL SITES

88%

of landing pages are Simple or Easy to access with Zyte API - the two lowest tiers. Just 12% reach Moderate or above.

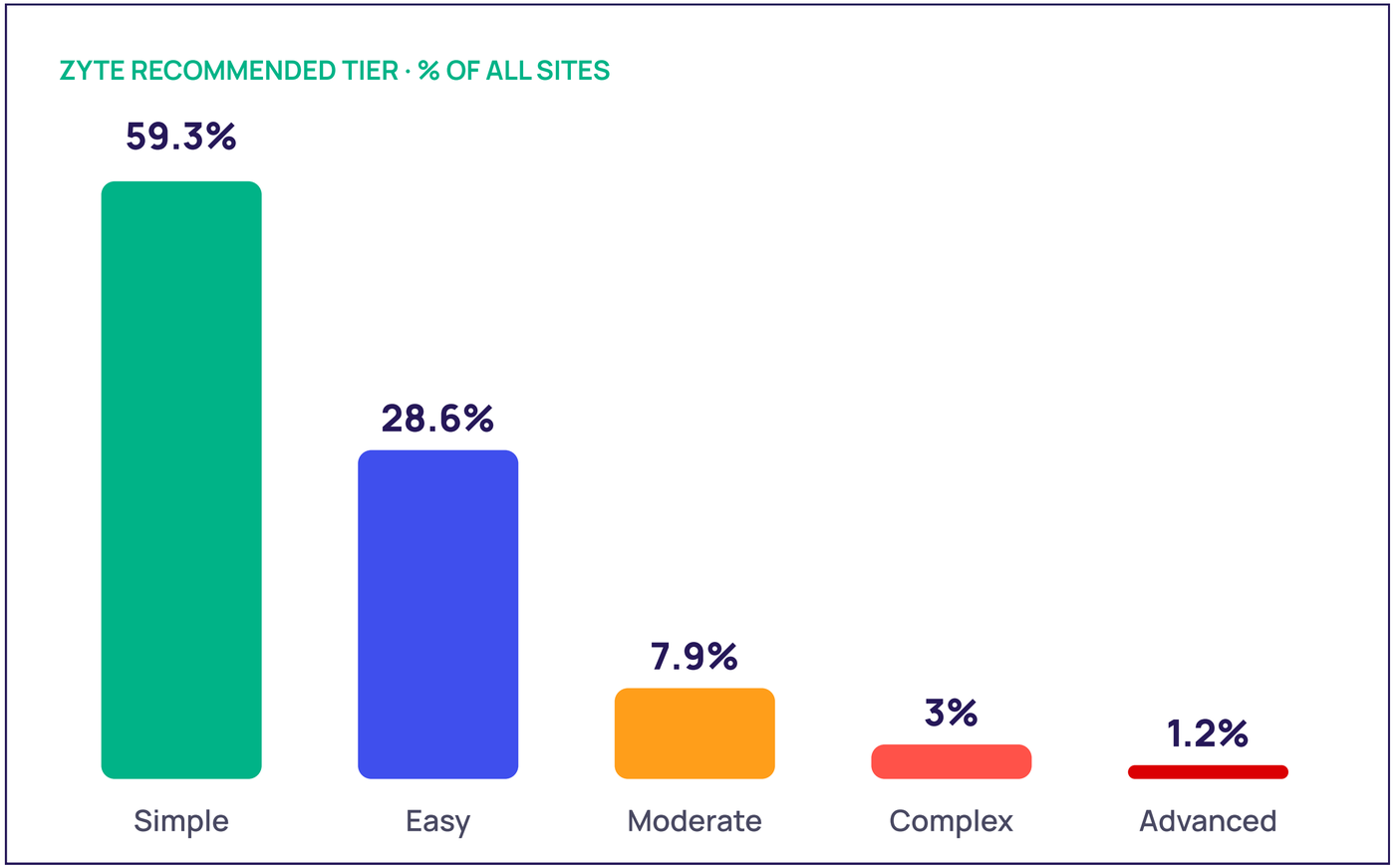
Most of the web is still cheap to access

The good news - despite growing barrier sophistication, the vast majority of the web remains cost-efficient to scrape with Zyte API.

To quantify this, we mapped every site to Zyte's five-tier access model.

The web is not uniformly hard. Despite barriers, 88% of landing pages still sit at Simple or Easy - the two lowest tiers. Just 12% reach Moderate or above.

The web has a long tail where complexity concentrates sharply toward value.



Infrastructure requirements influence cost

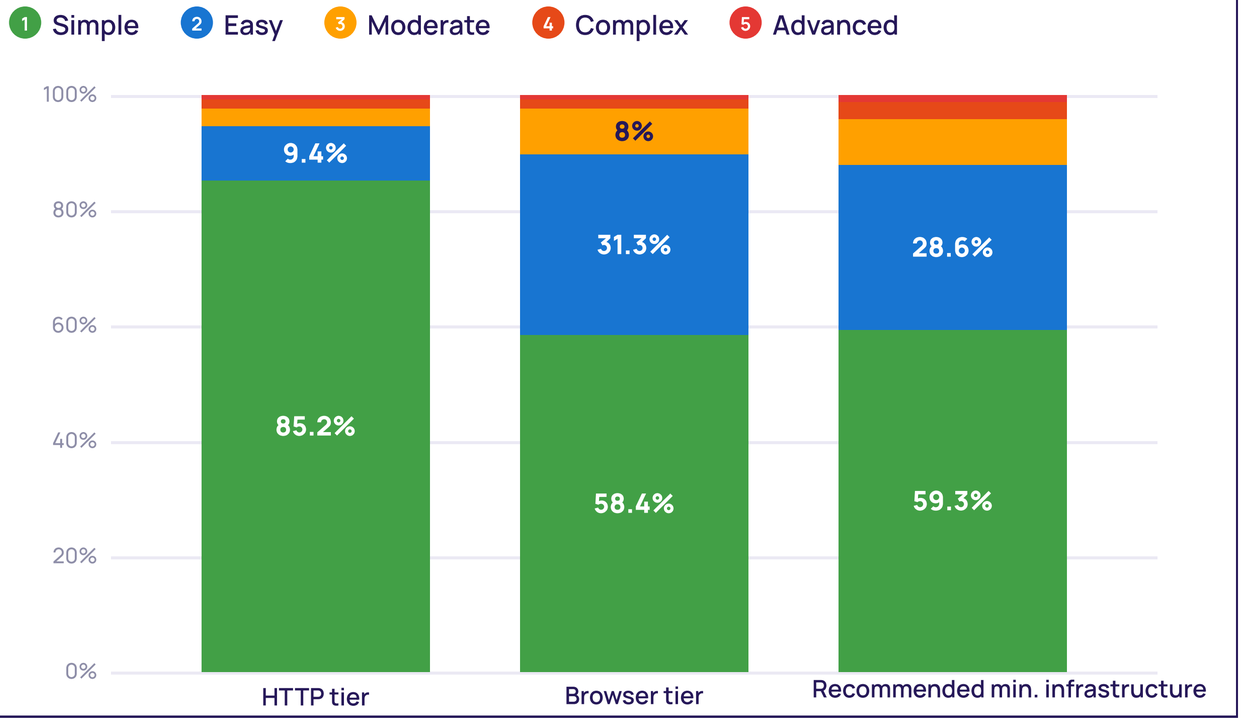
JavaScript drives complexity and cost.

When pages require JavaScript rendering, switching from HTTP-only to browser-based access demands more infrastructure fire-power.

Even so, this increase is still contained within Zyte API's Simple and Easy tiers.

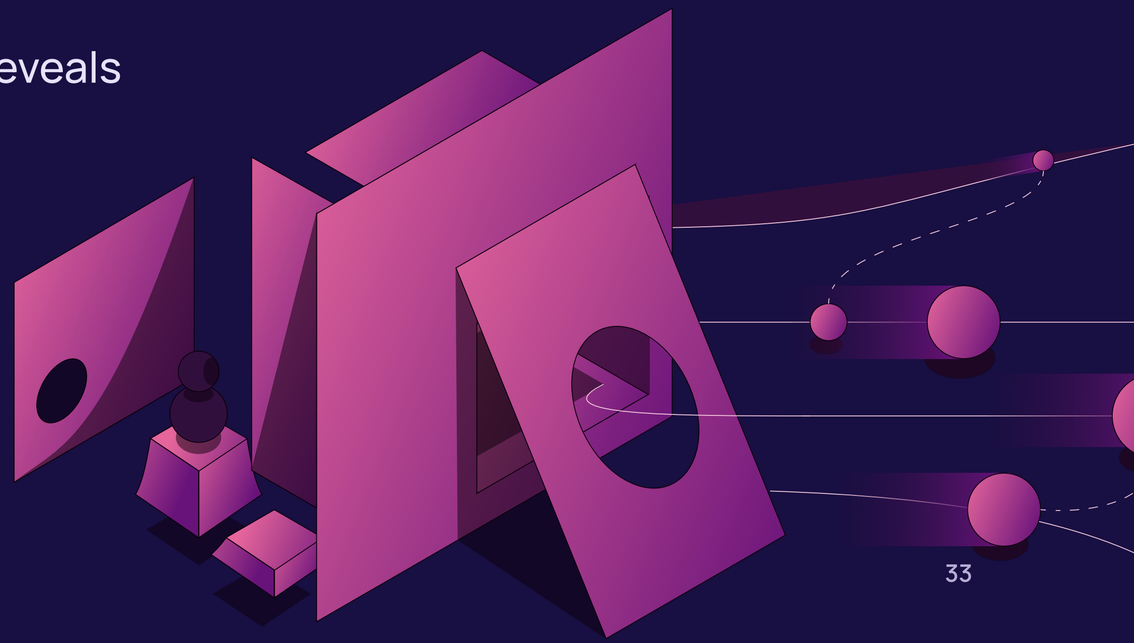
ZYTE TIER BY ACCESS METHOD

Browser access escalates the cost tier



robots.txt: The advisory layer

A humble text file hides on websites' servers - and increasingly reveals the policies and penalties facing automated visitors.



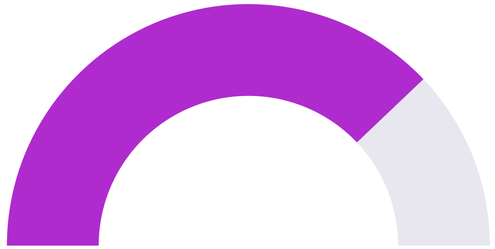
robots.txt: Three in four websites publish a file

Three-quarters of websites employ robots.txt, the file site owners use to convey preferred access policies to visitors.

robots.txt operates on an honor system. It communicates intent rather than executing technical enforcement.

Legitimate and ethical scraping setups are programmed to fetch and read the robots.txt file before hitting any other pages.

Recent guidelines ask general-purpose AI providers to respect the system while gathering training data.



75.8%

of sites have a valid robots.txt file

Content industries lead on granular bot management

The more content-focused a site's business model is, the more sophisticated its robots.txt will be.

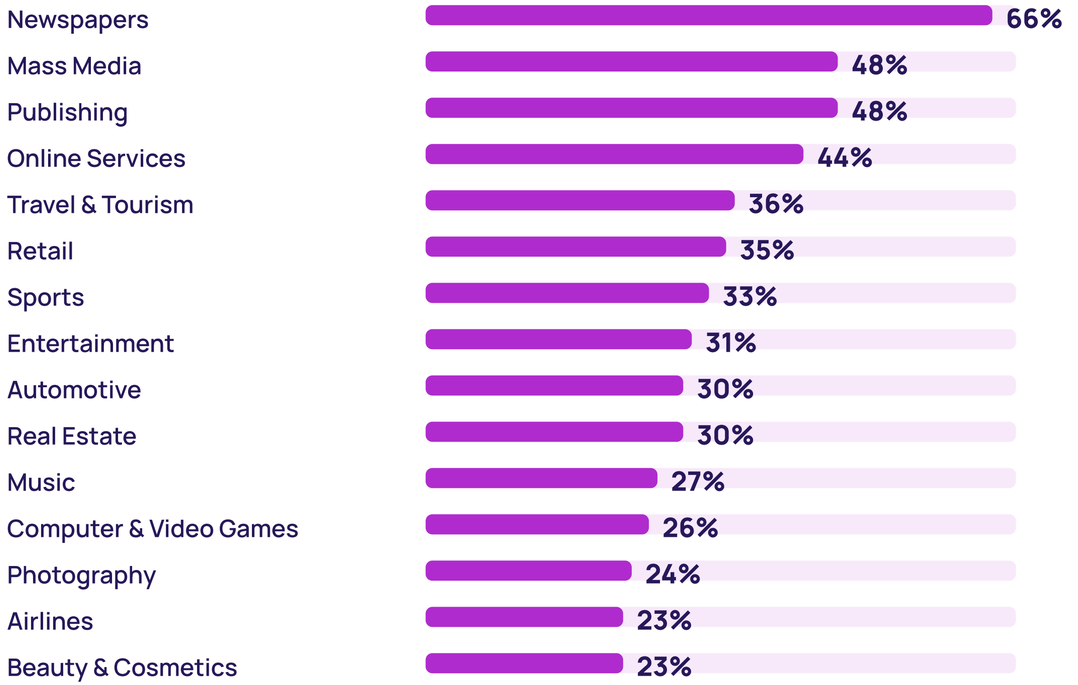
Newspapers, media, and publishing sites have the most sophisticated robots.txt files. They feature a curated list of allowed and disallowed agents.

Two third of newspapers name five or more more crawlers in their robots.txt files.

Retail and online services follow behind due to high perceived value of product data.

ROBOTS.TXT SOPHISTICATION · BY INDUSTRY · TOP 15

Content industries name the most crawlers



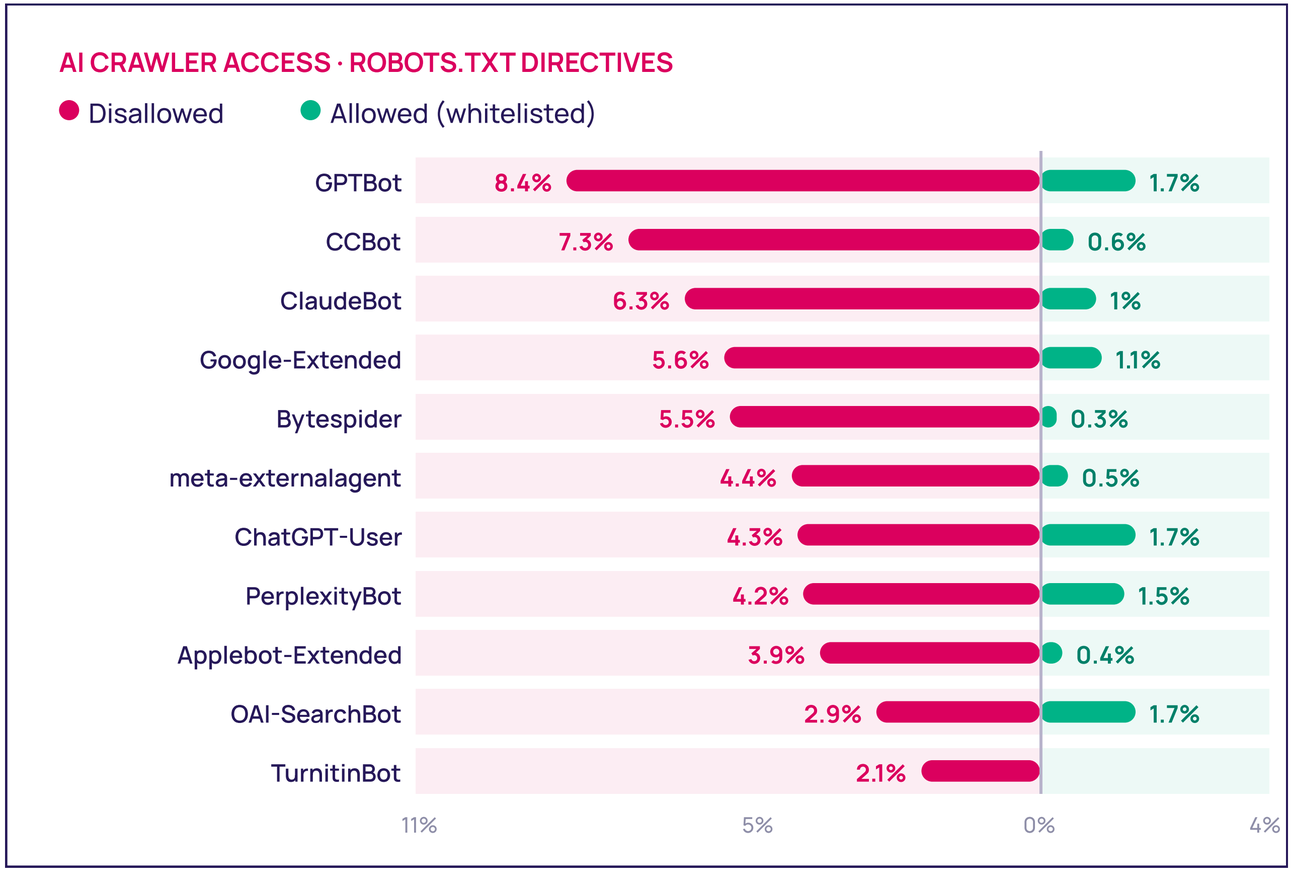
13% of robots.txt files target AI crawlers

Sites are actively trying to prevent their data from being ingested by LLMs.

Around 13% of the URLs in the dataset address AI crawlers in their robots.txt.

News and media sites have aggressive blacklisting, led by newspapers (69%), publishing (51%), and mass media (44%), specifically targeting unwanted crawlers.

As AI companies publish User-Agents, we expect the sophistication of robots.txt files to increase as sites scramble to update their whitelists and blacklists.

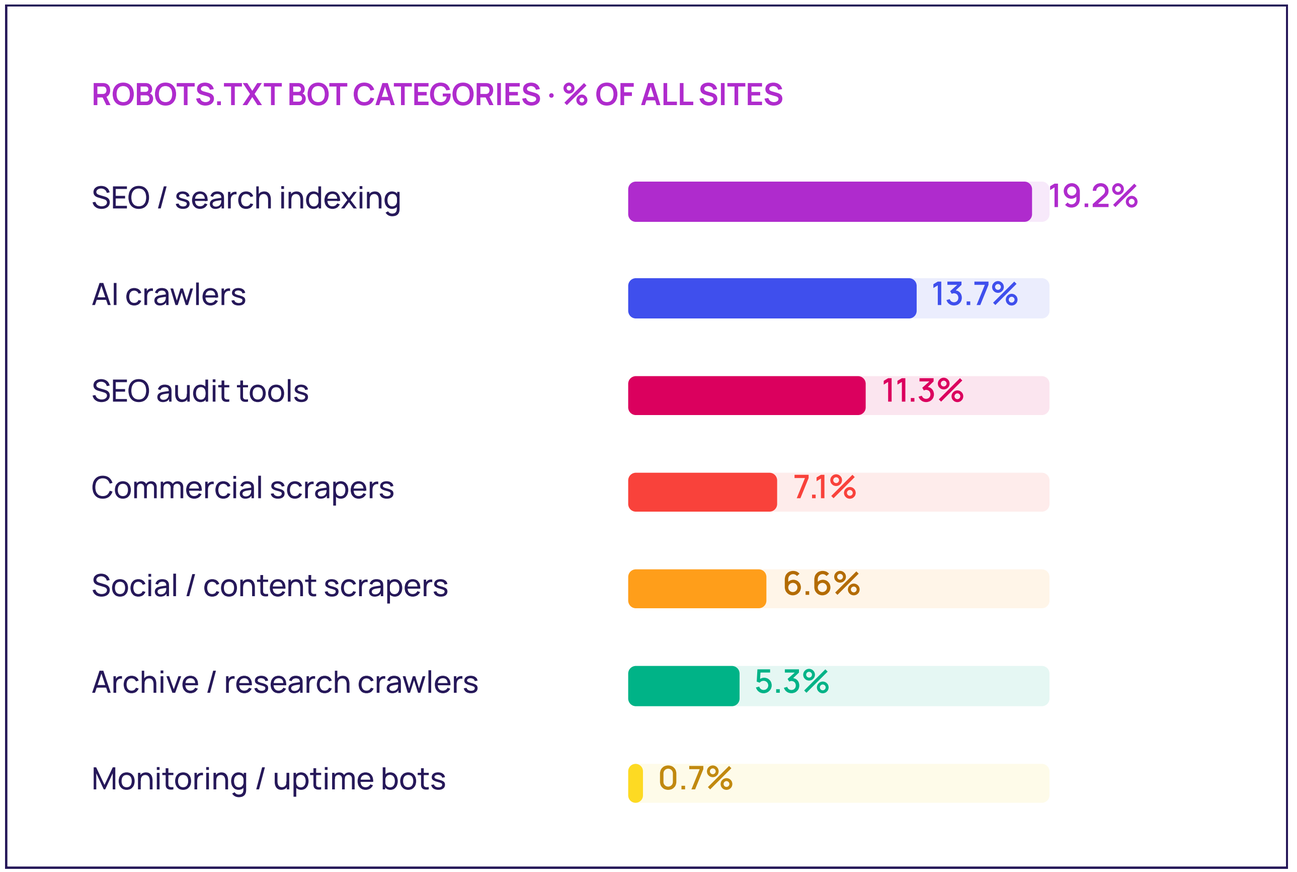


SEO bots are in robots.txt's sights, AI second

Search engines are more blocked than AI bots.

However, search site crawlers are more likely than LLMs to appear in robots.txt files.

This reflects the relative value that publishers place on these respective platforms as content re-use partners.



The state of robots.txt

The exclusion protocol text file keeps on going - but is not the only gatekeeper on the web.

Voluntary compulsion

robots.txt has historically been considered an informal request rather than a technical barrier to scraping.

Only a tiny fraction of sites (0.5%) gets by exclusively with a robots.txt file and no other technical defenses.

Robots don't dance alone

robots.txt is bolstered by other mechanisms.

The vast majority of robots.txt-publishing sites use a "belt and suspenders" approach, using overlapping strategies to further minimize risk.

Some 48.3% of all sites have robots.txt plus two or more technical defense.

Access control by industry

Web access barriers are not raised to the same height in all corners of the web. Their application follows business models and industry economics.

Where data is richest, defenses run deepest

Fragmentation by industry means there is no universal key to the web. Effective site access requires understanding different industries' unique postures and adapting approaches accordingly.

Hardest to access

Industries with contested, high-value data stack secondary defense on top of WAFs, while those with open, indexable content stop at firewalls.

For instance, real estate and automotive deploy meaningful barriers on top of baseline WAF to defend listing and pricing data against systematic extraction.

Easiest to access

At the other end, public sector, agriculture, and media and publishing rely almost entirely on CDN-bundled WAF - their business models depend on discoverability, not data exclusivity.

Barrier adoption across 110 industries

Different access requirements show up in 110 industries analyzed by State of Web Access — our heatmap rolls them into 20 groups.

	WAF	Antibot	CAPTCHA	JavaScript	Rate Limit	TLS	robots.txt	Blocks AI	Tier
Media & Publishing	97%	14%	24%	36%	16%	11%	86%	30%	2
Retail & E-commerce	93%	23%	27%	44%	31%	19%	77%	12%	2
Technology	93%	14%	21%	39%	18%	12%	77%	12%	1
Financial Services	90%	21%	14%	37%	16%	12%	74%	5%	2
Healthcare	94%	19%	25%	44%	26%	14%	80%	11%	1
Manufacturing & Logistics	90%	16%	22%	38%	18%	14%	71%	6%	1
Travel & Hospitality	85%	23%	16%	50%	29%	18%	68%	9%	2
Entertainment	96%	27%	24%	49%	30%	14%	75%	16%	1
Education	92%	17%	19%	38%	17%	10%	76%	13%	1
Design & Creative Arts	97%	18%	29%	38%	20%	12%	80%	11%	1
Public Sector	89%	11%	16%	35%	16%	12%	73%	7%	1
Legal & Professional Svcs	91%	14%	22%	39%	18%	12%	71%	5%	1
Non-profit & Civil Society	95%	22%	27%	40%	21%	14%	82%	10%	1
Agriculture	96%	12%	33%	33%	21%	12%	75%	7%	1
Energy & Utilities	86%	16%	15%	38%	16%	11%	63%	4%	1
Construction & Environment	92%	17%	19%	47%	23%	11%	76%	9%	2
Recruitment & HR	92%	19%	20%	35%	18%	15%	75%	6%	1
Automotive	92%	16%	9%	44%	38%	16%	80%	22%	2
Real Estate	92%	24%	19%	48%	39%	15%	75%	17%	2
Telecom	95%	19%	18%	45%	16%	11%	79%	6%	2

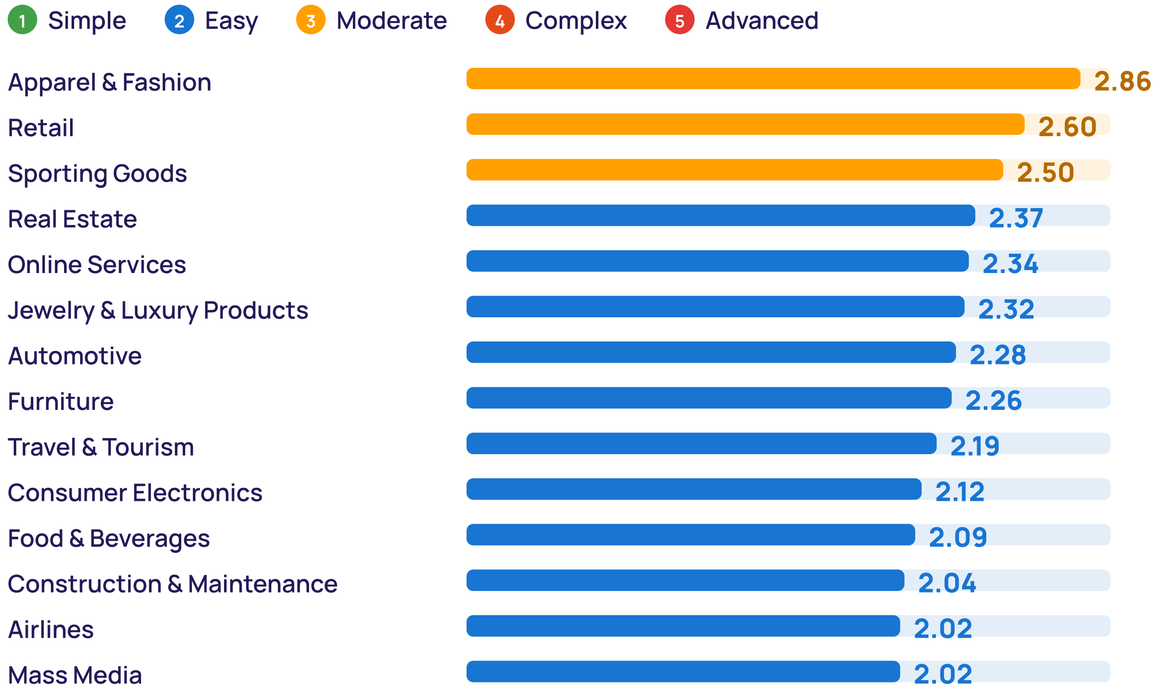
Fashion guards its shelves hardest

Ecommerce sites demand stronger infrastructure.

Across all Apparel & Fashion sites, the recommended access infrastructure requirement gives the industry a Zyte tier of 2.86 — deep into Moderate complexity.

Meanwhile, the entire bottom 15 of industries sits between 1.02 and 1.19, all solidly Simple.

High-value sectors demand the heaviest infrastructure



The contested battleground of the modern web

Digging deeper into specific industries reveals more granular approaches.

Listings-related categories

Measures employed by listings-related categories make them among the most complex sites to access. The retail category falls into the cluster of the most protected categories. Mass media and online services follow closely. They need to balance access from both legitimate visitors and undesirable user agents which might pose economic threats.

E-commerce and consumer-facing sectors

The e-commerce and consumer-facing sectors represent the contested battleground of the modern web. Unlike other sectors where information is meant to be fully public, e-commerce sites face a fundamental tension: they must be discoverable by search engines and accessible to customers while remaining opaque to competitors.

Real estate erects the tallest walls

Experienced web scrapers know real estate consistently ranks in the top tier of difficulty. Real estate sites face pressure to block automated access because their business model depends on controlling information asymmetry.

News sites prefer paywalls to protection

Landing pages in the media and publishing space rank among the least defended industries on the web. Publishers seem to defend their content through business model innovation, not web infrastructure - rather than technical barriers like WAF or JavaScript, media sites increasingly rely on registration walls that build first-party data. They use paywalls that monetize access directly, and they use membership models that create sustainable revenue independent of traffic.

The future of access

The web isn’t necessarily “closing” - it’s putting a price on access.

Where data is richest, defenses run deepest

> Over the last decade, the default posture of the internet has shifted from open-by-default to increasingly filtered.

Only a minority of popular landing pages are frictionless, and roughly half now layer multiple defenses.

There is no single switch that has been flipped - it's more like an interweaving system of gates that compound: infrastructure-layer controls like WAFs, architectural friction through JavaScript-heavy rendering, interactive checks like CAPTCHA, throughput throttles like rate limiting, and pre-HTTP screening through TLS fingerprinting.

Access is changing

That tech stack matters because web data is no longer a “free input”. Rather, it’s a strategic resource - and it is being treated that way by every party involved.

Scraping engineers

The era of one-off scrapers is over. Durable collection now looks like systems engineering, demanding:

- Multi-path retrieval (HTTP vs browser vs APIs).
- Instrumentation by layer (TLS, WAF, JavaScript, and CAPTCHA).
- Ongoing adaptation as vendors and sites update their playbooks.

The goal now cannot be to “get past a block,” it must be laying pipelines that stay reliable as access control evolves. It becomes increasingly complex for individual teams to handle.

Web data buyers

Access friction is a cost curve. Two suppliers can quote the same dataset and deliver radically different reliability, latency, and total cost of ownership depending on their ability to handle stacked defenses.

Access difficulty is a first-class requirement. That means monitoring coverage, change rates, failure modes, and the operational plan for revalidation over time - not just price per record.

Website owners

Access control is now product strategy. Barriers rarely operate in isolation; most deployments are layered, and the market is consolidating around a handful of major vendors. That concentration makes control mechanisms easier to buy - and easier for sophisticated actors to learn.

The strategic question is not as blunt as: “Can we block scrapers?” Increasingly, website owners are asking more specific questions: “Which automated access do we want to allow, under what terms, and with what safeguards?”

Data and tech leaders

The web is fragmenting into tiers of access. Some of it remains a commons; some of it is guarded by default.

Incentives, not geography, are decisive: the strongest stacks cluster where information is scarce, transactional, regulated, or directly tied to revenue. Any organization building AI, analytics, or market intelligence on web data needs to plan for this reality: access is an operational determinant, and resilience is a competitive advantage.

What we think happens next

We expect three forces to shape the next phase of the web.

01

Control will keep professionalizing

WAF and bot mitigation will continue to move closer to the edge - faster to deploy, cheaper to maintain, and more integrated across products.

02

Automation pressure will keep rising

It's not just coming from scrapers, but also from agents, monitoring bots, and AI systems generally. As automated demand grows, more sites will treat traffic quality more as a resource-management problem, less a security problem.

03

Access will become more explicitly negotiated

We expect more sites to formalize machine access via APIs, partnerships, allowlists, and policy controls - especially as AI-related crawling becomes a board-level concern.

The practical takeaway is simple: the future of web data belongs to the teams that treat access as an engineering discipline and a governance decision - not a workaround. The web is still there. But the rules of engagement are changing, and every stakeholder now has a choice: build brittle shortcuts, or build for resilience.

How does Zyte help?

Zyte’s solutions

Web data access is becoming harder. Zyte gives teams one automated way to access the modern web.

1 Web Application Firewalls

No more tuning every site by hand. Handle common network-layer frictions with automated proxy routing, request configuration, cookies, headers, and retries.

2 Antibot

Zyte’s library of 320,000 access strategies includes browser execution, session handling, routing, retries, and fingerprint-aware configuration.

3 CAPTCHA

Manage CAPTCHA-related access challenges automatically, as part of the request workflow, combining browser rendering, sessions, and routing.

4 JavaScript

Managed browser rendering and JavaScript execution. Get rendered page content without running a headless fleet.

5 Rate limiting

Auto-managed request delivery with adaptive routing, proxy infrastructure, retries, and session-aware configuration, reduces the need for custom rate-limit workarounds.

6 TLS fingerprinting

Avoid low-level connection engineering by using managed, browser-compatible request infrastructure designed for reliable public web data collection.

Always the lowest cost

Zyte automatically selects the minimum effective access configuration for each site, helping teams control proxy, browser, maintenance, and engineering costs.

zyte

Access the web's data

Zyte turns websites into data with industry-leading technology and services.
Over 15 years, Zyte has helped more than 4,500 businesses gather data from five trillion pages.

Products

Zyte API

Everything you need in one, AI-enabled toolkit. Automate discovery, unblocking and extraction. Hassle-free scraping, full control.

Get started

Agentic Web Data

Bring web data extraction into your IDE or coding assistant. Access, extract, and structure data as part of your normal workflow.

Explore our plugins

Scrapy Cloud

Scalable cloud hosting for your spiders. Provides a simple way to run your crawls and browse results.

Deploy now

Services

Zyte Data

The feed you need for your dream data. When scale and speed matter, let Zyte's in-house expert engineers build, deliver and maintain your custom web data pipeline.

Contact us

\+ See how Zyte helps access and unlock web data at real scale, claim a huge $200 Zyte API credit now.

Claim now

Methodology

This report is based on a large-scale scan of popular landing pages, designed to measure how often common access barriers appear and how they stack.

Scope, dataset and what we measured.

SCOPE AND DATASET

Landing page: For each URL, we scanned the landing page provided by SEMRush's top websites lists, producing two datasets:

- **Global dataset:** Because some sites appear in more than one industry category, the dataset contains 11,100 records in total — the figure cited in the global charts, since each industry analysis counts each record once. This gives a single, stable snapshot of the most popular sites on the commercial web and is the source for the global-level figures throughout this report.

- **Regional dataset:** 22,137 unique landing page URLs from SEMRush's regional trending websites list (snapshot: October 2025), sampling top URLs across countries and categories. A single site may be popular in multiple countries - we audited each of those sites once. Combining and deduplicating from the global and regional lists, we audited a total 24,898 landing pages.

Geography mapping: SEMRush country codes were then grouped into six regions: North America, Europe, Asia, South America, Oceania, and Africa.

Industry mapping: SEMRush category codes were mapped into 17 industries used for analysis in this report.

WHAT WE MEASURED

For each landing page, we scanned for six access barriers:

- Antibot / bot mitigation services
- CAPTCHA
- JavaScript dependence
- Rate limiting
- TLS fingerprinting
- Web Application Firewall (WAF)

We also analyzed robots.txt as a separate signal.

How we measured each barrier

Antibot / bot mitigation services

To measure how often landing pages deploy dedicated antibot services, we looked for vendor fingerprints in scripts, request patterns, and endpoints that commonly appear when commercial bot mitigation is active.

CAPTCHA

To measure CAPTCHA adoption, we looked for the telltale markers CAPTCHA vendors leave behind - HTML snippets, script loads, and network calls tied to known providers. This lets us reliably measure deployment at scale, including “invisible” CAPTCHAs that don’t always show a challenge to every visitor.

JavaScript dependence

We compared results of plain HTML fetches versus those of fully rendered pages. If rendering increased the meaningful HTML by more than 50%, we counted the page as “JavaScript-dependent”.

Rate limiting

To test whether sites actively throttle bursts of automated traffic, we sent a short burst of requests mixing gentle and faster pacing, then recorded whether the site started rejecting traffic (typically with HTTP 429 or 403) and how quickly that happened. This doesn’t recreate every real-world scraping pattern, but it does reveal whether a site has an aggressive “no bursts” posture.

TLS fingerprinting

Before any page loads, browsers “introduce themselves” in the TLS handshake. To detect whether sites screen clients at the TLS layer before any HTTP request is processed, we inspected that handshake to see whether sites appear to verify the client’s cryptographic fingerprint, ie, whether they’re checking if the request looks like a standard browser.

Zyte recommended tier

Each domain was tested twice with Zyte API - once via plain HTTP and once via headless browser - and assigned a recommended tier based on the minimum infrastructure required by whichever access method demanded more. Tiers run from Simple (plain HTTP, no fingerprint handling) through Easy (JavaScript rendering), Moderate (fingerprint evasion), Complex, and Advanced (full specialised proxy and evasion configuration). The recommended tier is a ceiling, not an average: a site that is Simple via HTTP but Easy via browser is rated Easy overall.

robots.txt

To assess how intentionally sites communicate rules for automated access, we fetched and parsed each site’s robots.txt and extracted signals like disallow rules, crawl delays, and per-agent allow/deny directives.

Tools, complexity proxy and caveats

Tools

- Scrapy Cloud
- Internal Zyte's tier API
- Modified version of caniscrape 0.1.8
- wafw00f 2.3.2's WAF database

Complexity proxy (Zyte tiers)

In addition to barrier presence, we use Zyte's tier system as a proxy for the cost and complexity of access at scale. For each domain, we take the higher of the HTTP tier and browser tier (scale one to five).

Key caveats

Landing page vs. deep pages: Many sites leave homepages open for SEO while defending deeper pages more aggressively. Our numbers reflect what is detectable on landing pages.

Fingerprint-based detection: When we say a barrier is "present," we mean "detectable via the fingerprints/signals we scanned for." Custom or niche solutions may be undercounted.

CAPTCHA readiness vs. served challenges: Detecting CAPTCHA infrastructure does not guarantee a challenge was served during our scan.

Regional dataset reflects the landing pages most often accessed from the country code, not the landing pages of websites from the country.

zyte

Dive deeper

Get even more data from State of Web Access with our web explorer.

Get the data

Zyte blog

Read it now | Sign up for the newsletter

Insights, updates and tips from the new data frontier.

Follow us

Bluesky | X | Mastodon | LinkedIn | Facebook | dev.to

Talk to us

Get in touch

How do you see the year ahead?
Tell us your project's requirements.