THE METHODOLOGY, IN FULL

0 of 25 earned an A

We scored 25 of the best-built web-sites on the internet against this rubric. Not one earned an A.

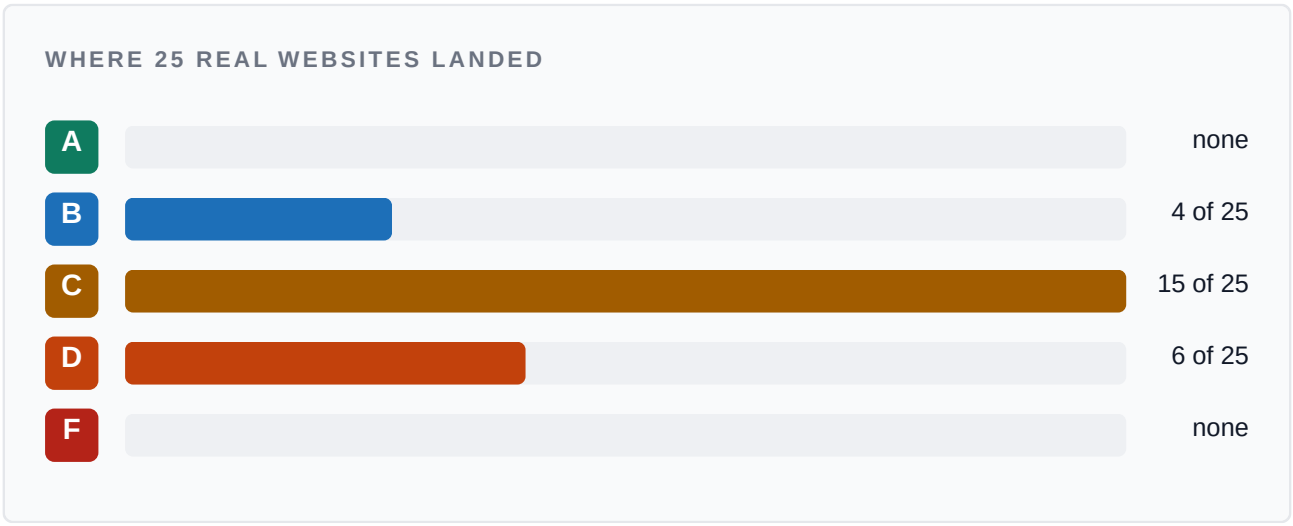
The best of them managed a B. Fifteen landed on a C. 9 scored below 70, the point at which this tool would tell an agency there is a conversation worth having.

CONTENTS

CHAPTER 01

Nobody gets an A

25 real websites, scored on the rubric this document describes. The results are the reason the rest of it is worth reading.



The set

The 25 domains are a fixed benchmark, not a survey. They were chosen to be hard to argue with — cloudflare.com, github.com, mozilla.org, shopify.com, stripe.com, wikipedia.org and nineteen others, ranging from engineering-led companies to ordinary small-business sites.

This file is not marketing material. It is the frozen rubric output the regression suite asserts against: change the scoring and the test fails. Which makes it the one dataset here that cannot be quietly tuned to flatter us — the file you would have to edit to improve these numbers is the file that would break the build.

Two of the 25 are ours, and they are the top two scores. We have left them in and named them — securafy.com and securafyai.com — because removing them after seeing the result is exactly the sort of thing this document exists to argue against. Every figure below holds without them.

What came back

Not one of the 25 earned an A. The best managed a B. The median composite was 74, the range 52 to 91, and 9 landed below 70 — the threshold at which this tool tells an agency there is a real conversation to have.

These are not neglected websites. Several are maintained by companies whose entire business is the web. The point is not that they are bad; it is that "well built" and "scores well across eight categories" turn out to be different things, and the gap between them is where the work is.

> **If sites built by engineering-led companies land on a C, a small business scoring a C is not behind. It is average — and average is a position, not a verdict.**

A FEW YOU WILL RECOGNISE

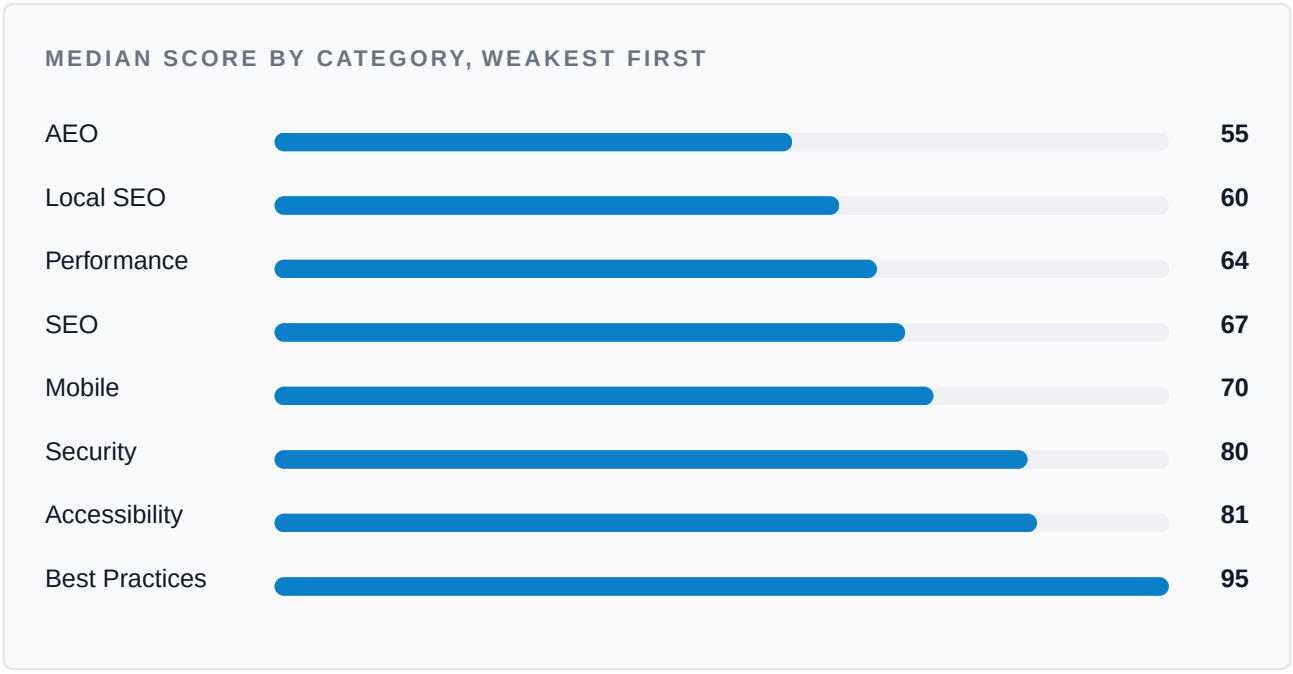
DOMAIN	SCORE	WEAKEST CATEGORY
github.com	74 C	AEO 40
cloudflare.com	77 C	AEO 50
mozilla.org	77 C	AEO 50
wikipedia.org	61 D	AEO 30
shopify.com	71 C	AEO 55

Where they lose it

The weakest category across the set is AEO, median 55, with 8 of 25 scoring under 50. That is the newest thing being measured here and the least attended to anywhere: whether an AI assistant can actually read, understand and cite the page.

The strongest is Best Practices, median 95, with 24 of 25 at 90 or better. It is worth 4% of the composite — the smallest weight of the eight.

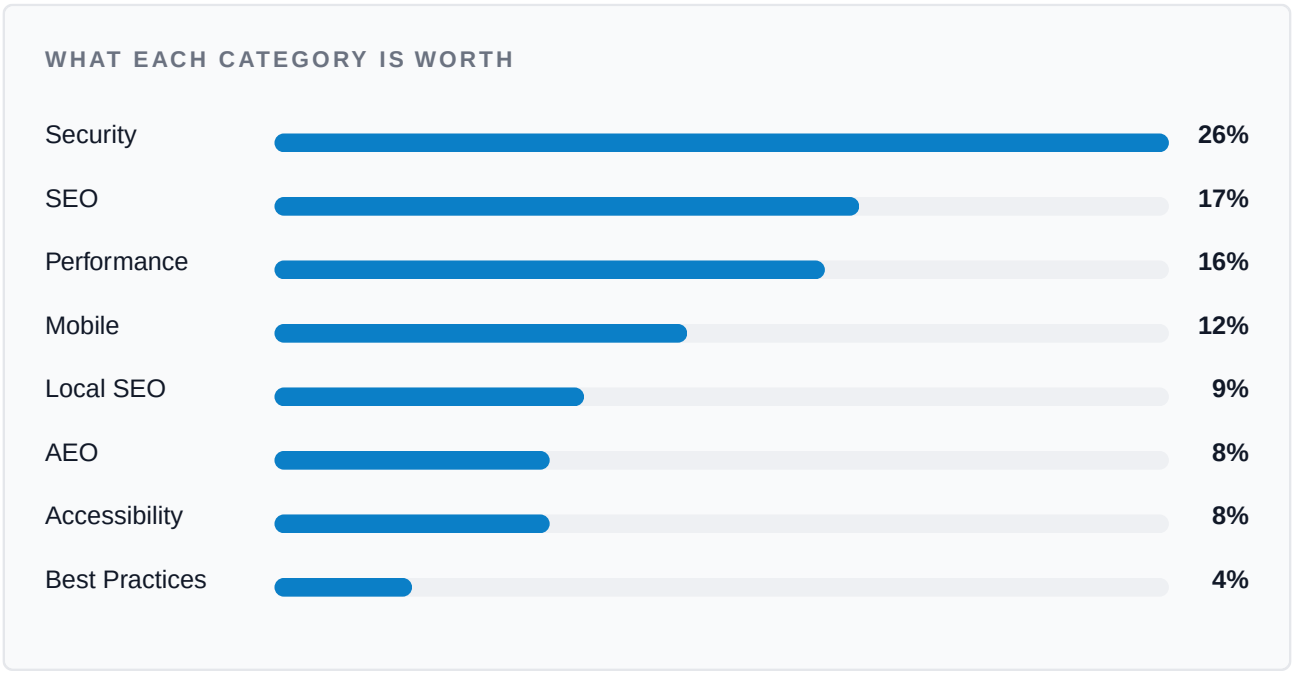
That pairing is the whole argument for weighting. A category almost everyone passes cannot distinguish anyone, and a rubric that let it count equally would be measuring conformity rather than quality.


MEDIAN SCORE BY CATEGORY, WEAKEST FIRST
AEO
55
Local SEO
60
Performance
64
SEO
67
Mobile
70
Security
80
Accessibility
81
Best Practices
95

CHAPTER 02

Why one number

A single score is a summary, not a verdict — and it is only useful if you know what it is summarising.



The problem with a checklist

Most website audits hand back a list. Forty items, each marked pass or fail, sorted in whatever order the tool happened to run them. The list is accurate and almost useless, because it answers a question nobody asked: "what is true about this website?" The question a business owner actually has is "what should I do on Monday?"

A score answers that question badly if it is an average of everything, and well if it is weighted by what actually costs the business something. A missing Permissions-Policy header and a domain that anyone on the internet can send mail as are both "one failed check". They are not remotely the same problem.

> **A grade is a way of ranking your attention, not a way of grading your worth.**

What the weights say

The eight categories do not count equally. Security carries 26% of the composite — the largest single weight — because its failures are the ones with consequences that outlive the website. A slow page costs you a visitor. A domain with no mail authentication costs you a customer who was phished by someone pretending to be you.

Best Practices carries 4%, the smallest, because it is housekeeping: real, worth fixing, and never the reason a business loses a deal.

An earlier version of this rubric averaged its categories evenly. That let a nearly empty page score in the high fifties by riding a fast, unremarkable Performance result and a Best Practices score it could not fail. The weights exist because the unweighted version flattered pages that deserved no flattery.

What the number is not

A composite below 70 is what this tool treats as worth starting a conversation about. That threshold was not chosen by taste — at a stricter setting it fired on roughly seven in ten real websites, including Cloudflare's and Mozilla's. A tool that tells you Cloudflare needs help has a credibility problem, not a lead.

The score is also not a comparison against an industry average, because there is no such thing at a useful resolution. What it does compare against is specific: the actual competitors AI assistants name when asked who to hire in this business's field, each one audited on this same rubric.

CHAPTER 03

How a grade is decided

Five bands, deliberately uneven, because the distance between a D and an F matters more than the distance between an A and a B.

THE BANDS

A	93–100	Nothing here needs attention this quarter.
B	80–92	Sound, with specific items worth scheduling.
C	65–79	Working, but losing ground on things visitors notice.
D	45–64	Several categories are actively costing this business.
F	0–44	Failures a visitor or a search engine will hit immediately.

The bands

The bands are not fifths. An A starts at 93 and a D starts at 45, which means the top band is narrow and the bottom two are wide. That is intentional: at the top, small differences are real and worth distinguishing; at the bottom, a site scoring 30 and a site scoring 44 have the same problem, which is that several categories are broken at once.

Every category is graded on the same bands as the composite, so a category grade and an overall grade mean the same thing. A B in Security is the same standard as a B overall.

Measured, or inferred

Two categories can be reported before they have been measured. Performance needs a live page-speed run, and Security needs DNS and TLS lookups; both take longer than anyone will wait for a first result. So a first scan returns provisional scores from what can be read off the page immediately, marks those categories inferred, and says so on the report.

A second pass then measures for real and re-scores. The numbers move, sometimes substantially, and the report is marked measured. This is the one place the tool will show you a number it intends to change — and it labels it rather than hiding the wait.

> **If a report says inferred, it is telling you it has not finished. A settled report always says measured.**

CHAPTER 04

Security

How well this site protects the people who visit it. 26% of the composite.

WHAT THIS CATEGORY CHECKS

- HTTPS
- SPF Record
- TLS Certificate
- HSTS Header
- X-Content-Type-Options
- Permissions-Policy
- DMARC Policy
- DKIM Signing
- CSP Header
- X-Frame-Options
- Referrer-Policy
- Software Version Disclosure

Measured over DNS and TLS, not read off the page

Every other audit tool in this category grades marketing and calls an SSL certificate security. This one asks a different question: what can someone do to this business using its domain, and what would the business never find out about?

Four of the twelve checks never touch the website. DMARC, SPF, DKIM and the certificate are queried directly over DNS and TLS. That matters because the most expensive failure here is invisible from the page: a domain with no mail authentication can be sent from by anyone, and the first the owner hears of it is a customer asking why they were invoiced twice.

The mail checks are weighted above every response header for exactly that reason. A missing Permissions-Policy is real and worth fixing. A domain anyone on the internet can send mail as is a different order of problem, and a rubric scoring them within two points of each other cannot support a conversation about managed security.

> **A published DMARC policy of p=none is monitoring, not protection. It watches the spoofing happen and does nothing about it.**

Why a valid-looking record can protect nothing

An SPF record ending in -all looks strict. Two conditions make it protect nothing, and neither is visible to the owner reading their own record: the chain can exceed the ten-lookup limit the specification imposes, in which case receivers must ignore the record entirely; or it can include a hostname that no longer resolves, which does the same thing.

This tool walks the chain and reports both. It also distinguishes an authoritative "no such record" from a lookup that simply failed — collapsing those two silently forgives every domain with no TXT records at all, which is the population most in need of the finding.

Every check here is independently verifiable. A dig against the mail policy or a single curl against the headers confirms what the report says, and a check that could not be run is reported as not checked, never as a failure.

CHAPTER 05

SEO

Whether search engines can understand and rank this site. 17% of the composite.

WHAT THIS CATEGORY CHECKS

- Title Tag
- H1 Heading
- Image Alt Text
- Schema Markup
- Meta Description
- Canonical URL
- Open Graph Tags

The part that has not changed

Search engines still need to work out what a page is about, which page is canonical when several are similar, and what to show in a result. The checks here are the ones that answer those questions, and they have been stable for a decade because the underlying problem has not moved.

This is also the category most likely to have been worked on already. An agency that has touched the site at all has usually fixed the title and the description. Which makes a low score here informative in a different way: it usually means nobody has looked at this site in years, and the other seven categories are about to confirm it.

Where it overlaps with being readable by machines

Schema markup appears in this category and again, differently, in AEO. That is deliberate rather than double-counting: a search engine uses structured data to decide how to display a result, and an assistant uses it to decide whether the page can be summarised and cited at all. Same markup, two different consumers, two different standards for "enough".

Open Graph tags are scored here for a similar reason. They are nominally about how a link looks when shared, but they are also the cheapest reliable statement a page makes about its own title, description and subject — which is why an absent set correlates so strongly with everything else in this category being absent too.

CHAPTER 06

Performance

How long a visitor waits before seeing anything. 16% of the composite.

WHAT THIS CATEGORY CHECKS

- Largest Contentful Paint
- Interaction to Next Paint
- Cumulative Layout Shift
- Total Blocking Time

Four numbers, measured on a real page load

These are Google's Core Web Vitals, and they are measured rather than estimated: a real page-speed run against the live URL. That takes twenty to forty seconds, which is why a first scan reports this category as inferred from what could be read off the HTML, then measures for real and re-scores.

Largest Contentful Paint is how long before the visitor sees the main thing. Cumulative Layout Shift is how much the page moves under them while it loads. Interaction to Next Paint is how long a tap takes to do anything. Total Blocking Time is how long the page is busy and unresponsive.

The provisional pass scores different things entirely — the size of the HTML, whether scripts block the head, whether caching headers are set. Those are reasonable predictors of speed and they are not speed, which is why they are replaced rather than averaged in once the real numbers arrive. A settled report shows the four measurements above and nothing else.

> **A report marked inferred has not finished measuring. A settled report always says measured — and the numbers do move.**

Why this is worth less than security

Performance carries real weight, but less than security, and the reason is what each failure costs. A slow page loses you the visitor who was not going to wait. A domain anyone can send mail as loses you a customer who was phished by someone wearing your name — and it keeps costing after the website is fixed.

There is also a measurement honesty point. Page speed varies by network, device and time of day, and a single run is a sample rather than a verdict. Weighting it below the checks that return the same answer every time is the conservative choice.

CHAPTER 07

Mobile

Whether this site works on a phone. 12% of the composite.

WHAT THIS CATEGORY CHECKS

- Viewport Meta Tag
- Responsive Images
- Viewport Zoom Disabled

Three checks, because three things actually break

A missing viewport meta tag means a phone renders the page at desktop width and scales it down, so everything is unreadably small. Disabling zoom means a visitor who needs to enlarge the text cannot — which is both a usability failure and an accessibility one. Non-responsive images mean a phone downloads the desktop-sized file over mobile data.

This is a short list on purpose. Most of what people call mobile usability is design judgement that no automated check can assess honestly, and inventing a score for it would produce a confident number about something never measured.

CHAPTER 08

Local SEO

Whether local customers can find and trust this business. 9% of the composite.

WHAT THIS CATEGORY CHECKS

- Google Business Profile
- Business Hours Listed
- Review Count
- Review Recency
- NAP Consistency
- Profile Photos
- Average Rating

The half of the internet that is not the website

For a business that serves a place, the Google Business Profile does more work than the homepage. It is what appears in the map, what carries the reviews, and what answers "are they open now" — and it is entirely outside the website an agency was hired to improve.

NAP consistency — name, address and phone matching between the site and the profile — is the check that most often fails quietly. A business that moved, changed a phone number, or was listed twice ends up with two versions of itself on the internet, and search engines resolve the ambiguity by trusting neither.

> **Review recency is scored separately from rating. Forty five-star reviews, none in three years, reads as a business that used to be good.**

CHAPTER 09

AEO

Whether AI assistants can read and cite this site. 8% of the composite.

WHAT THIS CATEGORY CHECKS

- Structured Data Coverage
- Author Signals
- FAQPage/HowTo Schema
- FAQ Schema
- Blog/Content Presence

Whether a machine can read you, and whether it will say so

This is the newest category here and the weakest across the benchmark — the median was the lowest of the eight, and a third of the set scored under fifty. It asks whether an AI assistant can parse the page into facts it is willing to state, and attribute them to a source it can name.

Structured data coverage is the largest part of it. An assistant summarising a business benefits enormously from being told, in machine-readable form, what the business is, where it is and what it offers, rather than inferring it from prose. Author signals matter for the same reason a citation does: an assistant asked to recommend somebody is more willing to name a source that identifies who is speaking.

What this category is not

It is not a measure of whether assistants currently name this business. That is a separate measurement entirely — asking ChatGPT, Gemini, Claude and Perplexity who they recommend in the industry, and recording who they actually said. This category measures readiness: whether the page could be read and cited if an assistant went looking.

The two correlate less than you would expect. A business can be widely recommended on reputation while its own site is unreadable to a crawler, and a technically immaculate site can go unmentioned because nothing else on the internet corroborates it.

CHAPTER 10

Accessibility

Whether people using a screen reader or keyboard can use this site. 8% of the composite.

WHAT THIS CATEGORY CHECKS

- Page Language
- Form Labels
- Landmarks
- Frame Titles
- Image Alt Text
- Heading Structure
- Skip Link
- Focus Order

Named as people, not standards

This category asks whether someone using a screen reader or a keyboard can use the site. It is deliberately phrased that way rather than as conformance to a specification, because "WCAG 2.1 AA" means nothing to the business owner reading the report and "someone using a screen reader cannot get past your menu" means something immediately.

It is also the category this kind of scan can say least about with confidence. Automated tools catch missing labels, unlabelled controls and structural problems; they cannot assess whether an interface makes sense to navigate without sight. The weight reflects that honestly rather than implying a full audit has happened.

> **An automated accessibility score is a floor, not a certificate. Passing it means the obvious failures are absent, not that the site is usable.**

CHAPTER 11

Best Practices

Housekeeping that search engines and browsers expect. 4% of the composite.

WHAT THIS CATEGORY CHECKS

- Robots Meta Tag
- Link Structure
- Twitter Card

The category almost everyone passes

Across the benchmark, twenty-four of twenty-five scored ninety or better here. It is worth the smallest weight of the eight, and those two facts are the same fact: a category nearly everyone passes cannot distinguish anyone.

It is also the shortest list — three checks, not the dozen Security runs. That is deliberate. Housekeeping worth scoring is housekeeping with a consequence, and most of what tools file under best practices is convention with no observable effect on anything.

That does not make these pointless. A robots meta tag accidentally set to noindex removes a site from search results entirely, and it is the single most damaging one-line mistake on this list. But damage that rare and that binary is better reported as a finding than folded into a score.

> **If a rubric weighted this equally with security, it would be measuring conformity rather than quality — and every site in the benchmark would score higher than it deserves.**