

AccessAudit — Benchmark Report

Benchmark: AccessAudit vs axe-core

<https://dequeuniversity.com/demo/mars/>

Generated: 28 June 2026

1. Coverage Comparison

Coverage measures how many unique WCAG rule categories each tool detects, regardless of verified ground truth.

22

AccessAudit — unique rule categories

110 total violations

13

axe-core only — unique rule categories

64 total violations

AccessAudit detected **9 additional rule categories** (+69% more than axe-core baseline). A lower AccessAudit compliance score (77 vs 86) is expected — higher detection sensitivity means more issues are reported.

2. Precision / Recall / F1 — Criterion Level

Metrics computed against **13 manually verified violation instances** mapped to **13 affected WCAG criteria** (source: deque_official). Matching is performed by WCAG criterion across a universe of **94 total WCAG rule categories**.

Criterion-level evaluation: A criterion is counted as detected if the tool reported at least one violation for that WCAG criterion. Criterion-level precision of 100% means every flagged criterion corresponds to a verified violation — not that every individual element-level finding is correct. Instance-level validation requires manual expert review of individual findings.

METRIC	ACCESSAUDIT	AXE-CORE	IMPROVEMENT
F1 Score $2 \times P \times R / (P + R)$	66.7%	50.0%	+16.7 pp
Precision $TP / (TP + FP)$	72.7%	71.4%	+1.3 pp
Recall $TP / (TP + FN)$	61.5%	38.5%	+23.0 pp
Severity-Weighted Recall $\sum (w \times TP) / \sum (w \times (TP + FN)) \mid$ crit=4, ser=3, mod=2, min=1	64.1%	41.0%	+23.1 pp
MCC $(TP \times TN - FP \times FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$	0.621	0.473	+0.148
True Positives (TP) Known violations correctly detected	8	5	+3
False Positives (FP) Detected but not in ground truth	3	2	1
False Negatives (FN) Known violations missed	5	8	-3
True Negatives (TN) Rules correctly not flagged (of 94 total)	78	79	—

3. Recall by WCAG Principle

Recall per WCAG principle shows which accessibility categories benefit most from AccessAudit's additional detection layers.

WCAG PRINCIPLE	GT ISSUES	AA FOUND	AXE FOUND	AA RECALL	AXE RECALL
1. Perceivable	6	3	2	50.0%	33.3%
2. Operable	5	3	1	60.0%	20.0%
3. Understandable	1	1	1	100.0%	100.0%
4. Robust	1	1	1	100.0%	100.0%

4. Per-Criterion Breakdown

Detection status for each WCAG criterion in the verified ground truth.

☐ Both tools detected ☐ AccessAudit only ☐ axe-core only ☐ Neither detected

WCAG 1.1.1

CRITICAL

Perceivable

axe-core

AA ✓

AXE ✓

Missing and inappropriate alt text on multiple images including spacer, offer images, and logo

WCAG 2.1.1

CRITICAL

Operable

Keyboard focus skips form fields, inaccessible interactive controls throughout booking widget

WCAG 1.2.3

MODERATE

Perceivable

Videos lack audio descriptions for visual content without narration

WCAG 1.2.8

MODERATE

Perceivable

No transcripts provided for any video content

WCAG 1.2.2

SERIOUS

Perceivable

First video lacks captions completely; second video has no captions for first minute

WCAG 1.3.1

SERIOUS

Perceivable



Custom DOM rule

AA ✓

Improper heading structure, missing landmarks, missing/orphaned form labels, fieldsets without legends

WCAG 1.4.3

SERIOUS

Perceivable



Contrast sampling

AA ✓

AXE ✓

Color contrast failures on Facebook login, navigation links, headings, and footer text

WCAG 2.2.2

SERIOUS

Operable

Special offers auto-scroll with no pause mechanism available

WCAG 2.4.1

SERIOUS

Operable



Keyboard simulation

AA ✓

No skip navigation link present

WCAG 2.4.4

SERIOUS

Operable



axe-core

AA ✓

AXE ✓

Links without discernible text — navigation, footer images are background images with no text

WCAG 2.4.7

SERIOUS

Operable



Keyboard simulation

AA ✓

Visual keyboard focus indicator broken or missing in multiple areas

WCAG 3.1.1

SERIOUS

Understandable



axe-core

AA ✓

AXE ✓

Language attribute missing from document

WCAG 4.1.2

SERIOUS

Robust



axe-core

AA ✓

AXE ✓

Empty buttons throughout booking widget detectable with automated tools

False Negatives — not detected by AccessAudit (5)

WCAG 2.1.1

CRITICAL

Keyboard focus skips form fields, inaccessible interactive controls throughout booking widget

Requires interactive testing or manual verification — automated detection not yet possible.

WCAG 1.2.3

MODERATE

Videos lack audio descriptions for visual content without narration

Requires interactive testing or manual verification — automated detection not yet possible.

WCAG 1.2.8

MODERATE

No transcripts provided for any video content

Requires interactive testing or manual verification — automated detection not yet possible.

WCAG 1.2.2

SERIOUS

First video lacks captions completely; second video has no captions for first minute

Requires interactive testing or manual verification — automated detection not yet possible.

WCAG 2.2.2

SERIOUS

Special offers auto-scroll with no pause mechanism available

Requires interactive testing or manual verification — automated detection not yet possible.

5. Computational Overhead

Time and API cost required for AccessAudit's additional detection layers compared to a baseline axe-core scan.

19.2s

AccessAudit scan time

Full hybrid scan

6.6s

axe-core scan time

Baseline scan

+12.6s

Overhead

Additional time

COST COMPONENT	DETAILS	ESTIMATED COST (USD)
Anthropic API	101,354 tokens (63,705 in / 37,649 out)	\$0.20963
Browserless	19.2s @ \$1.00/hr (cloud Chromium)	\$0.00535
Total (estimated)	Per single-page scan	\$0.21498

Time efficiency: **0.55s per +1 pp Recall gained** (12.6s ÷ 23.0 pp). Cost efficiency: **\$0.00935 per +1 pp Recall gained.**

Overhead breakdown: keyboard navigation simulation (30 Tab keypresses) · 320px reflow viewport test · custom contrast ratio sampling · form error simulation · DOM landmark and heading hierarchy analysis · skip-link detection · touch target size measurement. None available in a static axe-core scan.

Disclaimer: This is a preliminary single-page benchmark. Precision/Recall metrics are computed against 13 manually verified violation instances mapped to 13 WCAG criteria. Multi-page evaluation across diverse site types and instance-level validation are required for final dissertation confirmation. Automated tools detect approximately 40–60% of WCAG violations; manual expert review is required for full compliance verification.