

AccessAudit — Benchmark Report

Benchmark: AccessAudit vs axe-core

https://a11y-assessments.pages.oit.duke.edu/accessible-u/before_u.html

Generated: 28 June 2026

1. Coverage Comparison

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

24

AccessAudit — unique rule categories

105 total violations

13

axe-core only — unique rule categories

52 total violations

AccessAudit detected **11 additional rule categories** (+85% more than axe-core baseline). A lower AccessAudit compliance score (74 vs 86) is expected — higher detection sensitivity means more issues are reported.

2. Precision / Recall / F1 — Criterion Level

Metrics computed against **17 manually verified violation instances** mapped to **14 affected WCAG criteria** (source: duke_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 $2 \times P \times R / (P + R)$	96.3%	72.7%	+23.6 pp
Precision $TP / (TP + FP)$	100.0%	100.0%	+0.0 pp
Recall $TP / (TP + FN)$	92.9%	57.1%	+35.8 pp
Severity-Weighted Recall $\sum (w \times TP) / \sum (w \times (TP + FN)) \mid$ crit=4, ser=3, mod=2, min=1	90.7%	60.5%	+30.2 pp
MCC $(TP \times TN - FP \times FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$	0.958	0.729	+0.229
True Positives (TP) Known violations correctly detected	13	8	+5
False Positives (FP) Detected but not in ground truth	0	0	0
False Negatives (FN) Known violations missed	1	6	-5
True Negatives (TN) Rules correctly not flagged (of 94 total)	80	80	—

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	7	7	5	100.0%	71.4%
2. Operable	5	4	1	80.0%	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 ✓

Images missing alternative text

WCAG 1.4.4

CRITICAL

Perceivable

axe-core

AA ✓

AXE ✓

Viewport meta disables text scaling

WCAG 2.1.2

CRITICAL

Operable

Modal dialog does not trap focus

WCAG 4.1.2

CRITICAL

Robust

axe-core

AA ✓

AXE ✓

Invalid ARIA attributes and roles

WCAG 1.3.1

MODERATE

Perceivable

Custom DOM rule

AA ✓

AXE ✓

Missing landmark regions

WCAG 1.4.10

MODERATE

Perceivable

Reflow test (320px)

AA ✓

Horizontal scrolling at 320px viewport

WCAG 2.4.1

MODERATE

Operable

Keyboard simulation

AA ✓

No skip navigation link

WCAG 1.3.5

SERIOUS

Perceivable

Custom DOM rule

AA ✓

Personal data fields missing autocomplete

WCAG 1.4.1

SERIOUS

Perceivable

axe-core

AA ✓

AXE ✓

Links not distinguished from surrounding text

WCAG 1.4.3

SERIOUS

Perceivable

axe-core

AA ✓

AXE ✓

Insufficient color contrast ratios

WCAG 2.4.4

SERIOUS

Operable

axe-core

AA ✓

AXE ✓

Links without discernible text

WCAG 2.4.7

SERIOUS

Operable

Keyboard simulation

AA ✓

No visible focus indicator

WCAG 2.5.8

SERIOUS

Operable

Target size rule

AA ✓

Touch targets below 24x24px minimum

WCAG 3.1.1

SERIOUS

Understandable

axe-core

AA ✓

AXE ✓

Missing document language attribute

False Negatives — not detected by AccessAudit (1)

WCAG 2.1.2

CRITICAL

Modal dialog does not trap focus

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.

16.5s AccessAudit scan time Full hybrid scan	4.8s axe-core scan time Baseline scan	+11.6s Overhead Additional time
---	--	--

COST COMPONENT	DETAILS	ESTIMATED COST (USD)
Anthropic API	88,927 tokens (58,572 in / 30,355 out)	\$0.17660
Browserless	16.5s @ \$1.00/hr (cloud Chromium)	\$0.00457
Total (estimated)	Per single-page scan	\$0.18117

Time efficiency: **0.33s per +1 pp Recall gained** (11.6s ÷ 35.8 pp). Cost efficiency: **\$0.00506 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 17 manually verified violation instances mapped to 14 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.

AccessAudit — Benchmark Report

Generated: 28 June 2026 · Not a legal compliance certificate