

AccessAudit — Benchmark Report

Benchmark: AccessAudit vs axe-core

<https://www.w3.org/WAI/demos/bad/before/home.html>

Generated: 30 June 2026

1. Coverage Comparison

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

15

AccessAudit — unique rule categories

66 total violations

8

axe-core only — unique rule categories

50 total violations

AccessAudit detected **7 additional rule categories** (+88% more than axe-core baseline). A lower AccessAudit compliance score (84 vs 91) is expected — higher detection sensitivity means more issues are reported.

2. Precision / Recall / F1 — Criterion Level

Metrics computed against **18 manually verified violation instances** mapped to **18 affected WCAG criteria** (source: w3c_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 $2 \times P \times R / (P + R)$	51.9%	41.7%	+10.2 pp
Precision $TP / (TP + FP)$	77.8%	83.3%	-5.5 pp
Recall $TP / (TP + FN)$	38.9%	27.8%	+11.1 pp
Severity-Weighted Recall $\sum (w \times TP) / \sum (w \times (TP + FN)) \mid$ crit=4, ser=3, mod=2, min=1	41.5%	30.2%	+11.3 pp
MCC $(TP \times TN - FP \times FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$	0.485	0.426	+0.059
True Positives (TP) Known violations correctly detected	7	5	+2
False Positives (FP) Detected but not in ground truth	2	1	1
False Negatives (FN) Known violations missed	11	13	-2
True Negatives (TN) Rules correctly not flagged (of 94 total)	74	75	—

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	6	2	1	33.3%	16.7%
3. Understandable	4	1	1	25.0%	25.0%
4. Robust	2	1	1	50.0%	50.0%

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 or have inadequate text alternatives; CSS background images convey information without text equivalent

WCAG 2.1.1

CRITICAL

Operable

Not all functionality operable via keyboard; scripts remove focus on receive

WCAG 1.4.5

MODERATE

Perceivable

Images of text used where CSS text could achieve the same visual presentation

WCAG 2.4.6

MODERATE

Operable

Headings not programmatically determined; do not describe topic or purpose

WCAG 4.1.1

MODERATE

Robust

Incorrect use of start/end tags and attribute markup

WCAG 1.3.1

SERIOUS

Perceivable



Custom DOM rule

AA ✓

Information and relationships not programmatically determined; form labels not associated with controls

WCAG 1.3.2

SERIOUS

Perceivable

Meaningful reading sequence broken by CSS positioning and layout tables

WCAG 1.4.1

SERIOUS

Perceivable

Color used as the only visual means to distinguish links from surrounding text

WCAG 1.4.3

SERIOUS

Perceivable



axe-core

AA ✓

AXE ✓

Text does not meet minimum 4.5:1 contrast ratio

WCAG 2.4.1

SERIOUS

Operable

No mechanism to bypass blocks of repeated content

WCAG 2.4.2

SERIOUS

Operable

Page title does not identify contents

WCAG 2.4.4

SERIOUS

Operable



axe-core

AA ✓

AXE ✓

Link purpose cannot be determined from link text alone; null alt on image links

WCAG 2.4.7

SERIOUS

Operable



Keyboard simulation

AA ✓

Keyboard focus indicator removed by script; CSS borders hide visual focus

WCAG 3.1.1

SERIOUS

Understandable



axe-core

AA ✓

AXE ✓

Default human language of page cannot be programmatically determined

WCAG 3.2.1

SERIOUS

Understandable

Change of context initiated on focus via script

WCAG 3.2.2

SERIOUS

Understandable

Change of context on input without prior warning — select list changes context

WCAG 3.3.2

SERIOUS

Understandable

Labels not provided for form controls

WCAG 4.1.2

SERIOUS

Robust

 axe-core

AA ✓

AXE ✓

Name and role of UI components not programmatically determinable; focus state not available

False Negatives — not detected by AccessAudit (11)

WCAG 2.1.1

CRITICAL

Not all functionality operable via keyboard; scripts remove focus on receive

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

WCAG 1.4.5

MODERATE

Images of text used where CSS text could achieve the same visual presentation

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

WCAG 2.4.6

MODERATE

Headings not programmatically determined; do not describe topic or purpose

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

WCAG 4.1.1

MODERATE

Incorrect use of start/end tags and attribute markup

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

WCAG 1.3.2**SERIOUS**

Meaningful reading sequence broken by CSS positioning and layout tables

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

WCAG 1.4.1**SERIOUS**

Color used as the only visual means to distinguish links from surrounding text

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

WCAG 2.4.1**SERIOUS**

No mechanism to bypass blocks of repeated content

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

WCAG 2.4.2**SERIOUS**

Page title does not identify contents

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

WCAG 3.2.1**SERIOUS**

Change of context initiated on focus via script

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

WCAG 3.2.2**SERIOUS**

Change of context on input without prior warning — select list changes context

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

WCAG 3.3.2**SERIOUS**

Labels not provided for form controls

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.

62.2s AccessAudit scan time Full hybrid scan	12.8s axe-core scan time Baseline scan	+49.4s Overhead Additional time
---	---	--

COST COMPONENT	DETAILS	ESTIMATED COST (USD)
Anthropic API	61,575 tokens (38,930 in / 22,645 out)	\$0.15262
Browserless	62.2s @ \$1.00/hr (cloud Chromium)	\$0.01727
Total (estimated)	Per single-page scan	\$0.16989

Time efficiency: **4.45s per +1 pp Recall gained** (49.4s ÷ 11.1 pp). Cost efficiency: **\$0.01531 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 18 manually verified violation instances mapped to 18 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: 30 June 2026 · Not a legal compliance certificate