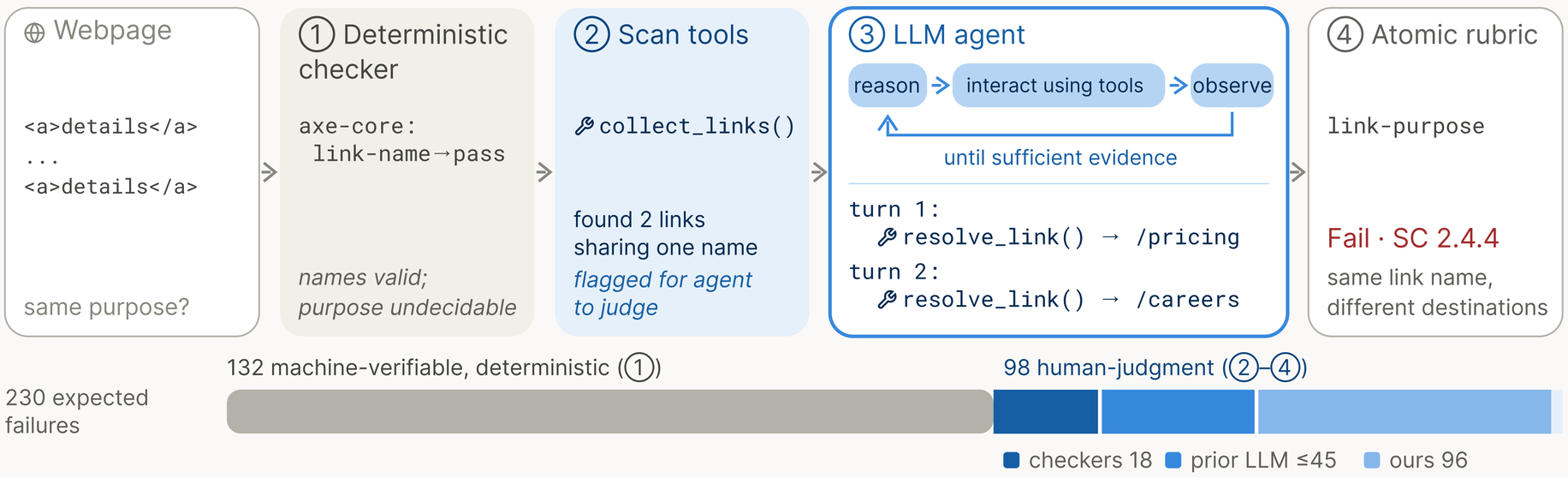
Targeted Interaction Agents for Web Accessibility Evaluation

Mingyuan Zhong, Ajit Mallavarapu, Mengqi Shi, Davin Win Kyi, James Fogarty, Jacob O. Wobbrock

University of Washington



Existing checkers cover 132 machine-verifiable failures. Our tools and rubrics target the 98 human-judgment failures and detect 96 of them.

Problem & Data

Where checkers fail

Recent LLM-driven checkers improve coverage but still miss failures and introduce many false positives. Our error analysis identifies three recurring gaps: semantic meaning, rendered perception, and interaction evidence.

Deterministic checkers miss failures in three areas:

Meaning
Semantic judgment is needed, such as deciding whether a label is descriptive.

Perception
Rendered evidence is needed, such as determining whether an image contains text.

Interaction
Interactive evidence is needed, such as testing keyboard traps or error messages.

ACT dataset & partition

We selected ACT test cases using the frequency of manually detected issues in Deque's large-scale web audit dataset.

Our dataset contains 770 ACT test cases covering 49 ACT rules and 20 WCAG Success Criteria. Of these, 230 are expected failures.

Human-judgment rules contain 42.6% of expected failures, yet no deterministic checker exceeds 18.4% recall on this subset.

Approach

System design

Deterministic component

We combine QualWeb, axe-core, and IBM Equal Access, to utilize their reliable deterministic components. In total, they identify:

53.0% of expected failures at 98.4% precision.

15 scan tools

Run once per page to record behavior and enumerate candidate issues using keyboard navigation, screen-reader traversal, and screenshots.

14 agentic tools

Called by the agent to gather evidence by resolving links, running OCR, resizing pages, and interacting with elements.

21 atomic rubrics for 14 SC

Each rubric decides one narrow question about one Success Criterion. It specifies the required evidence and decision method. The agent follows the rubric, calls tools as needed, and returns a verdict.

Example — SC 2.4.4 (Link Purpose) is decided by two rubrics:

`link-purpose`
purpose clear in context

`link-name-equivalence`
links with identical names serve equivalent purpose

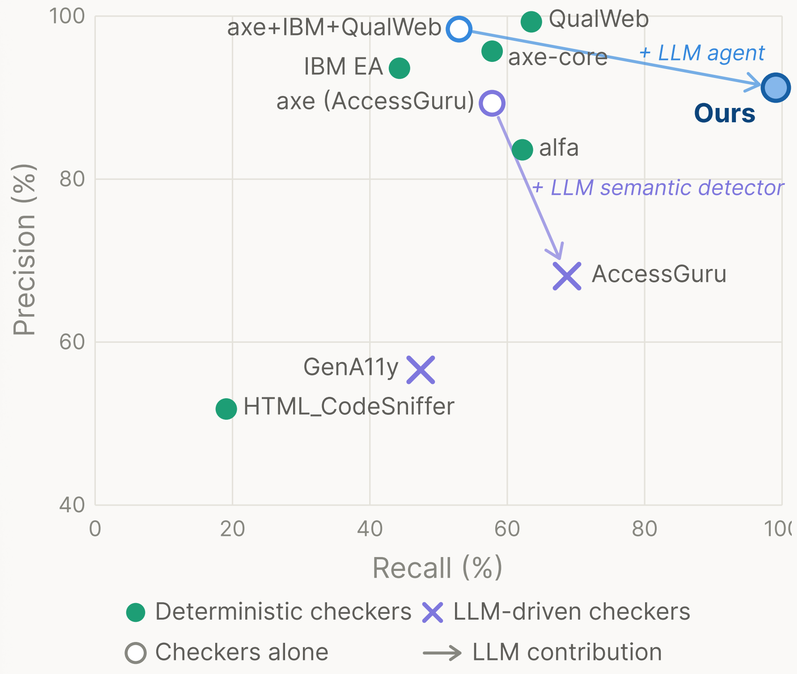
The remaining six Success Criteria are handled entirely by the deterministic component and do not require agent analysis.

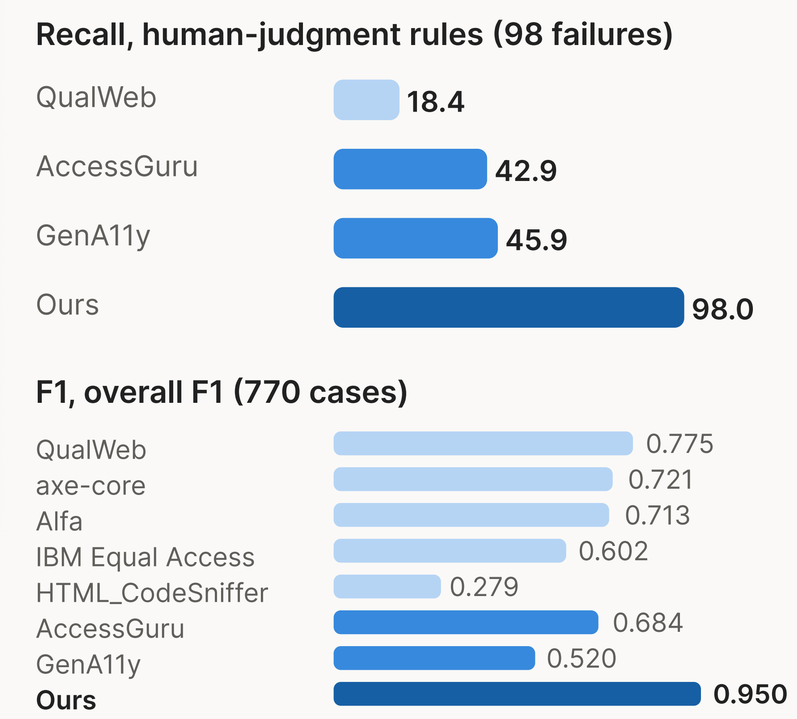
Evaluation

Results

On the 770-case ACT dataset, our system detects 99.1% of expected failures, compared to 68.7% for the best baseline (LLM-driven). Our overall precision of 91.2% is close to the best deterministic checkers (93–99%).

On human-judgment rules, our system detects 98.0% of failures at 86.5% precision (F1 = 0.92), while prior LLM checkers reach at most 0.50 F1.





Dataset: W3C ACT rules test cases [17]. Issue ranking: Deque Automated Accessibility Coverage Report [3].
Baselines: AccessGuru [5], GenA11y [6]. LLM used: Claude Sonnet 4.6.