Mind the Gap: Do Widely-Used SAST Tools Really Cover the Language-Specific Top CWEs?

Zhilin Li
Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences
Beijing, China
University of Chinese Academy of Sciences
Beijing, China
lizl@ios.ac.cn

Yiran Nie
Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
Hangzhou, China
nieyiran25@mails.ucas.ac.cn

Xianglong Qi
Beijing University of Technology
Beijing, China
qixianglong@emails.bjut.edu.cn

Jiwei Yan*
Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences
Beijing, China
Beijing Key Laboratory of Intelligent Software Engineering
Beijing, China
University of Chinese Academy of Sciences
Beijing, China
yanjiwei@otcaix.iscas.ac.cn

Jun Yan*
Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences
Beijing, China
Key Laboratory of System Software (Chinese Academy of Sciences)
Beijing, China
Beijing Key Laboratory of Intelligent Software Engineering
Beijing, China
Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
Hangzhou, China
yanjun@ios.ac.cn

Abstract

This is the data artifact for the paper "Mind the Gap: Do SAST Tools Cover High-Frequency CWEs across Programming Languages?". We are applying for the Artifacts Available badge. We have publicly released our comprehensive benchmark dataset, including 197,016 language-annotated CVE records, the CWE View 1000 hierarchical taxonomy, detailed checker rule mappings for nine popular SAST tools, and the full experimental evaluation matrices. All artifact files have been permanently archived on Zenodo for public access and verification.

CCS Concepts

• **Security and privacy → Software security engineering**; • **Software and its engineering → Software testing and debugging**.

*Corresponding author.

ASE '26, Munich, Germany

ACM ISBN 979-8-4007-2882-2/2026/10
https://doi.org/10.1145/3832783.3837435

Keywords

Static Analysis, Software Security, SAST Tool, CVE/CWE

1 Title

Mind the Gap: Do SAST Tools Cover High-Frequency CWEs across Programming Languages?

2 Link

https://zenodo.org/records/19122421

3 Purpose

Static Application Security Testing (SAST) tools are critical for identifying software vulnerabilities early. However, existing evaluations often focus on language-agnostic rankings like the MITRE CWE Top 25. This artifact provides the benchmark datasets, CWE taxonomy mappings, and experimental evaluation matrices used in our research paper to evaluate nine popular SAST tools regarding their coverage of language-specific high-frequency CWEs across C/C++, Java, and Python.

4 Badge

We formally apply for the Artifacts Available badge. Our artifact satisfies all requirements specified in the ACM Artifact Review

and Badging Version 1.1. The complete artifact package is publicly archived on Zenodo with a registered and persistent DOI, ensuring long-term availability and accessibility. All underlying datasets and resources are included with clear provenance information, allowing users to understand their sources and construction processes. In addition, the artifact provides comprehensive documentation, including usage guidelines and relevant instructions, and is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license to facilitate open access, reuse, and further research.

5 Technology

The artifact assumes basic familiarity with common data handling tasks, including opening and inspecting compressed archives, structured text/JSON files, and spreadsheet files (.xlsx).

The artifact can be used on any standard commodity laptop or desktop computer running either x86-64 or ARM architecture. A minimum of 2 GB RAM and 250 MB of available disk space is recommended (the compressed package is approximately 150 MB). No high-performance server, GPU, or specialized hardware is required.

The required software environment includes a standard operating system, an archive extraction utility such as tar or 7-Zip, and a spreadsheet viewer for inspecting the provided data files.

6 Provenance

The artifact is permanently archived and publicly accessible through Zenodo. It is assigned a persistent Digital Object Identifier (DOI: https://doi.org/10.5281/zenodo.19122421), ensuring long-term availability and citability.

7 Instructions

Reviewers can access and verify the artifact through the following steps. First, download the Top-CWE-Coverage.tar.gz package from the Zenodo repository and extract it using a standard archive utility (e.g., tar -zxvf Top-CWE-Coverage.tar.gz). After extraction, a successful setup should contain three subdirectories and one Excel file: CVEData/, which includes raw and annotated CVE records as well as sampling validation results; CWEView1000/, which provides the hierarchical parent-child relationship taxonomy of CWEs; ToolData/, which contains rule-to-CWE mappings for the nine evaluated SAST tools; and RQ Result.xlsx, which summarizes evaluation metrics and supporting data tables for all research questions. As a smoke test, reviewers can open RQ Result.xlsx using any standard spreadsheet viewer to confirm that the artifact is correctly extracted and the provided data can be loaded without corruption.

The artifact contents are organized to facilitate verification of the corresponding research questions. For RQ1 (Language Distributions), reviewers can inspect the CVE records in CVEData/ and the RQ1 summary sheets in RQ Result.xlsx. For RQ2 (Tool Coverage Gaps), reviewers can examine the SAST tool rule specifications in ToolData/ and cross-reference them with the CWE taxonomy provided in CWEView1000/. For RQ3 (Multi-Language Comparison), reviewers can refer to the cross-language checker evaluation matrices and related analysis results provided in RQ Result.xlsx.