Benchmark · Skylos 5.3.0 · 1 November 2026

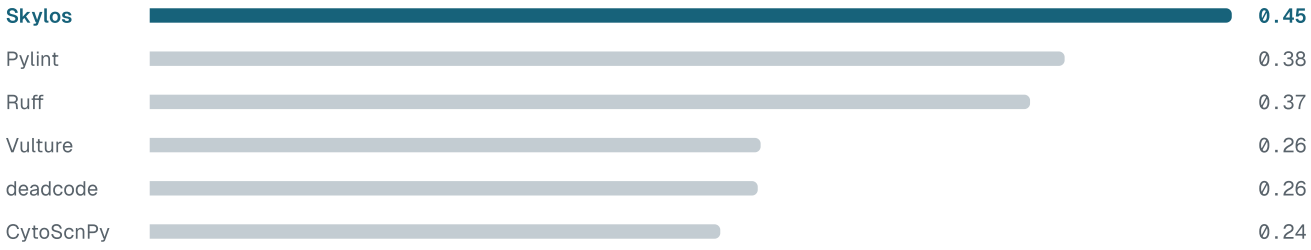
Skylos vs Vulture, Ruff, Pylint, Bandit and Semgrep

Highest overall score in dead code and security, on 43 open-source Python projects and 14,834 findings checked against the code.

Download the report (PDF) | Full dataset (ZIP)

Dead code

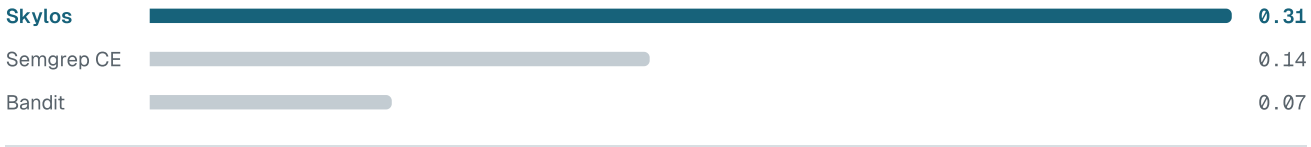
Overall score (F1)



Higher score than Pylint on 30 and than Ruff on 31 of 43 projects.

Security

Overall score (F1)



107 real issues found. Semgrep CE 30, Bandit 40.

What we scanned · 43 projects · 745,317 lines of Python · 26 blind, never scanned by Skylos before

labmlai/annotated_deep_learning_paper_implementations · django/asgiref · wbond/asn1crypto · MagicStack/asyncpg · authlib/authlib · OpenBMB/ChatDev · CloakHQ/CloakBrowser · docker/docker-py · fishaudio/fish-speech · googleapis/google-api-python-client · Pythagora-io/gpt-pilot · microsoft/graphrag · hankcs/HanLP · p-e-w/heretic · httplib2/httplib2 · khoj-ai/khoj · google/langextract · soxoj/maigret · ocrmypdf/OCRmyPDF · PDFMathTranslate/PDFMathTranslate · pre-commit/pre-commit · rapidfuzz/RapidFuzz · ScrapeGraphAI/Scrapegraph-ai · searxng/searxng · google-research/timesfm · ipython/traitlets · python-attrs/attrs · pallets/click · fastapi/full-stack-fastapi-template · pallets/flask · encode/httpx · locustio/locust · NanmiCoder/MediaCrawler · miguelgrinberg/microblog · FoundationAgents/OpenManus · pydantic/pydantic · danielgatis/rembg · psf/requests · Textualize/rich · encode/starlette · tqdm/tqdm · ultralytics/yolov5 · soimort/you-get

Head to head

Skylos | Menu | Pricing | Design partners | Sign in

Dead code

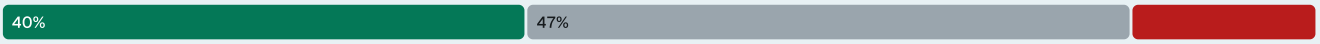
6 tools · 13,241 findings checked

Right | Kept on purpose | Wrong

Skylos 5.3.0

3,036 findings

F1 **0.45**



Found 896

Highest overall score.

Pylint 4.1.3

2,608 findings

F1 **0.38**



Found 668

Same accuracy, but only checks unused imports, variables, arguments and private members. Skylos finds more.

Ruff 0.17.3

734 findings

F1 **0.37**



Found 449

More accurate, but Skylos finds 2.0× more.

Vulture 2.16

5,124 findings

F1 **0.26**



Found 835

17% right vs 40% for Skylos. Skylos finds more.

deadcode 2.4.1

4,948 findings

F1 **0.26**



Found 810

18% right vs 40% for Skylos. Skylos finds more.

CytoScnPy 1.3.0

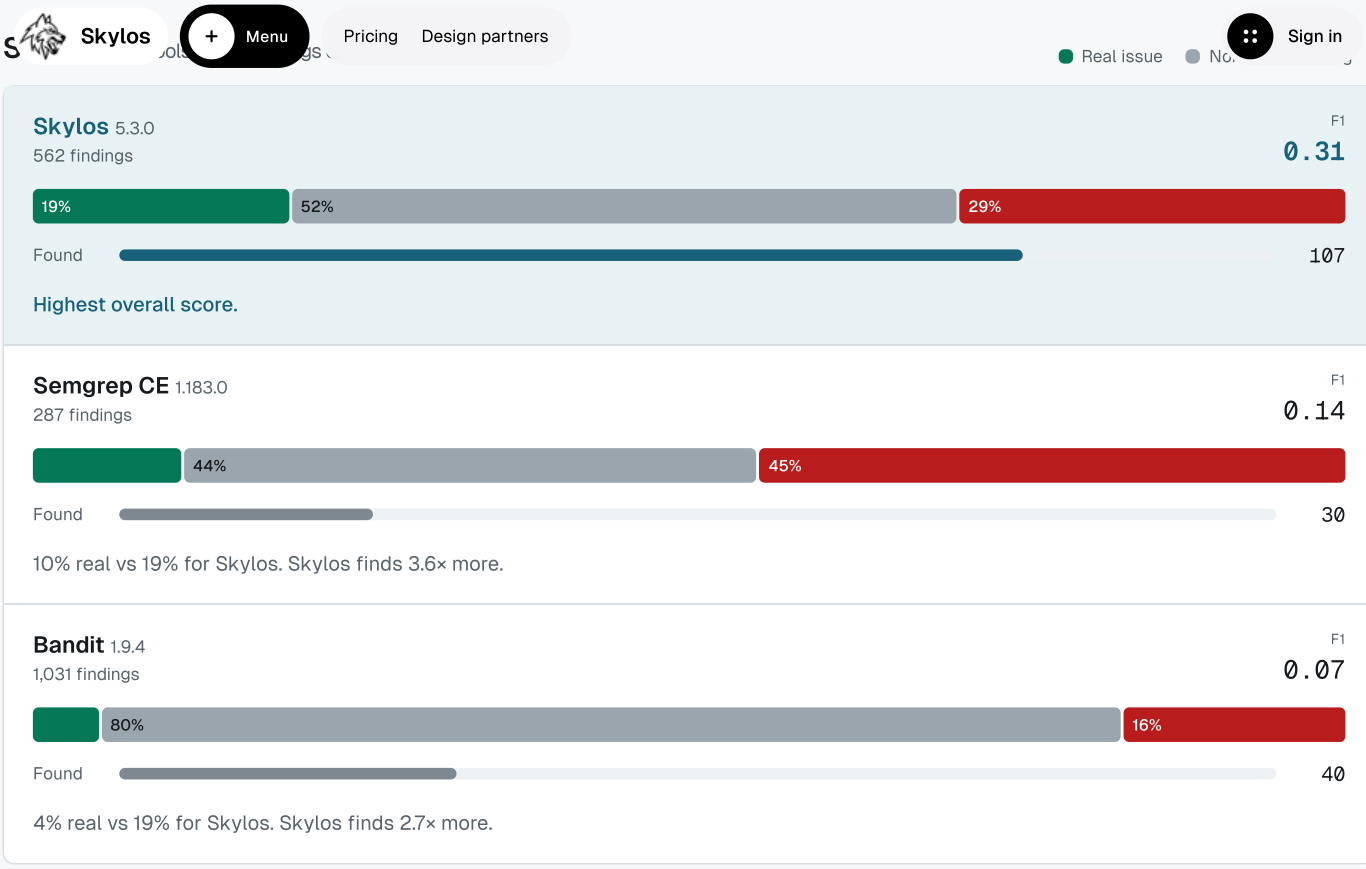
10,059 findings

F1 **0.24**



Found 1,301

Finds more, but 39% of it is wrong (Skylos: 13%).



76 of Skylos's 107 real issues are undeclared dependencies, a supply-chain check Bandit and Semgrep CE do not run.

Never seen vs seen before

Earlier rounds used some of these projects, so the never-seen ones are the fair test. Skylos scores highest on both.

Dead code · F1

Never seen · 26 projects		Seen before · 17 projects	
Skylos	**0.40**	**Skylos**	**0.67**
Pylint	0.34	Pylint	0.48
Ruff	0.29	Ruff	0.51
Vulture	0.25	Vulture	0.28
deadcode	0.25	deadcode	0.27
CytoScnPy	0.23	CytoScnPy	0.28

Security · F1

Never seen · 25 projects		Seen before · 17 projects	
Skylos	**0.30**	**Skylos**	**0.32**
Semgrep CE	0.11	Semgrep CE	0.22
Bandit	0.07	Bandit	0.07

Is it luck?

Two checks: how many projects Skylos wins, and how often it stays ahead when we re-draw the projects 4,000 times.

Dead code

vs Pylint



Projects won — 30 of 43, p 0.002

Re-draws won — 99%

vs Ruff



Projects won — 31 of 43, p < 0.001

Re-draws won — 98%

vs Vulture



Projects won — 32 of 43, p < 0.001

Re-draws won — 100%

vs deadcode



Projects won — 31 of 43, p < 0.001

Re-draws won — 100%

vs CytoScnPy



Projects won — 32 of 43, p < 0.001

Re-draws won — 100%

Each dot is one project where at least one tool reported something. Teal: Skylos scored higher there. Gray: the other tool did. Hollow: tied.

Security

vs Semgrep CE



Projects won — 13 of 42, p 0.17

Re-draws won — 99%

vs Bandit



Projects won — 17 of 42, p 0.005

Re-draws won — 100%

Each dot is one project where at least one tool reported something. Teal: Skylos scored higher there. Gray: the other tool did. Hollow: tied.

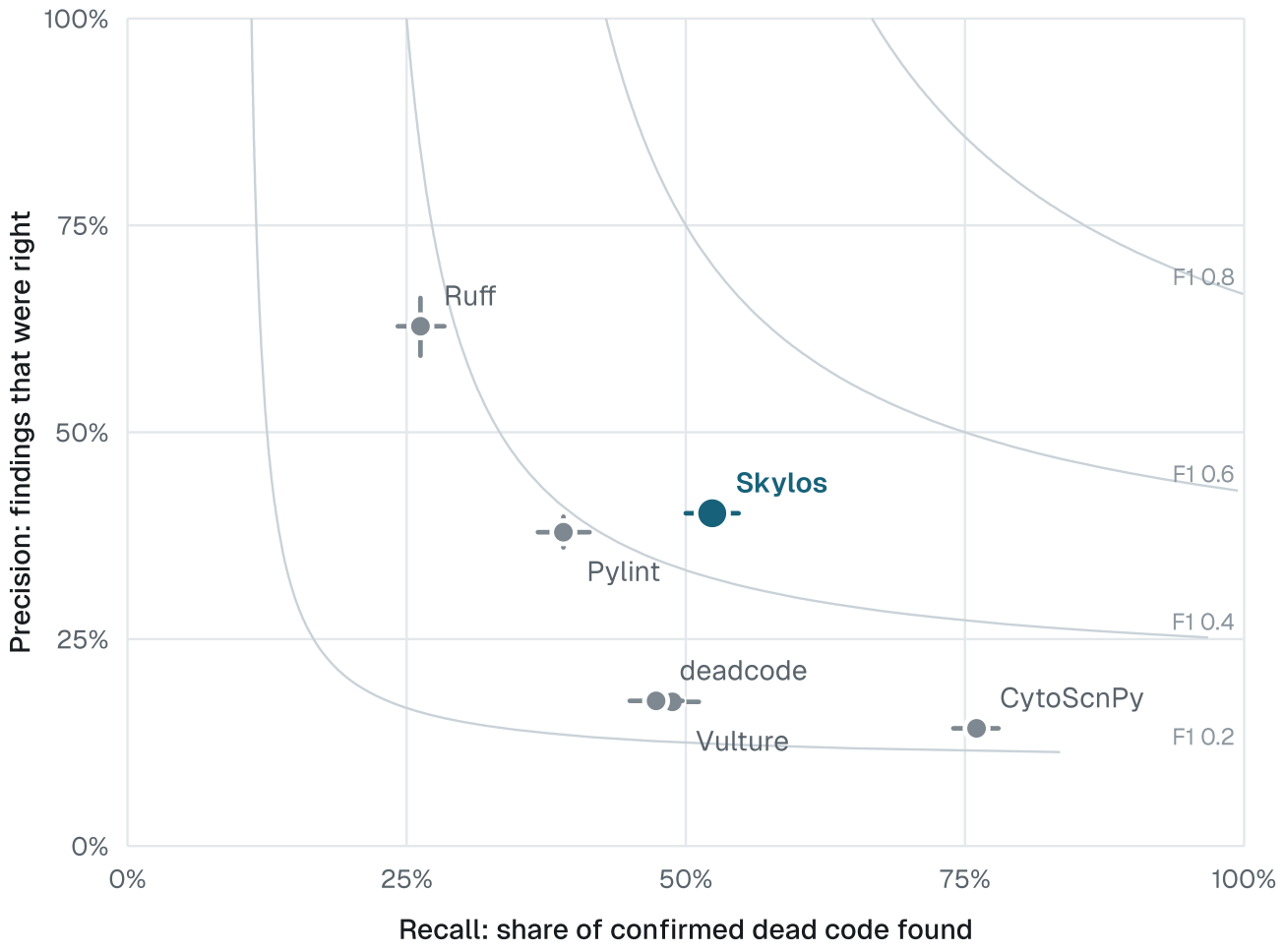
Appendix: full statistics and method

Confidence intervals, significance tests, per-repository results, repositories and method.

A1. Precision, recall and F1 with 95% intervals

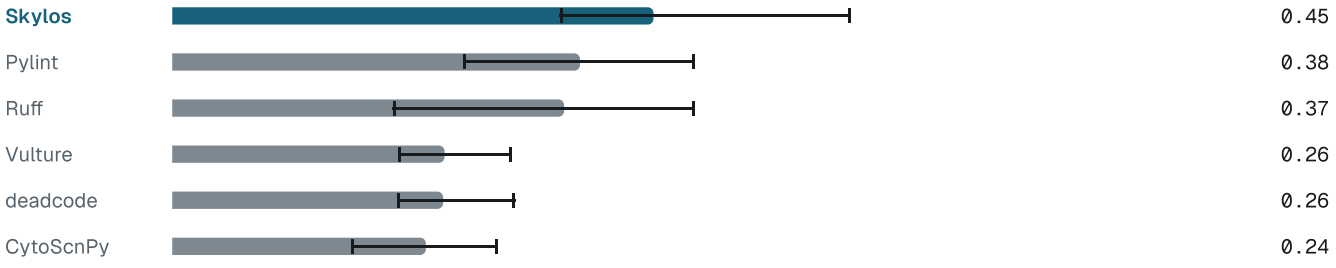
Precision against recall, with 95% intervals

Up and to the right is better. Curves mark equal F1. Hover a tool for exact values.



F1 by tool, with 95% intervals

Intervals resample whole repositories, so one large project cannot dominate.



A2. Every finding, by verdict

Precision is the share of a tool's findings that were right. **Recall** is the share of all confirmed dead code found by any tool that this tool reported. **F1** is their harmonic mean. "Rename only" findings are true, but the fix is renaming a leftover variable to _; they count as right, not as deletable dead code.

Skylos Menu Pricing Design partners Sign in

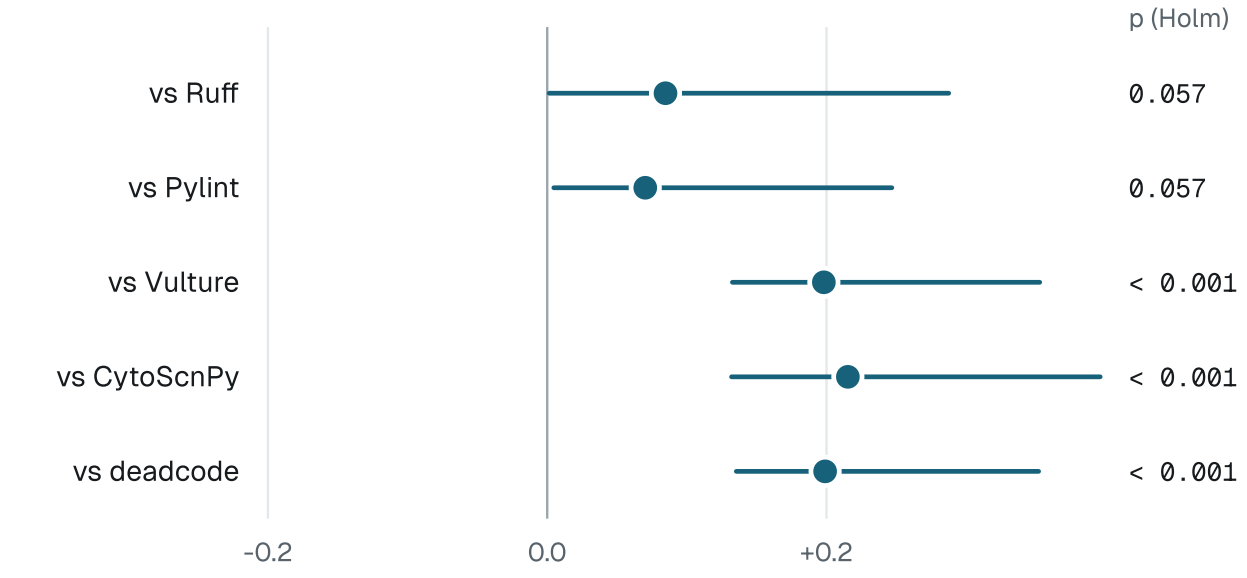
		Dead	Rename	Kept on purpose	Wrong	Precision [95% CI]	Recall [95% CI]	F1 [95% CI]
Skylos 5.3.0	3,036	896	317	1,407	397	40% [38-42%]	52% [50-55%]	0.45 [0.37-0.64]
Pylint 4.1.3	2,608	668	321	862	757	38% [36-40%]	39% [37-41%]	0.38 [0.28-0.49]
Ruff 0.17.3	734	449	12	53	220	63% [59-66%]	26% [24-28%]	0.37 [0.21-0.49]
Vulture 2.16	5,124	835	54	1,936	2,278	17% [16-18%]	49% [46-51%]	0.26 [0.21-0.32]
deadcode 2.4.1	4,948	810	54	1,756	2,307	18% [17-19%]	47% [45-50%]	0.26 [0.21-0.32]
CytoScnPy 1.3.0	10,059	1,301	125	4,715	3,888	14% [14-15%]	76% [74-78%]	0.24 [0.17-0.31]

A3. Significance tests

Each comparison pairs Skylos with one tool on the same findings and the same ground truth. F1 intervals and p-values come from 4,000 bootstrap resamples of whole repositories, which accounts for findings in one project being related.

Skylos minus each tool, with 95% intervals

Right of zero favours Skylos. A teal interval clears zero; gray means the difference is not established. p-values are Holm-adjusted for the 5 comparisons.



Skylos vs	Precision difference · Fisher p	Effect size (Cohen's h)	Found only by Skylos / only by tool · McNemar p	F1 difference · bootstrap p	Repos where Skylos F1 is higher / lower / tied
Ruff	-22.6 pts · < 0.001	-0.46	498 / 51 · < 0.001	+0.085 · 0.057	31 / 3 / 9 · sign test < 0.001
Pylint	+2.3 pts · 0.084	0.05	449 / 221 · < 0.001	+0.070 · 0.057	30 / 10 / 3 · sign test 0.002
Vulture	+22.8 pts · < 0.001	0.51	446 / 385 · 0.037	+0.198 · < 0.001	32 / 7 / 4 · sign test < 0.001

Skylos + Menu Pricing Design partners Sign in

Skylos vs	Precision difference · F	Effect size	Found only by Skylos / only by tool · McNemar p	F1 difference · bootstrap p	Repos where Skylos F1 is higher / lower / tied
CytoScnPy	+26.0 pts · < 0.001	0.60	179 / 584 · < 0.001	+0.215 · < 0.001	32 / 8 / 3 · sign test < 0.001
deadcode	+22.7 pts · < 0.001	0.51	444 / 358 · 0.005	+0.199 · < 0.001	31 / 8 / 4 · sign test < 0.001

p below 0.05 means the difference is very unlikely to be chance. Every p-value is Holm-adjusted: corrected for making several comparisons at once, so one lucky result cannot pass as significant.

A4. Repository by repository

F1 for each tool in each project.

Repository	Skylos	Ruff	Pylint	Vulture	CytoScnPy	deadcode
annotated_deep_learning_paper_implementations blind	0.28	0.21	**0.68**	0.16	0.33	0.16
asgiref blind	**0.67**	–	0.00	0.00	0.00	0.00
asn1crypto blind	**0.44**	–	0.24	0.13	0.03	0.07
asyncpg blind	**0.53**	0.17	0.23	0.40	0.15	0.36
authlib blind	**0.65**	–	0.00	0.22	0.02	0.23
ChatDev blind	**0.59**	0.53	0.54	0.50	0.58	0.51
CloakBrowser blind	0.60	0.47	**0.61**	0.20	0.41	0.25
docker-py blind	**0.54**	0.00	0.22	0.14	0.09	0.13
fish-speech blind	**0.83**	0.70	0.63	0.42	0.52	0.43
google-api-python-client blind	**0.72**	0.53	0.65	0.36	0.23	0.36
gpt-pilot blind	0.33	–	0.19	0.36	0.17	**0.37**
graphrag blind	**0.59**	–	0.05	0.38	0.34	0.40
HanLP blind	0.33	**0.35**	0.34	0.22	0.25	0.21
heretic blind	**0.50**	–	0.00	0.10	0.07	0.11
httplib2 blind	0.76	**0.91**	0.28	0.00	0.11	0.00
khoj blind	**0.50**	–	0.26	0.18	0.36	0.19
langextract blind	0.42	0.11	0.11	**0.71**	0.26	0.70
maigret blind	0.58	0.52	**0.61**	0.32	0.41	0.32
OCRmyPDF blind	0.27	–	0.05	0.25	0.21	**0.29**
PDFMathTranslate blind	**0.76**	0.27	0.08	0.26	0.38	0.26
pre-commit blind	–	–	**0.08**	0.00	0.03	0.00
rapidfuzz blind	0.00	–	0.00	0.17	0.06	**0.20**
Scrapegraph-ai blind	**0.73**	0.48	0.44	0.28	0.46	0.26
searxng blind	**0.38**	0.14	0.09	0.25	0.06	0.26
timesfm blind	0.32	0.48	**0.65**	0.17	0.44	0.17
traitlets blind	0.00	–	**0.13**	0.07	0.00	0.09
attrs	0.00	–	**0.27**	0.00	0.03	0.00
click	**0.74**	–	0.07	0.48	0.07	0.47
fastapi-template	**0.33**	**0.33**	–	0.03	0.05	0.03
flask	**0.67**	–	0.13	0.19	0.08	0.25
httpx	0.00	–	**0.18**	0.00	0.04	0.00
locust	**0.32**	–	0.29	0.19	0.11	0.19
MediaCrawler	**0.80**	0.65	0.65	0.34	0.60	0.33
microblog	**0.67**	0.00	0.00	0.00	0.00	0.00

Skylos + Menu Pricing Design partners Sign in

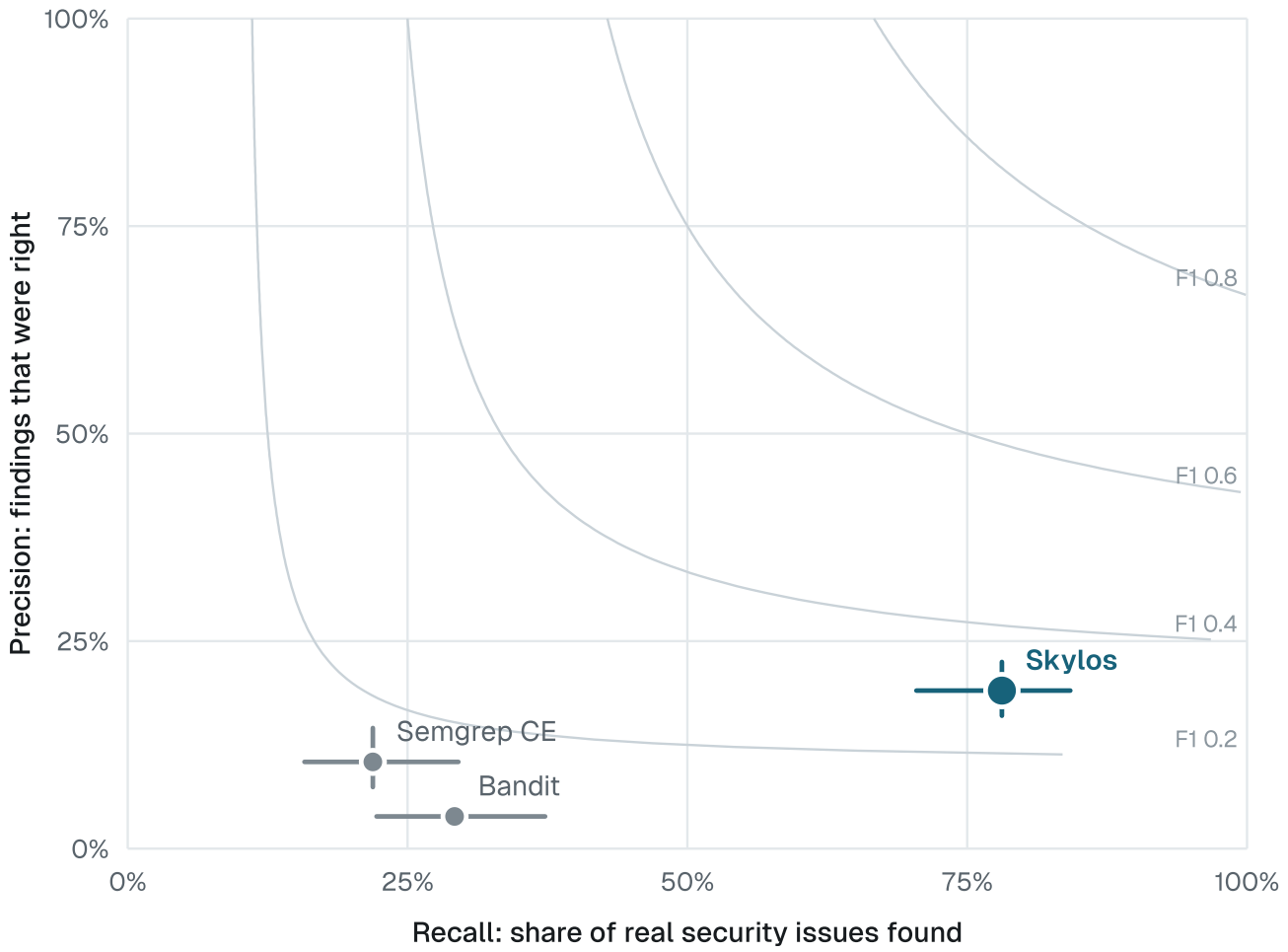
Repository	Skylos	Ruff	Pylint	Vulture	CytoScnPy	deadcode
pydantic	**0.24**	–	0.08	0.11	0.11	0.12
rembg	**0.75**	0.73	0.30	0.33	0.08	0.33
requests	0.00	0.00	**0.18**	0.09	0.07	0.08
rich	**0.27**	0.09	0.00	0.15	0.10	0.14
starlette	0.00	–	0.00	**0.11**	0.00	**0.11**
tqdm	n/a	n/a	n/a	n/a	n/a	n/a
yolov5	0.27	–	0.25	0.24	**0.33**	0.24
you-get	**0.75**	0.71	0.69	0.51	0.50	0.51

Darker is higher; bold marks the best tool in each repository; a dash means the tool reported nothing there; n/a means nothing in that repository was confirmed, so recall is undefined. Select a repository to list its findings.

A1. Precision, recall and F1 with 95% intervals

Precision against recall, with 95% intervals

Up and to the right is better. Curves mark equal F1. Hover a tool for exact values.



F1 by tool, with 95% intervals

Intervals resample whole repositories, so one large project cannot dominate.



A2. Every finding, by verdict

Precision is the share of a tool's findings that were right. **Recall** is the share of all real security issues found by any tool that this tool reported. **F1** is their harmonic mean.

Tool	Findings	Real issue	Noise	Wrong	Precision [95% CI]	Recall [95% CI]	F1 [95% CI]
Skylos 5.3.0	562	107	290	165	19% [16-22%]	78% [70-84%]	0.31 [0.18-0.42]
Semgrep CE 1.183.0	287	30	127	130	10% [7-15%]	22% [16-30%]	0.14 [0.07-0.22]

Skylos
Menu
Pricing
Design partners
Sign in

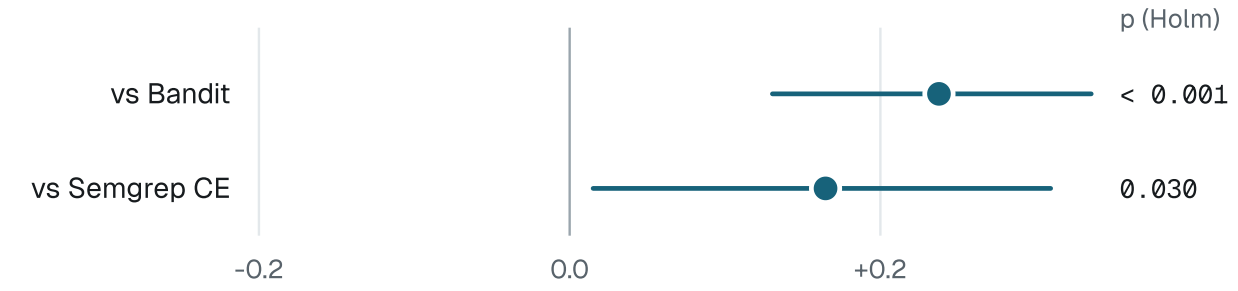
Tool	Findings	Real issue	Noise	Wrong	Precision [95% CI]	Recall [95% CI]	F1 [95% CI]
Bandit [illegible]	1,[illegible]	[illegible]	823	168	4% [3-5%]	29% [22-37%]	0.07 [0.[illegible]

A3. Significance tests

Each comparison pairs Skylos with one tool on the same findings and the same ground truth. F1 intervals and p-values come from 4,000 bootstrap resamples of whole repositories, which accounts for findings in one project being related.

Skylos minus each tool, with 95% intervals

Right of zero favours Skylos. A teal interval clears zero; gray means the difference is not established. p-values are Holm-adjusted for the 2 comparisons.



Skylos vs	Precision difference · Fisher p	Effect size (Cohen's h)	Found only by Skylos / only by tool · McNemar p	F1 difference · bootstrap p	Repos where Skylos F1 is higher / lower / tied
Bandit	+15.2 pts · < 0.001	0.51	87 / 20 · < 0.001	+0.238 · < 0.001	17 / 3 / 22 · sign test 0.005
Semgrep CE	+8.6 pts · 0.001	0.24	87 / 10 · < 0.001	+0.165 · 0.030	13 / 6 / 23 · sign test 0.17

p below 0.05 means the difference is very unlikely to be chance. Every p-value is Holm-adjusted: corrected for making several comparisons at once, so one lucky result cannot pass as significant.

A4. Repository by repository

F1 for each tool in each project.

Repository	Skylos	Bandit	Semgrep CE
annotated_deep_learning_paper_implementations blind	0.60	0.06	0.06
asn1crypto blind	n/a	n/a	n/a
asyncpg blind	n/a	n/a	n/a
authlib blind	n/a	n/a	n/a
ChatDev blind	0.13	0.04	0.00
CloakBrowser blind	n/a	n/a	n/a
docker-py blind	n/a	n/a	n/a
fish-speech blind	0.77	0.00	0.00
google-api-python-client blind	n/a	n/a	n/a
gpt-pilot blind	n/a	n/a	n/a
graphrag blind	0.78	0.21	0.00
HanLP blind	0.11	0.10	0.32
heretic blind	0.00	0.00	0.67
httplib2 blind	0.00	0.33	0.00

Repository	Skylos	Bandit	Semgrep CE
Skylos	0.55	0.23	Sign in
langextract blind	n/a	n/a	n/a
maigret blind	**0.12**	0.09	0.00
OCRmyPDF blind	n/a	n/a	n/a
PDFMathTranslate blind	**0.40**	0.08	0.00
pre-commit blind	n/a	n/a	n/a
rapidfuzz blind	n/a	n/a	n/a
Scrapegraph-ai blind	**0.44**	0.15	0.29
searxng blind	0.06	0.05	**0.15**
timesfm blind	0.05	0.10	**0.25**
traitlets blind	n/a	n/a	n/a
attrs	n/a	n/a	n/a
click	**0.67**	0.00	**0.50**
fastapi-template	–	–	n/a
flask	n/a	n/a	n/a
httpx	n/a	n/a	n/a
locust	0.18	0.00	**0.25**
MediaCrawler	0.00	**0.04**	0.00
microblog	**0.50**	0.00	0.22
OpenManus	**0.88**	0.27	0.22
pydantic	n/a	n/a	n/a
rembg	**0.57**	0.00	0.00
requests	–	n/a	n/a
rich	n/a	n/a	n/a
starlette	–	n/a	n/a
tqdm	n/a	n/a	n/a
yolov5	0.41	0.16	**0.51**
you-get	**0.23**	0.02	0.17

Menu Pricing Design partners

Darker is higher; bold marks the best tool in each repository; a dash means the tool reported nothing there; n/a means nothing in that repository was confirmed, so recall is undefined. Select a repository to list its findings.

Raw data: Dead-code findings (CSV) Security findings (CSV) Statistics (JSON)

A5. The 43 repositories

Each one pinned to the commit we scanned. **Blind** repositories had never been run through Skylos before this benchmark.

Repository	Set	Python lines	Dead-code findings	Confirmed dead code	Security findings	Real issues	Commit
labmlai/annotated_deep_learning_paper_implementations	Blind	39,513	386	68	62	13	2be1800
django/asgiref	Blind	1,987	94	1	0	0	03dcd19
wbond/asn1crypto	Blind	20,031	154	5	28	0	a767d2d
MagicStack/asyncpg	Blind	11,135	167	11	27	0	13b0191

Skylos + Menu Pricing Design partners Sign in

	Set	Python lines	Dead-code findings	Confirmed dead code	Security findings	Real issues	
authlib/authlib	Blind	23,266	466	11	50	0	df5fb3e
OpenBMB/ChatDev	Blind	29,056	234	102	56	2	464a681
CloakHQ/CloakBrowser	Blind	9,772	53	13	50	0	b860957
docker/docker-py	Blind	14,216	222	7	10	0	e734fd5
fishaudio/fish-speech	Blind	10,116	91	37	36	10	0f8a12e
googleapis/google-api-python-client	Blind	13,253	109	25	31	0	c1bdb24
Pythagora-io/gpt-pilot	Blind	18,550	479	53	20	0	9b763fd
microsoft/graphrag	Blind	39,714	409	94	20	9	c600ff8
hankcs/HanLP	Blind	57,016	3,775	445	152	5	281e8ef
p-e-w/heretic	Blind	7,075	39	1	19	1	ecde93f
httplib2/httplib2	Blind	2,405	65	6	5	1	dbbba89
khoj-ai/khoj	Blind	34,760	718	102	103	40	ae229ca
google/langextract	Blind	16,850	73	18	25	0	3493510
soxoj/maigret	Blind	11,598	138	39	35	2	b8747b9
ocrmypdf/OCRmyPDF	Blind	25,407	338	32	49	0	b457e9e
PDFMathTranslate/PDFMathTranslate	Blind	6,266	99	13	32	4	3bdb146
pre-commit/pre-commit	Blind	7,168	149	2	28	0	8cb5845
rapidfuzz/RapidFuzz	Blind	8,867	37	1	25	0	0932842
ScrapeGraphAI/Scrapegraph-ai	Blind	16,623	90	19	55	14	194055e
searxng/searxng	Blind	53,298	1,153	42	117	4	ffa88e3
google-research/timesfm	Blind	16,558	139	36	54	1	e51928e
ipython/traitlets	Blind	8,722	132	2	7	0	a576f24
python-attrs/attrs	Reference	7,483	85	2	8	0	c8bb9bf
pallets/click	Reference	11,136	156	9	24	2	04ef3a6
fastapi/full-stack-fastapi-template	Reference	1,736	73	1	1	0	b1125d8
pallets/flask	Reference	9,499	144	5	9	0	2ac8988
encode/httpx	Reference	8,827	108	1	5	0	b5addb6
locustio/locust	Reference	12,649	149	9	28	2	2054782
NanmiCoder/MediaCrawler	Reference	25,076	440	167	51	1	8f58399

Skylos + Menu Pricing Design partners Sign in

	Set	Python lines	Dead-code findings	Confirmed dead code	Security findings	Real issues	
miguelgrinberg/microblog ↗	Reference	1,843	138	1	15	2	a975ef6
FoundationAgents/OpenManus ↗	Reference	11,429	112	44	16	7	3309bf4
pydantic/pydantic ↗	Reference	59,256	410	19	22	0	08d49ca
danielgatis/rembg ↗	Reference	4,732	130	5	14	2	cd8f366
psf/requests ↗	Reference	5,634	143	2	4	0	ef439eb
Textualize/rich ↗	Reference	39,043	311	20	20	0	9d8f9a3
encode/starlette ↗	Reference	8,488	180	1	4	0	7b2948f
tqdm/tqdm ↗	Reference	4,483	61	0	11	0	75bdb6c
ultralytics/yolov5 ↗	Reference	16,137	157	18	131	11	af08ba1
soimort/you-get ↗	Reference	14,644	635	222	134	4	049548f

A6. How we measured

Projects

The 26 blind repositories were picked by a fixed rule before any tool ran: 16 in two rounds down the most-starred Python repositories on GitHub, in star order, and 10 libraries down the most-downloaded PyPI packages, from rank 201. Every round skipped anything we had cloned, benchmarked or mentioned before, and kept repositories pushed in 2026 with packaging metadata at the root, 5,000 to 60,000 lines of Python and a test folder. The other 17 are reference repositories from our earlier benchmarks.

What counts

Shipped Python code only: not tests, docs or examples. Findings from all tools are pooled, so the same symbol on the same line is judged once, without knowing which tool reported it. Every verdict records its evidence: the line that uses the code, or why nothing does.

- **Dead code**: nothing uses it; a maintainer would accept removing it.
- **Rename only**: an unpacked or loop variable that is never read.
- **Kept on purpose**: unreferenced but deliberately kept, such as public API, compatibility aliases or signatures a framework requires.
- **Wrong**: the code is used. Deleting it would break something.
- **Security**: real issue (worth fixing in this project), noise (true but harmless here) or wrong (the claim is false).

Statistics

- Precision and recall: Wilson score 95% intervals.
- F1 and all differences: 4,000 bootstrap resamples of whole repositories, so related findings in one project are not treated as independent.
- Precision differences: Fisher's exact test, with Cohen's h as effect size.
- Recall differences: exact McNemar test on the same ground truth.
- Repository wins: two-sided sign test.
- All p-values are Holm-adjusted across the comparisons with Skylos in each test.
- Recall is measured against everything any tool found and the review confirmed, the usual pooled ground truth.

Tools and commands

Skylos 5.3.0: `skylos . --json`

Ruff 0.17.3: `ruff check --select F401,F841 .`

Pylint 4.1.3: `pylint --disable=all --enable=unused-import,unused-variable,unused-argument,unused-private-member`

CytoScnPy 1.3.0: `cytoscnpy . --json`

deadcode 2.4.1: `deadcode .`

Bandit 1.9.4: `bandit -r . (B101 assert_used skipped, as most projects do)`

Semgrep CE 1.183.0: `semgrep scan --config p/default (rules in the security category)`

Each tool ran with its default settings on the same checkout, the latest release on the run date. Skylos counts only what it lists by default. Engine: Skylos 5.3.0, commit `f4bff656`.

Skylos was improved on some reference repositories in earlier rounds, so the blind set is the fairer test. Choose **Blind repos** at the top to see it on its own. Looking for the fixture regression checks? Regression evidence

© 2026 Skylos