

garak 0.17.1 in Practice: What NVIDIA’s EU AI Act Mapping Buys You, and What It Does Not

A scanner finishes, prints its summary, and some of those findings now carry EU AI Act risk-category tags. The release note that shipped them says the tags organize results and are not a legal compliance determination.

No, garak’s EU AI Act tags are not compliance evidence. NVIDIA shipped v0.17.0 on 9 September 2026 with reference tags associating probe results with risk categories, then v0.17.1 on 16 October 2026 with the hit@k metric and Python 3.14 support. Two releases in 37 days. The tags sort your findings. The evidence an auditor wants, you still write yourself.

The 0.17 line added risk-category tags and a hit@k field

in report.jsonl

garak is a free, Apache-2.0 red-team scanner you install with `pip install garak`. It fires probe families at a model and scores the responses with detectors. A fresh v0.17.1 install reportedly lists 191 probe classes and 117 detectors, though the counts vary by release and by whether you count classes or runnable configurations. Generators cover OpenAI, Anthropic, Hugging Face, Ollama, Bedrock, Azure, Cohere, Groq, LiteLLM, REST and NeMo Guardrails, so pointing it at whatever you already run is usually a config line rather than an integration project.

The two interesting changes in this window are the risk-category tags and hit@k. hit@k reports whether at least one policy-violating result shows up in the first k attempts. It appears in terminal output and per probe-detector pair in `report.jsonl`. That second part matters more than the metric itself: a per-pair field in a machine-readable file is something you can diff across runs and across model versions.

> **Install and version facts**
> garak is Apache-2.0 and free; the real cost is model-provider API calls and compute. v0.17.0 supported Python 3.11 to 3.13 and dropped 3.10; v0.17.1 added 3.14. If you are pinned to 3.10, this line is not for you.

Does garak's EU AI Act tagging count as compliance evidence?

No, and NVIDIA says so in the release note. The tags associate probe results with risk categories so you can organize output, and NVIDIA explicitly describes them as not a legal compliance determination. The distinction is not pedantic. Article 55 asks providers of GPAI models with systemic risk to conduct and document adversarial testing, report serious incidents and maintain adequate cybersecurity. A scanner produces raw material for the first of those. It does not decide scope, it does not define your threat model, and it does not write the documentation.

THE CLAIM	THE REALITY
Run garak, export the report with its EU AI Act tags, and you have adversarial-testing evidence for Article 55.	NVIDIA's own v0.17.0 note calls the mapping a way to organize results against risk categories and not a legal compliance determination. Article 55 also covers incident reporting and cybersecurity, which no scanner touches.

The timing explains why this keeps getting misread. Article 50 transparency obligations have applied since 2 August 2026, and systems placed on the EU market before that date run out of transition on 2 December 2026. When a deadline is that close and a tool prints the right regulation's vocabulary, the temptation to treat the output as the deliverable is strong. I wrote separately about what Article 50 marking actually requires before 2 December; the short version is that those obligations are about what users see, not what your scanner logs.

Detector scoring was still being patched in the 16 October release

v0.17.1 shipped fixes for mitigation-prefix matching, unexamined-file scoring, skipped multi-turn attempts and safe-token suppression. Each of those is a scoring correctness bug. A mitigation-prefix matching fix means some refusals were being read wrong; a skipped multi-turn attempt means some probes were not completing. That is normal for a tool at this stage, and it is also a reason to date-stamp every report you keep. A finding from a 0.16 run and a finding from 0.17.1 are not the same measurement.

No independent adoption figure and no peer-reviewed benchmark of garak detector accuracy appears in the release material. Detector output varies by probe, by model, by language and by output format. Treat the pass/fail column as a prompt for human review rather than a verdict.

WHAT WORKS

- Breadth: 191 probe classes and 117 detectors in one Apache-2.0 package, no license negotiation.
- hit@k lands per probe-detector pair in report.jsonl, so run-over-run comparison is mechanical.
- Generator coverage from OpenAI and Anthropic to Ollama and raw REST means little glue code.
- Release cadence is live: two releases in 37 days, Python 3.14 within weeks of need.

ROUGH EDGES

- The EU AI Act tags are a classification aid by NVIDIA's own wording, not a determination.
- Detector scoring was still being patched in the 16 October release.
- No peer-reviewed accuracy benchmark; detector behaviour varies by model and language.
- Python 3.10 dropped at v0.17.0, so older pipelines need an upgrade first.

garak leads the 90-day release window on stars, at 9,967

The same window saw Microsoft PyRIT v1.1.0 on 4 September 2026 (MIT), Giskard giskard-agents v1.0.6 on 25 October 2026 (Apache-2.0), UK AISI Inspect AI 0.3.288 on 7 November 2026 (MIT), and Tencent AI-Infra-Guard at v4.6.9 on 5 November 2026 after releases on 10 September, 17 September, 1 October and 29 October.

TOOL	VERSION IN WINDOW	DATE	LICENSE	GITHUB STARS
NVIDIA garak	v0.17.1	16 Oct 2026	Apache-2.0	~9,967
Tencent AI-Infra-Guard	v4.6.9	5 Nov 2026	Apache-2.0	7,447
Giskard giskard-agents	v1.0.6	25 Oct 2026	Apache-2.0	5,947
Microsoft PyRIT	v1.1.0	4 Sep 2026	MIT	4,777
UK AISI Inspect AI	0.3.288	7 Nov 2026	MIT	3,209

Stars measure attention, not detector quality. What they do tell you is which project you can hire for, and which one has enough eyes on it that scoring bugs get reported. On that narrow reading, garak is ahead.

MY TAKE

The EU AI Act tagging is useful for a reason nobody is advertising: it gives a cross-functional

team a shared vocabulary. When a security engineer and a compliance lead look at the same report.jsonl and both see a risk category, the argument shifts from "is this relevant" to "is this acceptable". That is a real saving. I expect scanner output to show up in regulatory submissions anyway, and I expect the first serious challenge to it to turn on exactly the detector-accuracy gap the release notes leave open. Though I could be wrong about which gap gets tested first.

The textGrain watermark has the same missing benchmark

On 12 November 2026, IEEE Spectrum reported that OpenAI's textGrain watermark can be weakened by editing and may fail on texts under 150 words, with no EU performance benchmark for text watermarks. Same shape as garak's tags: a mechanism that looks like it satisfies an obligation, with no agreed measure of whether it does. Tooling moves faster than the benchmarks that would tell you whether the tooling works. My take is that any automated marking or detection layer shipped in the next year will be re-evaluated, so build it to be swapped.

Where I would put garak in a pipeline: as a regression gate in CI, run on every model or prompt change, with hit@k tracked per probe-detector pair and the raw `report.jsonl` archived with the version number. That is defensible. Exporting the EU AI Act tags as your Article 55 adversarial-testing evidence, with no scoping document and no human review of the hits, is not. The gap between those two uses is where the work is, and no release note will close it for you.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a

background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →