



OpenAI Ships EU Text Watermarking, but Only Approved Researchers Get the Detector

OpenAI said on 27 October 2026 it will start watermarking AI-generated text in the EU. The detector goes to approved researchers and expert organisations only, while the Transparency Code asks for public access.

OpenAI announced on 27 October 2026 that it will begin watermarking AI-generated text in the EU, and that its detection tool will initially reach only approved researchers and expert organisations, despite the Transparency Code requiring public access. With the Article 50 marking deadline on 2 December 2026, that gap decides whether machine-readable marking is something a European user can check, or a vendor claim you take on trust.

The same day, the European Commission endorsed the first harmonised standard under the AI Act, covering an AI quality-management system, developed by CEN and CENELEC Joint Technical Committee 21. Compliance with it grants a presumption of conformity for



the relevant obligations. The standard, EN 18286:2026, addresses quality-management requirements for high-risk AI systems, per [European Scientist](#). The high-risk obligations themselves have been postponed until at least December 2027.

So one branch of the regime got a process standard with a two-year runway. The other branch, transparency, got a watermark whose reader is gated. That asymmetry is the story.

! The catch

A marking scheme that only a closed list of parties can read shifts verification from the user to the vendor. If you cannot independently check the mark, the mark tells you nothing about the specific document in front of you. It tells you what the provider says about its own output.

Two further pieces of context from the same [CDT Europe bulletin](#). The Council of the EU agreed its negotiating position on the Digital Omnibus on 21 October 2026, after member-state negotiations. And the AI Office has requested serious-incident reports and information on companies' security practices following 2026 loss-of-control cases, per [Connic's enforcement summary](#), which places those duties under Article 55 for systemic-risk general-purpose providers and Article 73 for certain high-risk systems. EVP Henna Virkkunen and the AI Scientific Panel held a special meeting on 9 October 2026 on those incidents and on model monitorability.

Nothing in the reporting says whether a researcher-gated detector satisfies Article 50

I want to be precise about what is unconfirmed. The announcement covers watermarking of generated text in the EU. The detector access policy is described as initially limited, which leaves room for it to widen. Whether a regulator reads a researcher-gated detector as adequate machine-readable marking is not something any source here settles.

My reading of the design constraint: text watermarking is weaker than image provenance in a way that makes gating rational for the provider. A detector you publish is a detector adversaries can query until they learn how to strip the signal. That is a genuine engineering trade-off rather than bad faith. It is also the trade-off I would expect to lose in Brussels, because the Transparency Code was written for verification by the public rather than for a



research programme.

WHERE I LAND

My take: the gated detector is defensible as security engineering and indefensible as transparency compliance. If I were advising OpenAI, I would publish a rate-limited, logged public verification endpoint, accept the robustness loss, and argue the detector is one signal among several rather than proof. If I were advising a European buyer, I would treat the watermark as unverifiable until a third party outside the approved list demonstrates a read.

This is the same failure mode I wrote about with [C2PA content credentials, where five of six platforms strip the metadata](#). A marking standard survives only if the part that reads the mark is as widely deployed as the part that writes it. Watermarked text that travels through a copy-paste, a CMS and a PDF export has the same fragility problem, and now it also has a reader problem.

Your Article 50 obligation does not depend on OpenAI's detector policy

If you ship AI-generated text into the EU, you are the provider or deployer of your own output. Three concrete moves.

First, do not treat an upstream vendor's watermark as your compliance artefact. Keep your own marking and disclosure at the point where text reaches a user, and document it. I set out the mechanics in my [guide to shipping Article 50 marking before 2 December 2026](#).

Second, write the incident path now rather than after an event. The AI Office has already asked for serious-incident reports and security-practice information this year. Article 55 and Article 73 duties need a named owner, a clock and a log, which is cheap to build in advance and expensive to improvise.

Third, read the harmonised standard even if you are not high-risk. Presumption of conformity under EN 18286:2026 is the first time the AI Act offers a concrete, auditable target instead of a principle. The December 2027 postponement of high-risk obligations gives you a long runway to align quality-management processes without a deadline



overhead.

Anthropic's US proposal aims at catastrophic harm, Article 50 aims at consumer information

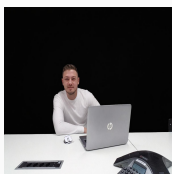
The regulatory picture outside the EU is pulling in a different direction, which matters if you build across jurisdictions. Anthropic published a US federal safety proposal on 17 October 2026, updated 28 October, that would let a federal court pause a frontier model at the Commerce Department's request, with thresholds of imminent risk of death or injury to more than 50 people, or more than \$1 billion in economic damage, via bioweapon, cyberattack or loss of developer control, as reported by the [Brussels Times](#). The two regimes will not harmonise, and a product sold in both markets will carry both sets of machinery. Switzerland, meanwhile, still had no overarching AI-specific legislation as of 23 October 2026, with a consultation draft expected by end of 2026, according to the [Legal 500 country guide](#). Swiss-domiciled companies selling into the EU inherit the EU rules anyway.

I expect the detector access question to be settled by pressure rather than by rule. My prediction for the next six to twelve months: either OpenAI widens access to something resembling a public endpoint, or the Commission names researcher-only detection as insufficient in a transparency enforcement context. I do not think the status quo holds through 2027, partly because the Digital Omnibus negotiation is still live and transparency is one of the easier things to tighten when high-risk duties have already slipped a year.

One thing I am genuinely unsure about: whether robust text watermarking is technically achievable at all at the strength the Code implies. Nothing in the available reporting tells us how OpenAI's scheme performs against paraphrasing or translation, and that silence is worth noticing. If you are making a procurement decision that leans on it, ask the vendor for the false-positive rate and the survival rate after a round-trip translation, in writing. If that conversation would be useful to have with someone outside your vendor relationship, [I am happy to talk it through](#).



OpenAI Ships EU Text Watermarking, but Only Approved Researchers Get the Detector



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 →