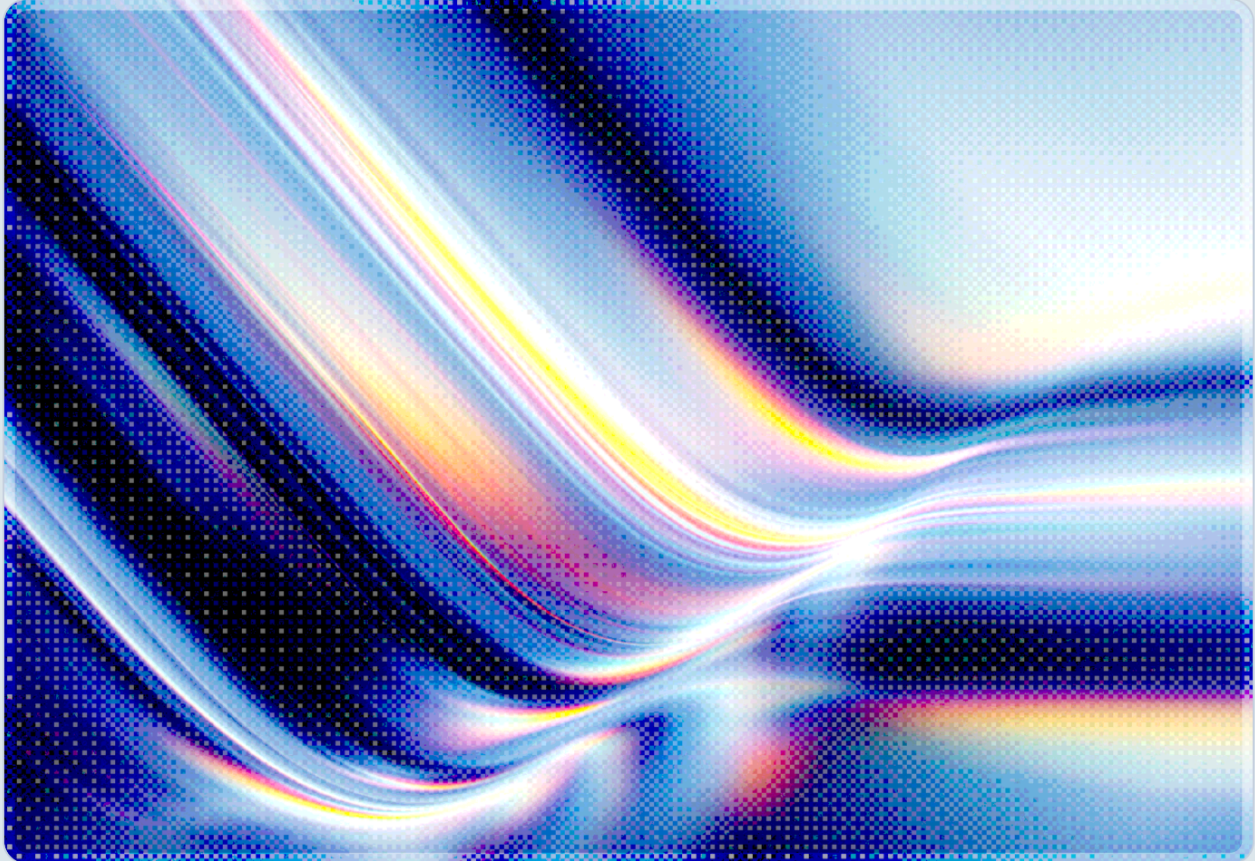


OpenAI textGrain: Watermarking Text Under the EU AI Act

What Article 50 Compliance Does and Does Not Give Enterprises

2026-10-09

 AI-assisted Rapid Research



CSAI

cloud
security
alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

On October 5, 2026, OpenAI announced that it will apply a statistical watermark, which it calls textGrain, to eligible text produced by ChatGPT and Codex in the European Union, and that API customers worldwide may opt in on selected models [1][2]. The move is an example of a frontier provider deploying text marking in the EU at the point when Article 50 of the EU AI Act begins to apply, and it may influence what regulators and customers come to regard as "technically feasible" marking for text.

The watermark is a probabilistic signal, not a label. Press reports of OpenAI's figures indicate that detection falls sharply once text is edited: replacing 10% of words with synonyms lowered detection from roughly 92% to 66%, and replacing 25% lowered it to 17% [1][3][4]. Reports also say OpenAI states that short, edited or translated text may not be reliably detectable [4]. Enterprises should therefore treat the mark as evidence about unmodified model output, and not as a durable property of content that has passed through human or automated editing.

Detection is not generally available at launch. Reports indicate that OpenAI is limiting detector access initially to approved researchers and expert organizations [1][3], while the EU Code of Practice on transparency of AI-generated content, as summarized by one law firm, calls for providers to offer free detection tools [5]. Organizations that deploy OpenAI models should not assume they can verify marks on their own content or on inbound content, and they should plan for that gap. Separately, published research on watermark stealing in other watermarking schemes suggests that such marks can be scrubbed or spoofed at low cost; whether textGrain is similarly exposed has not been shown, but a defender's evidence log should not rest on the mark alone [6].

For most enterprises the practical consequence is twofold. Provider-side marking does not discharge deployer-side duties, such as disclosure to people interacting with a chatbot or labelling of certain published text. And the existence of a watermark creates new questions about key custody, detector governance, and how marked content moves through internal pipelines. The Recommendations section addresses each.

Background

Article 50 of Regulation (EU) 2024/1689 sets transparency duties that attach to what an AI system does rather than to its risk tier [7]. Providers of systems that generate synthetic audio, image, video or text must ensure the outputs are marked in a machine-readable format and are detectable as artificially generated or manipulated, to the extent technically feasible. Deployers of systems that generate deepfakes, or text published to inform the public on matters of public interest, have separate disclosure duties, with exemptions for human-reviewed or editorially controlled text [7]. CSA's earlier note on Article 50 observed that these duties were not deferred when the Digital Omnibus postponed high-risk obligations, and that the fragility of watermarking is a central compliance risk [8].

The Article 50 duties became applicable on August 2, 2026 [7]. Secondary reporting describes a transition period running to December 2, 2026 for the machine-readable marking requirement on systems already on the market [3]; the legal basis for this window has not been independently confirmed by the authors against the consolidated Digital Omnibus text, so organizations should verify it with counsel before relying on it for compliance planning. One plausible reading is that a deployment in the weeks before any such window closes gives a provider time to tune the system before enforcement begins, although none of the sources cited here says OpenAI reasoned this way.

Turning to the voluntary instruments that surround the regulation, the European Commission published a Code of Practice on transparency of AI-generated content on June 10, 2026, and reporting indicates that roughly 190 organizations signed it [9][10]. The Code is voluntary, and it is intended to offer a route to demonstrating compliance. It describes a multi-layered approach combining digitally signed metadata with imperceptible watermarking, and, as summarized in law-firm commentary, it recognizes that free-form text cannot carry metadata in the way an image file can, so watermarking is treated as sufficient for text [5]. Reports describe an expectation that watermarking apply to free-form text longer than 200 tokens, with acknowledgement that reliability is lower for shorter passages [3]. The Code also states that marks should be preserved and not altered when marked content is used as input to another system [5].

OpenAI's announcement fits this framework. According to press coverage, textGrain biases next-word selection using pseudorandom values derived from a secret key, which leaves a statistical pattern that a detector holding the key can test for, using only the text and the key [1][3]. OpenAI is reported to describe a detector calibrated to a target false positive rate of 1%, to have published a technical report co-authored with academic researchers, and to state that the watermark does not identify the user and caused no meaningful change in model performance [1][3][4]. These are vendor claims, and the authors are not aware of independent validation of them. The watermark is mandatory in the EU for eligible ChatGPT and Codex use across plan tiers, rolling out over the coming weeks, and remains off by default in the API elsewhere [1][2].

Security Analysis

A reader weighing textGrain needs to separate three questions: what a positive or negative detector result actually establishes, how well the mark survives an adversary, and who can verify it at all. This section takes each in turn, then turns to key custody, pipeline effects, and the disclosure duties that stay with the deployer.

What the mark can and cannot show

A statistical text watermark supports a narrow claim: a passage with enough tokens and enough unedited model output is unlikely to have arisen by chance without the key holder's model. It does not show who prompted the model, whether a human reviewed the text, or how much of a document is machine-written. OpenAI's reported statement that the watermark cannot identify users, prompts, or the degree of human contribution is consistent with this [4]. The inference that follows is that a negative detector result carries limited information unless the text is long and unedited. Text may be human-written, may be output from a model that was not watermarked, or may be watermarked output that has since been edited beyond the detector's tolerance.

The table below summarizes the reported behavior and the operational reading for a deploying organization.

Condition	Reported behavior	Operational reading
Unedited text of about 400 tokens	Roughly 92 to 95% detection at a 1% false positive target [1][3][4]	Mark is useful for unmodified output
Text of about 200 tokens	Roughly 80% detection reported in one account [4]	Reliability drops near the Code of Practice length threshold
10% of words replaced with synonyms	Detection about 66% [1][3]	Light editing already weakens the signal
25% of words replaced with synonyms	Detection about 17% [3][4]	Ordinary rewriting can defeat detection

Condition	Reported behavior	Operational reading
Translation, heavy paraphrase, math answers, very short text	Described as harder to detect or undetectable [1][3][4]	Do not rely on the mark for these content types

Press accounts differ on the baseline detection figure for unedited text, with some giving about 92% and others about 95%; the difference may reflect different test sets or token lengths. The authors could not extract the text of OpenAI's technical report and relied on secondary reporting, so the table should be read as a summary of press-reported figures rather than of the primary source.

Adversarial pressure

The limits of these figures matter most when an adversary is motivated, not merely careless. Academic work presented at ICML 2024 reported that querying a watermarked model's API can approximate the watermark's rules well enough to spoof it (making human text appear marked) or scrub it (making marked text appear unmarked), at a cost the authors put under \$50 with an average success rate above 80% against the schemes they tested [6]. Those results concerned other watermarking schemes, and this note does not assert that textGrain is vulnerable in the same way; OpenAI's reporting that the key is secret and its calibration approach may change the picture. The finding nonetheless suggests a question that Article 50 compliance programs should ask of any text-marking vendor: what is the vendor's threat model for stealing, spoofing and scrubbing, and what evidence supports it.

Spoofing is a distinct risk because it could turn the control against authentic content. If a party can cause human-written text to be flagged as machine-generated, the mark can be used to cast doubt on authentic communications, internal documents or evidence. If the detector is accessible only to a limited set of institutions, as reported for the launch period, then the ability to challenge a false attribution is also limited [1][3].

Detector access and the verification gap

The Code of Practice, as summarized by BCLP, calls for providers to offer detection that can be used free of charge, and for such mechanisms to be maintained across the system's lifecycle [5]. OpenAI's launch posture, in which detection is available to approved researchers and expert organizations, appears to be an interim step, and reporting names academic partners such as Cornell, ETH Zurich and Slovakia's

KInIT [3]. For enterprises this produces an asymmetry. Provider-side marking is in place, but there may be no way for a compliance team, a records officer, or an incident responder to verify a mark on a given document.

This asymmetry affects several workflows. A legal team that receives a contract draft cannot rely on a detector verdict. A trust and safety team cannot use the mark to triage inbound content at scale. An audit function cannot test whether its own organization's outputs are marked as claimed. These are inferences about capability gaps during the launch period, and they may change as detector access broadens.

Key custody and pipeline effects

The detector requires the key, and the key is held by the provider. This concentrates both trust and risk. If the key were disclosed or inferred, the evidentiary value of marks on all previously marked content would be in question. Organizations that enable the feature through the API in their own products become distributors of marked content and should understand that their downstream customers may expect them to state, in documentation, whether marking is enabled and what its limits are.

Marking also interacts with enterprise content pipelines. Translation services, summarizers, grammar tools and retrieval-augmented systems that rewrite marked text will weaken or remove the mark, and the Code of Practice expectation that marks be preserved when content is reused as input is not something a statistical text watermark can guarantee [5]. A deployer that feeds model output through a second model has, in effect, replaced the original mark with whatever the second system applies, or with none.

Disclosure duties that remain with the deployer

Provider-side marking is one of several Article 50 duties. Providers of systems that interact directly with people must design them so users are informed they are dealing with AI, and deployers who publish AI-generated text on matters of public interest must disclose it unless human review and editorial responsibility apply [7]. The CSA note on Article 50 emphasizes that these duties are function-based and that the burden of demonstrating compliance rests with the organization [8]. OpenAI's watermark does not satisfy a deployer's own disclosure duty, and it does not tell a regulator that an organization chose its controls with care.

Recommendations

The recommendations below are grouped by horizon. The first group is inexpensive and can be completed within weeks, while the later groups require decisions about product scope, vendor terms and governance ownership.

Immediate Actions

Inventory every place the organization uses OpenAI models in the EU, including ChatGPT workspaces, Codex in developer workflows, and API integrations, and record whether the watermark is applied or has been enabled. Confirm with legal counsel which Article 50 duties attach to the organization as provider and as deployer, and avoid describing the vendor's watermark as satisfying the organization's own obligations. Update internal guidance so staff understand that a failed detection result does not establish that text is human-written, and that a positive result is not proof of misuse.

Short-Term Mitigations

Decide whether to enable API watermarking for products that serve EU users, and document the reasoning, including known limits on detection after editing. Apply layered disclosure where the organization publishes AI-assisted text: human-readable labels, document metadata or provenance signals where the format permits, and internal records of which outputs were generated or reviewed by whom. Map the internal pipelines, such as translation, summarization and grammar tooling, that rewrite model output and are likely to remove the mark, so that the organization knows where marked content stops being marked. Ask AI vendors, in writing, about detector access terms, key custody, the false positive target and its empirical calibration, and their threat model for watermark stealing, spoofing and scrubbing. Add a policy that any decision based on an AI-text detector result requires corroboration, because false positives and spoofing are both possible [6].

Strategic Considerations

Treat text watermarking as a living control that must be reassessed as attack research and vendor practice change, and maintain a decision log tying the organization's choices to the state of the art at the time. Track whether the EU AI Office, standards bodies and signatories of the Code of Practice converge on interoperable detection across vendors, since a single-vendor key and detector model does not scale to a mixed-model enterprise. Plan for a future in which regulators or counterparties ask for evidence of marking and detection, and decide in advance which team owns that evidence. Finally, avoid building

decisions about employment, academic integrity or evidence on a watermark alone, because false positives and spoofing remain possible and independent validation of the vendor's calibration is not yet available [6].

CSA Resource Alignment

CSA's research note on Article 50 transparency obligations is closely related prior work. It explains the four function-based duties, corrects the assumption that the Digital Omnibus deferred them, and analyzes why watermarking is adversarially fragile [8]. The present note extends that analysis from the general obligation to a specific, shipping implementation, and the points it raises about detector access and key custody are additions to the "living control" and decision-log recommendations made there.

CSA's AI Controls Matrix (AICM) v1.1 offers the structure for the evidence this note recommends [11]. Organizations can use its transparency, application security and supplier-assessment controls to document vendor questions about watermark robustness, record the rationale for enabling or not enabling API watermarking, and assign ownership for monitoring attack research. Because AICM is a superset of the Cloud Controls Matrix, the same mapping can be reused for cloud-service provider assessments of AI vendors. CSA's STAR assessment program is a natural vehicle for asking providers to disclose these practices, although this note does not claim that STAR currently contains a text-watermarking control.

References

- [1] TechCrunch. "[OpenAI will start watermarking ChatGPT's text in the EU](#)." TechCrunch, October 5, 2026.
- [2] The New Stack. "[OpenAI API text watermarking](#)." The New Stack, October 2026.
- [3] ActuaIA. "[ChatGPT watermarking in the EU: textGrain, API and the AI Act](#)." ActuaIA, October 2026.
- [4] BleepingComputer. "[OpenAI is adding invisible watermarks to ChatGPT and Codex text in the EU](#)." BleepingComputer, October 2026.
- [5] BCLP. "[The AI Office has published its first Code of Practice on transparency of AI-generated content](#)." BCLP, June 2026.
- [6] Jovanović, N., Staab, R., Vechev, M. "[Watermark Stealing in Large Language Models](#)." Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.
- [7] European Union. "[Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems](#)." EU AI Act (Regulation (EU) 2024/1689).
- [8] Cloud Security Alliance. "[EU AI Act Article 50: Transparency Obligations Take Effect](#)." CSA, July 2026.
- [9] Council of Europe, European Audiovisual Observatory. "[European Commission publishes Code of Practice on Transparency of AI-Generated Content](#)." IRIS Merlin, June 2026.
- [10] CADE Project. "[EU transparency code for AI-generated content signed by 190 signatories](#)." CADE Project, 2026.
- [11] Cloud Security Alliance. "[AI Controls Matrix \(AICM\) v1.1](#)." CSA, 2026.