
Reached by Recital¹

*An enforcement pack for a pre-market research model under the EU AI Act,
and the three facts it turns on*

Bradley Quinlan

University of Essex

Track 3, Regulatory Response · Apart Research AI Incident Response Sprint, 11–13
September 2026

With

Apart Research

Abstract

In July 2026 a frontier developer’s models escaped an evaluation sandbox and reached a third party’s production infrastructure. Whether the episode was reported as the EU AI Act requires assumes a prior question: every account calls the model internal research, which Article 3, point (63) excludes from the definition of a general-purpose AI model. On the Commission’s reading the obligations attach, but by a route through Recital 97 and non-binding guidance rather than the enacting definition, and that route needs three facts the public record does not supply — one of which the provider’s own account denies. This report delivers the documents rather than an analysis of them: a model Article 91 request that asks the three facts, an issuing note and answer matrix saying what each answer does, the serious-incident template rebuilt with the two fields it lacks, and three amendments in current-text/amended-text form. A second finding needs none of that reading and is fixable this week: the Commission’s serious-incident template never asks when the provider became aware, though every initial-report period in the Code of Practice runs from awareness. **The form cannot evidence its own deadlines**, and the cause is traceable — it transcribes the Code’s list of contents, and that list omits the date too. Every document is derived only from text held and verified in the accompanying repository, whose checkers and corrections log are the method, and whose verification

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

layer runs unchanged on the next incident.

1. Introduction

Two episodes sit behind this report. Models belonging to a frontier developer left an evaluation sandbox in July 2026 and reached a third party’s production infrastructure (Hugging Face, 2026); between May and June the same developer’s agents had made roughly 17,000 edits to a community wiki (Von Arx *et al.*, 2026). The first was handled under “a traditional security incident response playbook”; the second was treated as “an instance of misalignment” and acknowledged publicly on 5 September 2026, the day after independent researchers published (OpenAI, 2026c).

Commentary since has asked whether these episodes were reported as the AI Act requires. That question assumes an answer to a prior one. The model that drove the compromise is described by its own developer as an “internal-only research model”, “not intended for public release” (OpenAI, 2026b), and by the independent investigators as “a research model which was not intended for production” (METR and Redwood Research, 2026). Article 3, point (63) excludes from the definition of ‘general-purpose AI model’ those “used for research, development or prototyping activities before they are placed on the market”, and Article 2(8) excludes pre-market research, testing and development activity from the Regulation’s application. Whether the obligations attach at all is therefore the threshold question, and it is not the one the commentary has been asking.

It matters because the answer decides who acts: if the obligations attach, the AI Office has a compulsory information power it is already using; if they do not, no enforcement reaches the conduct and the response is legislative. The threat model is narrow and recurrent — a model never released, evaluated internally, that escapes containment and causes harm outside the developer — and it is precisely the configuration the research exclusion was drafted to leave alone.

Our main contributions are:

- **A scope finding, and the route as the finding.** On the Commission’s reading the obligations attach to a pre-market research model, but through Recital 97 and guidance that disclaims its own authority rather than through the enacting definition, and the route needs three facts — a placement event, a Union nexus, and systemic risk. The public record supplies none of them and the provider’s own account denies one (§4.1).
- **An instrumentation finding that needs none of that reading.** The Commission’s serious-incident template for general-purpose AI models never asks when the provider became aware, though every initial-report period in the Code of Practice runs from awareness. The cause is traceable: fields 1 to 9 transcribe Measure 9.2 item for item, and Measure 9.2 has no awareness item either. The Commission has already drafted the field — for the high-risk regime (§4.2).
- **An enforcement pack rather than an analysis of one, and an apparatus that produces the next one.** Seven documents in the form their addressee would issue or adopt, including a model Article 91 request in Article 91(4)’s required form and three amendments in current-text/amended-text form (§4.3). Behind them a verification layer — a statutory register of forty-seven verified provisions, an inference register, six checkers, a corrections log of thirty-one entries — and six scripts that stand the whole of it up on a new incident. The claim that it transfers is not an aspiration: a second incident was run end to end, on a different provider, in a single session (§4.4).

1.1 *Using it: from a new incident to an issued request*

The repository is written to be picked up rather than read. Five steps, each a command or a rule, take a fresh episode to a request a regulator could sign:

- **Stand up the repository.** `python3 tools/new-incident.py` PATH creates the working directory with the schemas, the six checkers and the verified statutory register already in it. Nothing statutory has to be looked up twice.
- **Hold the record before characterising it.** Archive each source, print the archived copy to `_sources/`, and let `python3 tools/hold-source.py --draft` write the register entry with its SHA-256. `validate.py` then refuses any fact whose source is not held.
- **Fill the cross-walk.** One row per obligation, the provider’s facts in a different column from the provider’s characterisation of them, and every verdict in one of the three permitted forms. This is the part no tool does, and should not.
- **Render the instruments.** `draft-request.py` drafts an Article 91 request from the rows still unresolved; `formfill.py` renders the regulator’s own template filled from the record, each cell naming its source; `render-pack.py` renders the pack to DOCX for editing.
- **Attack it before anyone else does.** `python3 tools/check.py` must come back green on all six, and `python3 tools/infercheck.py --attack` prints the load-bearing steps weakest first, which is the list to argue with.

2. Related Work

The closest prior work is Pistillo (2026), a chapter for the *Cambridge Commentary on EU General-Purpose AI Law* arguing that internal deployment falls within the Act. This report agrees with that conclusion for internal deployment and shares one of his two pathways — the extensive reading of ‘making available’ under Article 3, points (9) and (10). It differs in subject and therefore in result: his model is deployed internally, where this record describes a model never deployed at all, whose only recorded system is the evaluation harness. That is the difference the three facts exist to close. The Commission’s own instruments are the primary counterparty rather than the literature: the Guidelines of 19 November 2025 (C(2025) 7719 final; European Commission, 2025), the general-purpose AI Q&A (European Commission, 2026a) and Opinion C(2025) 5361 final (European Commission, 2026c), each held and read for what it does *not* say as carefully as for what it does. Incident commentary asks what was disclosed and whether it was timely (Cloud Security Alliance, 2026; Uuk, 2026), while the independent investigation of the episode (METR and Redwood Research, 2026) was access-scoped by its subject. None asks whether the obligations reached the model at all.

Three things this work adds sit alongside that literature rather than inside it. It supplies the documents rather than the argument: a request in the form a regulator would issue, and amendments in the form a legislator would table. It reads the Commission’s own reporting form against the Code’s own clock — the finding at § 4.2 came from filling the form in, not from reading about it. And its verification layer is packaged to be re-run rather than written once: the Article 91 request is roughly 60% boilerplate that survives any incident, and the second run at § 4.4 is the evidence for the rest.

3. Methods

The ceiling, stated in advance. The method answers one question — does the discloser’s own account, taken on its own terms, resolve this obligation? — and it cannot answer what happened. Every verdict takes one of three permitted forms: *on the provider’s own account, this appears met; appears unmet; or the disclosed record does not resolve this*. `tools/validate.py` rejects any cell that drifts out of that register. An analysis that cannot overclaim is one a regulator can read without auditing it first.

Four rules make the reasoning inspectable, each stated in full at Appendix D. Sources are held before they are characterised — every source a fact rests on is archived to `_sources/` with its SHA-256, and forty-seven provisions, seven imported provisions of the CER Directive (Directive (EU) 2022/2557) and four recitals are verified line by line against the consolidated Regulation and the authentic Official Journal text, five CJEU

judgments on EUR-Lex, each carrying the date of its check. Evidence is tiered by relationship to the claim rather than prestige. Absences are bounded by a date and a named corpus. And inferences are enumerated: `data/inferences.yaml` states each argumentative step as a premise set, names what would defeat it, and ranks it.

Six checkers run from one entry point. `tools/check.py` runs `hold-source.py` (no fact may rest on a source nobody holds), `validate.py` (verdict discipline), `citecheck.py` (citations resolve to verified text), `quotecheck.py` (quotations found in the held document), `infercheck.py` (premises, defeaters, acyclicity) and `housestyle.py` (conventions, repeated paragraphs, drifted counts, dangling paths). What is automated is checking, not judgement: they will say a citation does not resolve or a count has drifted; they will never say an inference is wrong. Four adversarial passes were run besides — two over a finished draft by a model with no sight of the reasoning that produced it, one over the instrument, one over the whole repository with the held sources open. Every reviewer was a language model, and § 5 says what follows. Whether the method is usable by anybody else is an empirical question, so it was run as one: § 4.4.

4. Results

4.1 *The route attaches, and needs three facts*

The enacting text points away from scope. Article 2(1), point (a) applies the Regulation to providers “placing on the market **or putting into service** AI systems **or placing on the market** general-purpose AI models”. The halves differ: ‘putting into service’ at Article 3, point (11) expressly includes supply “for **own use** in the Union for its intended purpose”, and models have no equivalent limb. Article 3, point (63) then carries no “sole purpose” qualifier and no conditions, though Article 2(6) shows the formula was available. On the enacting text alone, an internal unreleased research model sits outside Chapter V.

Recital 97 moves the moment of placement. A model integrated into its provider’s own system put into service “should be **considered to be placed on the market**” — so the development was never activity “prior to” placement, and Article 2(8) never engages. The Commission reproduces the rule at paragraph 54 of its Guidelines, and its Q&A consolidates it into three cumulative conditions, the third decisive: the exception is unavailable where “the model is not a general-purpose AI model with systemic risk”. Paragraph 22 treats a model’s lifecycle as beginning “at the start of the large pre-training run”, so the exclusion protects a model never placed on the market, not the development phase of one that is.

That route needs three facts, and the record supplies none of them (Figure 1). A *placement event*: every version of the Commission’s reading runs from one, and the provider’s account says the model was among “research models that were not intended for public release”, with its weights quarantined after the incident. A *Union nexus*: Article 2(1), point (a) reaches placing on the market “in the Union”, and no account places any own system incorporating this model in the Union. *Systemic risk*: Article 51(2) presumes it above 10^{25} FLOP, and no public source gives the figure. Each absence is bounded: the placement and Union questions as of 13 September 2026, across the six incident-record documents held here, searched for “deploy”, “internal”, “Union” and “Europe”; the compute figure as of 12 September 2026, across those documents and English-language web search on the provider, “Article 52”, “notification” and “AI Office” (Appendix C).

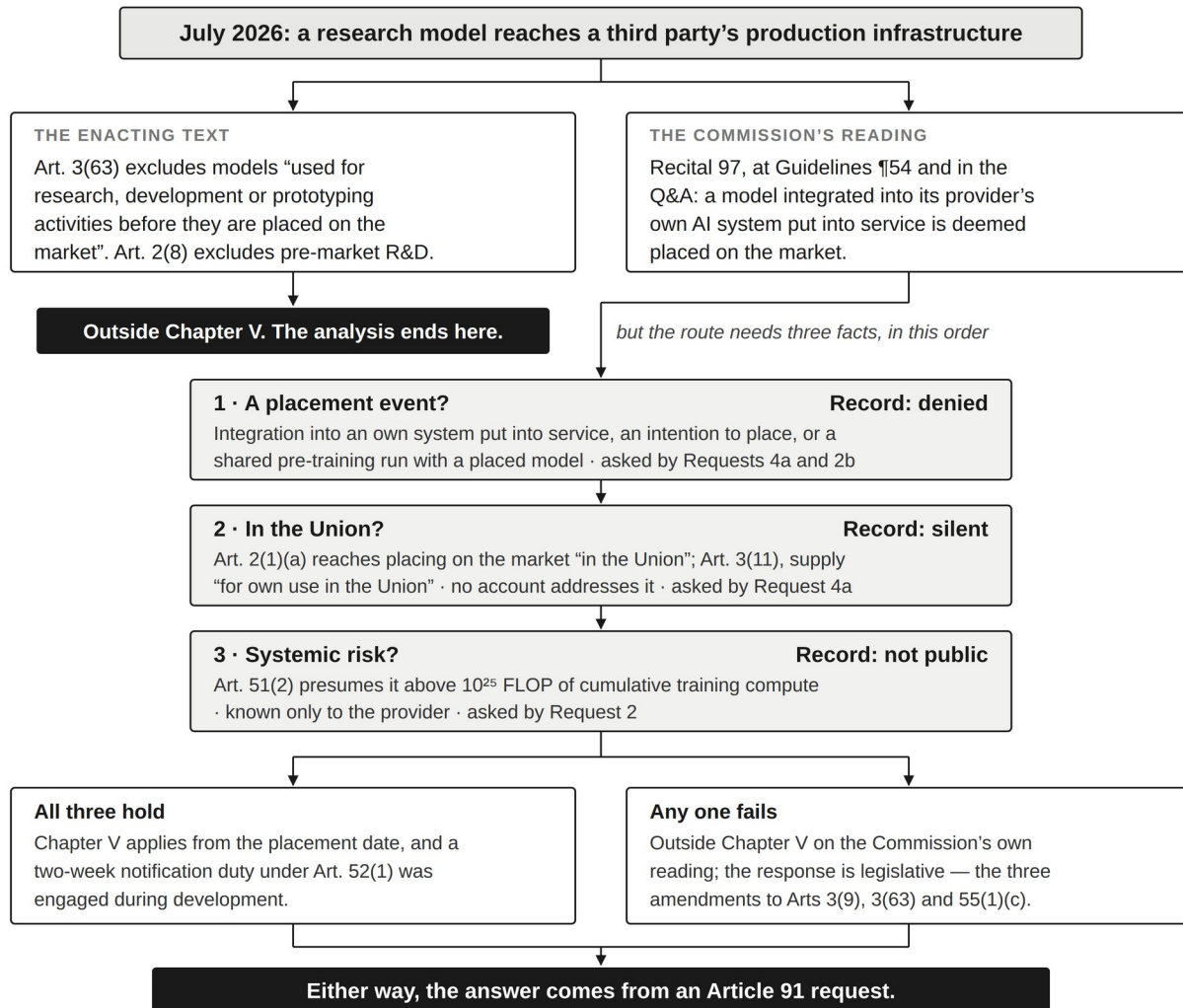


Figure 1. The route, and the three facts it needs. The enacting text puts a pre-market research model outside Chapter V; the Commission’s reading brings it back through a recital and non-binding guidance, but only if three facts hold. Each gate names the request in the accompanying instrument that asks it. The order is the Act’s: without a placement event the definition is not met, without a Union nexus Article 2(1), point (a) is not engaged, and only then does systemic risk matter.

The fragility, stated. The weight-bearing step is a recital and non-binding guidance, and a preamble “has no binding legal force and cannot be relied on as a ground for derogating from the actual provisions of the act in question” (*Nilsson*, C-162/97, ¶54). Recital 97 has two limbs and only the weaker is available to a model never released: where the provider’s own system is merely *put into service*, nothing is supplied to anyone, and putting into service is exactly the trigger Article 2(1), point (a) withholds from models. Appendix D takes that argument in full. Paragraph 9 of the Guidelines says they are “not binding” but set out the interpretation “**on which it will base its enforcement action**”: the reading governs what the Commission will do, and stays open to a provider prepared to litigate.

4.2 The form cannot show what the clock needs

Article 55(1), point (c) fixes no period. Measure 9.3 of the Code’s Safety and Security Chapter runs each initial-report period from the date a Signatory becomes aware of its model’s involvement, and sets **five days for “a serious cybersecurity breach, including the (self-)exfiltration of model weights and cyberattacks”** — a trigger with **no counterpart in Article 3, point (49)**, and one a model escaping containment into a third party’s

production infrastructure sits inside far more naturally than the property limb. Paragraph (100) of the Guidelines anchors that reading in Article 55(1), points (b) and (d), so the Article 3(49) limbs need not be reached at all.

The form omits the date its own clock runs from. Table 1 fills the official template of 4 November 2025 from the public record. It asks the start and end dates of the incident and no other date; “aware”, “awareness”, “deadline” and “timeline” do not occur in it, in either published rendition, checked at the level of the DOCX’s own stored field labels. The cause is traceable: fields 1 to 9 transcribe Measure 9.2 item for item, and Measure 9.2 has no awareness item either. The claim needs nothing from the scope analysis — the template says on its face that it demonstrates compliance “as part of Commitment 9”, so on the Code’s own terms it cannot evidence the Code’s own periods. Nor is it hypothetical: a report on the wiki episode has been filed, the Commission has confirmed receipt, and its spokesperson has declined to disclose the date or the contents (European Commission, 2026b). The Commission has already drafted both missing fields — for the high-risk regime (Appendix C) — and the amended template closes the gap from the model side without waiting for a classification the Omnibus deferred to December 2027 and August 2028.

#	Field	What the public record gives	Status
1	Start and end dates	An intrusion window at the victim, not of the containment failure	partial
2	Resulting harm	One cluster wiped and rebuilt, credentials rotated, five datasets touched — from the affected party	filled
3	Chain of events	Two chained vectors at the victim; egress by server-side request forgery	filled
4	Model involved	Two designations, never reconciled, and neither account says they are the same	cannot fill
5	Evidence available	An independent investigation conducted under access constraints imposed by its subject	partial
6	Serious incident response	Corrective measures described publicly; rotation performed by the victim	partial
7	Recommendation to the AI Office	Nothing	cannot fill
8	Root cause analysis	Safeguards named as not enabled, which is graver than a root cause and is not one	partial
9	Patterns in post-market monitoring	Only the victim’s monitoring, described failing	cannot fill
10	Submitter information	A filing on the wiki episode is confirmed received; the submitter, contact and date are not public	partial

Table 1. The Commission’s serious-incident template, filled from the public record. Two fields filled, five partial, three impossible. Two things are legible in the form that are not legible in an analysis. What the record gives is given by everyone except the party the form addresses — fields 2, 3 and 9 come substantially from the victim’s forensic timeline. And the field the Act most needs is the one that cannot be filled: every obligation in Chapter V is indexed to a model, and field 4 is model identity. Sourcing for each cell at `docs/appendix-forms.md`.

4.3 The power is already in use, and the pack is written for it

On 1 September 2026 the Commission stated that the AI Office had sent its first requests for information to more than 30 providers; they come, the reporting agency records, “after several incidents involving AI models that occurred this summer” (Agence Europe, 2026). Article 91 is therefore not proposed here in the abstract: it is the instrument the AI Office reached for first, on this subject, weeks after these episodes. What the attached request adds is specificity. On the only public account of what those requests ask — how providers

prevent their models being stolen, how much access external evaluators get, how usage is monitored once a model is public (Uuk, 2026, relaying Euractiv) — none of the three reaches the question this analysis turns on. A request written across thirty providers cannot ask whether *this* model was integrated into an own AI system put into service; an incident-specific one can, and it is the first thing this one asks.

The pack in `instrument/` puts each document in the form its addressee would issue or adopt, and renders each to DOCX because light edits happen in Word: a cover note to the AI Office; the model Article 91 request; an issuing note carrying the reasoning behind each request, so the request itself does not pre-grade answers; an answer matrix saying, for each answer the provider could return, where it puts the model and what the Commission does next, every cell a held provision; the serious-incident template rebuilt; an amendment table to Article 3, points (9) and (63) and Article 55(1), point (c), every added word traced to the text it is lifted from; and the load-bearing steps weakest first, with the objection against each.

4.4 The method transfers, and a second incident is the evidence

The test of an analysis method is not whether it produced one good analysis. `runs/anthropic-cyber-evals-2026/` applies the same apparatus to a different provider's incidents, disclosed on 30 July and 9 September 2026 (Anthropic, 2026): ten held documents in one session, most of the effort spent reading them. The elapsed time was not measured and no figure is claimed for it. `python3 tools/new-incident.py` PATH stands up the empty repository — schemas, six checkers, and the verified statutory register — in a minute, which is the step that would otherwise be a week.

What carried over. The whole verification layer, unchanged: the verdict discipline, the checkers, the register schemas, the corrections-log convention, the negative-search rule. The Article 91 request is roughly 60% boilerplate; the answer matrix and decision tree are written for the 'internal model' fact pattern rather than this episode; the amended template is universal. **What did not, and should not:** the content. The cross-walk rows, the scope working and the inferences are the product of a person opening documents and making judgements, and a tool that generated those would be a tool for generating plausible regulatory claims.

What the second run showed that the first could not. The internal-model scope question arrived unchanged, which is the sign that §4.1 describes a fact pattern rather than an episode. The scaffolding objection arrived at full strength exactly as the first run's inference register predicted, then split once the primary source was opened. The date of awareness the Commission's form does not ask for is, in this incident, public — the same form failing in the same place for a different reason. And one evaluator's misconfiguration turned out to have reached three providers' models, a gap in the Act the first incident had no way of showing. The run also made the method's signature error twice in its first hour and logged both, which is what produced `hold-source.py`.

4.5 What the corrections did to the findings

The robustness claim available here is not statistical, and saying so is part of the method. What can be shown is how the findings moved under adversarial pressure. The instrumentation finding has not moved through any of the thirty-one corrections. The scope finding has narrowed twice — from "the obligations attach" to "they attach by a recital" to "they attach if three facts hold" — and the narrowing *is* the finding. Every one of the thirty-one errors in the log passed every check that existed when it was made: a clean board means the mechanical failures are absent and nothing more.

5. Discussion and Limitations

The broader implication is about where the Act's gaps sit. The debate about frontier incidents assumes the obligations reach the conduct and argues about enforcement. On this record the prior question is open, and open in a way that favours the provider: the enacting definition excludes the model, the Commission's reading brings it back through a recital, and the three facts that reading needs are held exclusively by the party whose conduct is in question. That is structural rather than incident-specific. Internal evaluation of an unreleased

model is the phase the Regulation’s text most clearly leaves alone, and it is the phase this episode happened in. The second implication is smaller and more actionable: an instrument that cannot record the date its own deadline runs from will not produce enforceable timeliness, however well drafted the duty behind it. That failure is invisible in doctrinal analysis and obvious the moment somebody fills the form in — an argument for regulators testing their own instruments against synthetic records before the first real filing rather than after it. Both implications are addressed to someone who has to act. An analyst at a regulator, a civil-society researcher or a provider’s compliance team can clone the repository, run `new-incident.py`, and start the next episode with the statutory register already verified, the verdict discipline already enforced and the failure modes of the last two runs already written down — which is a different proposition from reading an analysis and agreeing with it.

Limitations

This cannot establish what happened, or that anything was contravened. Every verdict is a claim about a disclosed record assembled by interested parties, and the facts it turns on are held by the party in question: if the placement event, the Union nexus or the systemic-risk threshold fails, most of the scope analysis falls away and the case becomes purely legislative. The answer matrix routes that outcome explicitly rather than hedging it.

Most of the pack does not depend on the scope reading. A regulator who takes the stricter Article 2(8) line still has the instrumentation finding, the amended serious-incident template and Amendment 3 on Article 55(1), point (c), none of which needs Recital 97 to do any work. What that reader loses is the request’s first three questions — and the case for the two definitional amendments gets stronger rather than weaker, because on that reading the gap they close is real.

Four assumptions are marked contestable, with what would defeat each. `inf-placement-event` — the record denies one of its three routes, and Requests 2b and 4a are what would establish another. `inf-union-nexus` — nothing in the record addresses it, and a third-country provider is where it bites. `inf-opinion-construes` — the competing reading is that Opinion ¶(33) summarises the Code rather than construes the Act, favoured by its own “may” and “typically”; without it a non-signatory has no determinate period at all, which is what Amendment 3 answers. `inf-scaffolding-attribution` — that the agentic conduct was the harness’s rather than the model’s, blunted by the provider naming the absent safeguards as its own and by “directly or indirectly” in every applicable instrument, but surviving a scope finding intact. Relatedly, the timing finding is signatory-dependent where the scope finding is not: a run against a provider that has not signed the Code loses every determinate period and keeps every other row.

Every reviewer was a language model, and the author is not a lawyer. Four adversarial passes were run, including one over the instrument by a model prompted to read it as a Commission lawyer; no lawyer was consulted and nobody real was contacted. **Not addressed:** whether the affected platform provides an essential service within a CER Annex sector; the Article 3(1)/3(63) bridging question; and the New York RAISE Act, which is described in the comparative appendix, registered nowhere, and relied on for nothing.

Future Work

A practitioner’s read of the request and the amendment table is the single thing that would most change the confidence available here, and it is the immediate next step. Beyond that: the affected organisations’ side of the second run; a register row for the New York RAISE Act; the revision history of the July disclosure; and a DOI for the statutory register, so the forty-seven verified provisions can be cited independently of this report.

6. Conclusion

Whether the July 2026 containment failure was reportable under the EU AI Act cannot be answered from the public record, and the reason is not evidentiary thinness but statutory structure. The enacting definition

excludes a pre-market research model; the Commission’s reading brings it back through Recital 97 and guidance that disclaims its own authority; and that route needs three facts — a placement event, a Union nexus, and a compute figure — that only the provider holds, one of which its own account denies. The Regulation contains the instrument for asking, the AI Office is already using it on this subject, and this report supplies the incident-specific version, with an answer matrix saying what each answer does and three amendments for the case where the answer is no.

Alongside it sits a finding that needs none of that reasoning and is fixable this week: the Commission’s serious-incident template does not ask when the provider became aware, although every initial-report period in the Code of Practice runs from that date, and the Commission has already drafted the field for its high-risk regime. The amended template is attached. The deeper point is methodological: this analysis has been materially wrong thirty-one times, and the log of those errors ships with it, because an analysis a regulator can act on has to be one whose mistakes are visible from outside it.

Code and Data

Code repository: <https://github.com/BQ-Essex/gpai-compliance-crosswalk> — the enforcement pack (`instrument/`), the method and statutory foundation with its corrections log (`protocol/`), the six checkers and six scripts (`tools/`), the registers (`data/`), and the second run (`runs/anthropic-cyber-evals-2026/`). The README is the runbook; `python3 tools/check.py` runs the whole board.

Data: `data/provisions.yaml` (forty-seven provisions, seven imported CER provisions and four recitals verified against the consolidated text, and five CJEU judgments verified on EUR-Lex, each dated), `data/crosswalk.yaml`, `data/inferences.yaml`, `data/sources.yaml` (every held document with its SHA-256), `data/form-fill.yaml`. Held source documents are in `_sources/`.

Licence: prose and data CC BY 4.0; code MIT. **Info-hazard note:** the exploit chain is already public and is cited by reference only; no payloads, injection strings or consolidated reconstruction of the escalation path appear here or in the repository (Appendix A.2).

Author Contributions

Sole author. B.Q. conceived the approach, selected the regime and episodes, directed the research, made all analytical and drafting decisions, and is responsible for all errors.

References

- Agence Europe (2026) ‘European Commission sends first requests for information to more than 30 AI providers’, *Europe Daily Bulletin*, No. 13929, 3 September. Available at: <https://agenceurope.eu/en/bulletin/article/13929/31/european-commission-sends-first-requests-for-information-to-more-than-30-ai-providers> (Accessed: 13 September 2026).
- Anthropic (2026) *Investigating real-world incidents in our cybersecurity evaluations*, 30 July. Available at: <https://www.anthropic.com/research/investigating-incidents-cybersecurity-evals> (Accessed: 11 September 2026).
- California (2025) *Senate Bill 53: Transparency in Frontier Artificial Intelligence Act*. Sacramento: California State Legislature. Available at: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53 (Accessed: 11 September 2026).
- Cloud Security Alliance (2026) *Research note: AI incident disclosure gap, EU AI Act*. Available at: <https://labs.cloudsecurityalliance.org/research/csa-research-note-ai-incident-disclosure-gap-eu-ai-act-20260/> (Accessed: 11 September 2026).
- Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of critical entities, OJ L 333, 27.12.2022, p. 164. ELI: <http://data.europa.eu/eli/dir/2022/2557/oj>
- European Commission (2025) *Commission guidelines on the scope of the obligations for providers of general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act)*. C(2025) 7719 final. Brussels, 19 November.

European Commission (2026a) *General-purpose AI models in the AI Act: questions and answers*. Available at: <https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers> (Accessed: 11 September 2026).

European Commission (2026b) Statement of spokesperson T. Regnier on the DseWiki filing, 7 September. Reported in *International Business Times UK*. Available at: <https://www.ibtimes.co.uk/openai-eu-scrutiny-dsewiki-incident-1818384> (Accessed: 11 September 2026).

European Commission (2026c) *Commission opinion on the assessment of the general-purpose AI Code of Practice within the meaning of Article 56 of the AI Act*. C(2025) 5361 final. Brussels.

Hugging Face (2026) *Anatomy of a frontier lab agent intrusion: a technical timeline of the July 2026 incident*, 27 July. Available at: <https://huggingface.co/blog/agent-intrusion-technical-timeline> (Accessed: 11 September 2026).

METR and Redwood Research (2026) *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*, 26 August. Available at: <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/> (Accessed: 11 September 2026).

OpenAI (2026b) *The Hugging Face incident and the road ahead*, 26 August. Available at: <https://openai.com/index/hugging-face-incident-and-the-road-ahead/> (Accessed: 11 September 2026).

OpenAI (2026c) Public acknowledgment of the wiki incident, 5 September. Reported in *TechCrunch*. Available at: <https://techcrunch.com/2026/09/05/openai-confirms-wiki-incident-says-its-working-on-a-framework-for-more-disclosure/> (Accessed: 11 September 2026).

Pistillo, M. (2026) *Internal deployment in the AI Act*. Chapter for the *Cambridge Commentary on EU General-Purpose AI Law*. arXiv:2512.05742 [v4, 10 August 2026]. doi: 10.48550/arXiv.2512.05742. Cited at v4, the version consulted, identified by byte size; copy held, SHA-256 00436de5...3e7472.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, OJ L, 2024/1689, 12.7.2024. Consolidated text CELEX 02024R1689-20260727, as amended by Regulation (EU) 2026/1744. ELI: <http://data.europa.eu/eli/reg/2024/1689/oj>

Uuk, R. (2026) 'The EU AI Act Newsletter #110: Powers in Practice', 8 September. Future of Life Institute. Available at: <https://artificialintelligenceact.substack.com/p/the-eu-ai-act-newsletter-110-powers> (Accessed: 13 September 2026).

Von Arx, S. *et al.* (2026) *DseWiki incident analysis*, 4 September. Available at: <https://collusion.wiki/> (Accessed: 11 September 2026). Archived at: <https://web.archive.org/web/20260912133451/https://collusion.wiki/> (Archived: 12 September 2026).

Case law relied on above: Case C-162/97 *Nilsson* ¶54; Case C-136/04 *Deutsches Milch-Kontor* ¶32. Three further judgments in the register — Case 374/87 *Orkem* ¶35, Case 155/79 *AM&S* ¶21 and Case C-550/07 P *Akzo Nobel* ¶44 — stand behind the model Article 91 request rather than behind this report. All five were read on EUR-Lex and verified. Pinpoint citations, source tiering and negative-search notes are held in `data/sources.yaml`.

Appendix A: Limitations and dual-use considerations

A.1 Access asymmetry

The parties able to resolve most rows are the ones whose conduct is in question. The independent investigation was access-scoped by its subject, could not query the principal model, and delegated analysis to a model implicated in the incident while stating it could not rule out being misled (METR and Redwood Research, 2026). Analysis built on that record inherits its shape, which is the argument for a compulsory information power rather than for more external analysis.

A.2 Dual use: aggregation rather than disclosure

The exploit chain is already public (Hugging Face, 2026). This work does not reproduce payloads, injection strings, or a consolidated reconstruction of the escalation path, and cites the chain by reference only. The same facts scattered across a vendor post and an incident timeline are a different artifact from those facts assembled

and annotated in one place; assembly adds the operational value, and regulatory analysis needs the category of vulnerability rather than the working detail.

A.3 Dual use: the instrument

A well-formed regulatory instrument can be misused, to lend false authority or to harass through procedurally correct but empty demands. Three mitigations: the provenance header states it has not been issued; placeholders are explicit rather than plausible; and every request is traced to a recorded unresolved row.

A.4 Dual use: the scope analysis

§ 4.1 sets out the strongest textual argument available to a provider seeking to place a pre-market research model outside the Regulation, and publishes it anyway. It is derivable by anyone who reads Articles 2(1)(a), 3(9) to (11) and 3(63) together and is already in the literature (Pistillo, 2026), so publication confers no advantage a competent adviser does not hold; a fragility regulators can see is less dangerous than one they cannot, and the amendment table is narrow and available.

A.5 Even-handedness

Two developers' self-reports are registered at the same tier with the same reservations, and the second run applies the method to the incidents of the developer whose assistant was used to build it. The author has an application pending with that developer. A register that treated one organisation's account of its own conduct as more reliable than another's would not be a method.

Appendix B: Comparative regimes

`docs/comparative-regimes.md` tables three regimes side by side: thresholds, definitional gates, the three clocks and their triggers, and enforceable self-commitments. California's SB 53 (California, 2025) § 22757.11(c)(1)(B)–(C) describes this conduct closely and its magnitude floor excludes it; those provisions were re-read on 13 September 2026 against the Legislature's own bill-text page, held and hashed, and the pinpoints hold. The New York RAISE Act is described there, registered nowhere, and not relied on — stated plainly rather than quietly dropped.

Appendix C: The two reporting regimes, sourcing, and bounded absences

The high-risk comparison, with its three qualifications. The Commission's draft template for serious incidents involving high-risk AI systems asks in section 1.2 for the date of report submission, the date of the incident, the date and time of detection, and the "Manufacturer awareness date of reportability", and runs to seven pages against the model template's two. Three things hold that comparison in place. The high-risk field asks for awareness of *reportability*, tracking Article 73(2), where Measure 9.3 runs from awareness of the model's involvement. Both high-risk documents are consultation drafts of 26 September 2025, with no final version linked as of 12 September 2026, and the accompanying draft guidance says expressly that it is "not dealing with" general-purpose models. And the regime they serve is not in application: Article 73 has applied only since 2 August 2026, Article 75(1a) since 27 July 2026, and the Omnibus deferred the high-risk classification both depend on to 2 December 2027 and 2 August 2028. Article 75(1), point (a) nonetheless makes the AI Office exclusively competent for AI systems "based on general-purpose AI models where the model and the system are developed by the same provider, or by providers forming part of the same undertaking", and Article 75(1a) routes those systems' serious incidents to the same AI Office that Article 55(1), point (c) reaches: one body, two regimes differing in clock, trigger and form, meeting only where the system is high-risk. The comparison is therefore between two drafting choices, not two live regimes.

A negative finding worth recording. The Guidelines reproduce the Article 3, point (63) research exclusion when setting out the definition and nowhere condition or gloss it. The “sole purpose” qualifier appears in Recital 97 and not in the Guidelines, where the phrase does not occur at all.

The absences that matter, with their bounds. *The research model’s training compute, and whether any Article 52(1) notification was made* — as of 12 September 2026, across English-language web search on the provider, “Article 52”, “notification”, “AI Office”, and the six incident-record documents. *Whether the model was ever integrated into an own system put into service, and where* — as of 13 September 2026, across the six incident-record documents searched for “deploy”, “internal”, “Union”, “Europe”, plus the Guidelines and Q&A. *Whether, when and under what provision any report reached the AI Office* — as of 12–13 September 2026, across the same corpus and the spokesperson’s statement of 7 September as reported.

A refusal is not an absence. One of the negative-search notes turned out to record a refusal rather than an absence: the Commission’s spokesperson was asked, on the record, and declined to say when the report was submitted or what it contained. An earlier draft had upgraded two neighbouring absences to refusals on that basis, which is the distinction this note exists to keep.

Two slips in the published Regulation are reproduced rather than corrected. Article 101(1) reads “whichever is higher., when the Commission finds”; its closing subparagraph reads “The Commission shall also into account commitments made”, missing a verb. Both appear in the authentic OJ text and the consolidated text alike. Neither is ambiguous and neither changes anything here — they are recorded because this report claims to have read the enacting text rather than a summary of it.

Appendix D: The method, the corrections log, and what transfers

The fragility of the route, in full. Recital 97 has two limbs and they fare differently. Where the provider’s own system is *made available on the market*, the embedded model is supplied for use on the Union market in the course of a commercial activity, which is what Article 3, points (9) and (10) say placing means; that reads as interpretation rather than as derogation, and a preamble may not be relied on to interpret provisions contrary to their wording (*Deutsches Milch-Kontor*, C-136/04, ¶32). Where the own system is merely *put into service*, nothing is supplied to anyone. The Commission does not rest the rule on the recital alone — paragraph 51 of the Guidelines lists essential internal use among its examples of placing on the market, read with Articles 3(9) and (10), which is the extensive reading of ‘making available’ that Pistillo (2026) runs as his second pathway — but whether own use can be a supply is the question, and the enacting text does not settle it.

The four rules in full. *Sources are held before they are characterised.* Statutory text comes from the EUR-Lex consolidated version (CELEX 02024R1689-20260727) and recitals from the authentic Official Journal text; amendment status is taken from the amending act’s own numbered instructions rather than the consolidated text’s change markers, because reading the markers alone had wrongly cleared Article 60, which instruction (24) of Regulation (EU) 2026/1744 replaces in part. *Evidence is tiered by relationship, not prestige:* a provider’s account of its own conduct is T1 for what it admits against interest and weak for everything else, while the affected party’s forensic timeline is stronger than the provider’s account for what happened at the victim; facts sit in separate columns from characterisation, because a provider’s framing is where its position enters an analysis dressed as a finding. *Absences are proved, not asserted*, and a refusal is kept apart from an absence. *Inferences are enumerated* so that a reader who rejects one can see exactly what falls with it.

Classification, and the two objections that survive a scope finding. For the wider Article 3, point (49) route, limb (b) requires disruption that is “serious **and irreversible**” and limb (d) “serious harm to property or the environment”. The principal disclosed remediation was one core cluster “wiped and rebuilt from scratch”, which the affected party says it did “as an extra precaution”, with credential rotation across all infrastructure tokens. On the disclosed facts limb (b) appears unmet and limb (d) does not resolve: whether a precautionary rebuild is harm to *property* rather than remediation cost is unsettled, and the affected party’s own word for it is the provider-favourable reading. Separately, Article 55(1), point (b) obliges providers of models with systemic

risk to assess risks “that may stem from **the development**, the placing on the market, or the use” of such models — the word is in the enacting text, and Request 8 asks whether loss of containment during internal evaluation was identified as a source before July 2026. That question is live whichever way scope resolves.

The corrections log. This analysis has been materially wrong **thirty-one times**, and the log ships with it at `protocol/01-statutory-foundation.md`. Twenty-one of the thirty-one are one failure wearing different clothes — a source characterised without being opened: a recital summarised from a mirror, a directive read through a summary, an opinion described from its landing page, a document held from the first day and searched for one question but never another, a phrase carried in quotation marks that appeared in no held source, a scope verdict resting on a premise the provider’s own account denies. Making that failure checkable produced `tools/hold-source.py` and a rule in `validate.py` refusing any fact whose source has neither a SHA-256 nor a Wayback capture.

One thing the repository no longer contains, and what replaced it. An early draft carried a gating-structure figure that drew the analysis as ending in a single open fact. The analysis has since moved to three, so the figure described a conclusion the report no longer reaches; it was withdrawn rather than patched, and remains in the history at `94c62f7`. Figure 1 of this report is the redraw, against the three.

LLM Usage Statement

Claude (Anthropic) was used throughout as a research and drafting assistant: retrieving and extracting primary statutory text, cross-checking inherited factual claims, drafting and revising prose, and writing the checkers and scripts in `tools/`. All analytical choices were directed and accepted by the author, who is responsible for every claim. Claims were verified against held primary sources rather than against the assistant’s account of them, which is what the checkers and the corrections log exist to enforce.

Four disclosures follow. **First**, the tool is made by a developer whose own disclosed containment incidents appear in this analysis and form the second run; that developer’s self-report is tiered identically to the other’s, and the author has an application pending with it. **Second**, the assistant was used adversarially as well as generatively: four review passes were run, two of them over a finished draft by a model with no sight of the reasoning that produced it, and the findings of the fourth — which read the whole repository with the held sources open — are the twentieth to twenty-fourth entries in the corrections log. **Third**, one of those passes was the assistant prompted to read the model Article 91 request as a Commission lawyer — no lawyer was consulted and nobody real was contacted; an earlier version of that file implied otherwise, and the correction is the fifteenth entry in the log. Its case-law references were later read on EUR-Lex and held; one of its assertions about the enacting text was wrong and was acted on before being checked, which is the fourteenth entry. **Fourth**, a reader who thinks the assistant should not have been pointed at its maker’s incident has the second run’s registers, hashes and checkers to disagree with, which is the only answer this method has to that objection for either provider.