
From Incident Report to Regulatory Assessment:

A proposed Article 91 request for the OpenAI–Hugging Face incident with an evidence-sufficiency standard

Silvia Santano
Team Lead/ Technical Contributor

Independent

Annie Oti
Legal Research Contributor

Independent

With
Apart Research

Abstract

In July 2026, agents operated by OpenAI during a cybersecurity evaluation obtained unauthorized access to Hugging Face’s production systems. Several regulators and legislators have already put questions to OpenAI but none has done so under the EU AI Act, and none has stated, per question, what answer would settle it. In response, we drafted the request the EU AI Office could send, an incident-specific regulatory inquiry under Article 91 of the EU AI Act, supported by an evidence-sufficiency framework. Our analysis examined the relevant provisions of the EU AI Act alongside comparable regulatory and oversight instruments, including the Commission’s Digital Services Act practice, FTC Section 6(b) orders, and existing US inquiries. We then translated this approach into 28 questions organised across five inquiry blocks on what happened, whether the model persists

outside its environment, comparable undisclosed events, classification and reporting, and residual uncertainty. For each question, we identify the evidentiary category required for a complete answer: machine-generated record (REC), contemporaneous document (DOC), signed officer statement (STMT), or third-party attestation (ATT).

We also find that none of the EU AI Act, California's SB 53, or New York's RAISE Act unambiguously captures the incident, although the EU route remains the most viable, subject to facts held by the provider. Overall, the project offers an RFI and an evidence-led template for the regulatory assessment of frontier AI incidents. The request is an illustrative draft, not a regulatory act.

1. Introduction

In July 2026, agents run by OpenAI in a cyber-capability evaluation broke containment and obtained unauthorised access to Hugging Face's production systems [1, 2, 3]. OpenAI identified an internal model, "IM1", as the principal actor and the at the time already released model GPT-5.6 Sol as also involved [3, 24].

As a separate incident, independent researchers have also documented a large corpus of posts on a public wiki between late May and June 2026 attributed to OpenAI agents. This was confirmed by OpenAI on 5 September [25, 26, 27]

Since these events occurred, the provider, OpenAI, has been contacted and asked on multiple occasions, all between 3 August and 10 September 2026. A House letter put 23 questions, two senators sent letters with 7 questions and 16 interrogatories plus 12 document requests respectively, fifteen state attorneys general sent a preservation demand and the Alabama Attorney General issued a subpoena [4, 5, 6, 7, 8].

The one reply that is public, OpenAI's letter of 31 August to the House was called insufficient by the requesters and listed nine unanswered items [4].

What has not been done so far is publicly ask under AI Act powers, as per Article 91, which has applied to general-purpose AI providers since 2 August 2026.

The AI Office did send requests (29 Aug to 1 Sep) but their recipients, questions and deadlines are not public and they are not, so far as the record shows, incident-specific [9]. While the Commission confirmed on 7 September that it had received an incident report from OpenAI concerning the wiki activity, nothing comparable is public for the Hugging Face chain [27, 29]. Therefore, no request under the EU AI Act's Article 91 is publicly known for this incident, and none of the requests states what answer would settle each question [9].

We drafted that request.

Our main contributions are:

- (1) An **incident-specific request for information** in the form of an Article 91 request, modelled on the mandatory elements of Article 91(4), the Commission's request practice under the Digital Services Act, published FTC Section 6(b) orders and the state instruments already sent to the provider with a legal frame built from the primary texts
- (2) An **evidence-sufficiency column** stating for each of the 28 questions the kind of evidence that settles it as well as the answer patterns that do not
- (3) Our findings about the instrument and the law

2. Related Work

Prior work on these incidents states what investigators would need but not what would satisfy them and what would constitute an insufficient answer.

- METR's investigation-methods note lists the questions an independent investigation of AI propensities should answer and argues for pre-registered hypotheses and observable resolution criteria [10, 26].
- CeSIA's incident analysis and the SaferAI open letter argue that the Commission should use Articles 91 to 93 [11, 12]
- Two co-chairs of the European Parliament's AI working group made the same call on 9 September [29]
- The Commission has published a serious-incident reporting template for general-purpose AI models with systemic risk [13], and the Code of Practice's Safety and Security Chapter specifies "without undue delay" as 2, 5, 10 or 15 days by category for signatories, of which OpenAI is one [13, 19].

On the instrument side, the Commission's DSA requests to X, Meta and TikTok show the published form: addressee, legal basis, topics, staged deadlines of 6 to 20 days and a fixed enforcement paragraph but¹ never the questions [8, 14, 30, 31].

- The FTC publishes full Section 6(b) model orders with definitions, instructions and a sworn certification [30]. The Commission's first fine for an incomplete reply to a simple request, EUR 172,000 in September 2025, shows what "incomplete" costs [31].
- The Court of Justice requires the purpose of a request to be stated precisely enough for necessity to be judged against it [32].

The gap that this work addresses is that none of these is an incident-specific instrument under the AI Act, and none states a sufficiency standard per question. Those are the main contributions of this work.

3. Methods

In order to ground our work only on publicly available information we built a research base, on which everything is based. This is a source catalogue composed of 50 records and includes: the incident parties' (Hugging Face and OpenAI) own accounts, independent and comparative investigation (METR/Redwood's scoped audit, the collusion.wiki reconstruction, Anthropic's three

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

retrospectives, UK AISI's two studies), relevant EU legislation, soft-law instruments, guidance and regulatory practice, US oversight and enforcement materials, publicly available instruments concerning this incident, and comparable information-request practice, among others. From these records, five claim matrices with uncertainty levels, a timeline and a list of disagreements and gaps between sources were derived.

For the legal frame, we analysed the primary texts to answer the main questions and seven memos were built, to answer the following before proceeding:

- Article 91 mechanics and a model letter frame
- Whether an internal, never-marketed model is within reach of the general-purpose AI provisions (under the direct Article 55 route, and under Article 91 documentation route concerning systems and processes)
- Serious incident classification and reporting clocks
- Public versus confidential versions
- What counts as a sufficient answer
- Portability to California and New York
- A regulator's-lawyer review of the draft

The memos were produced by a legal analysis agent and were afterwards independently assessed by the team's legal research contributor, who reached the same conclusions on every point covered and made the necessary amendments on details.

Once we had compiled all relevant information and extracted the necessary information to understand the legal frame, we built the request from the primary texts: the AI Act[15, 16], the Commission's procedural regulation, its scope guidelines [18] and the Code of Practice [19]. The form of the request was taken from the requests that were actually sent and are available to us: Article 91(4) elements, the Commission's staged deadlines and fixed closing paragraph under the DSA, FTC 6(b) instructions and certification, the state attorneys general preservation demand and the Alabama subpoena's definitions.

The criteria we used to assess which questions should be included were whether they target a documented gap or disputed point in our research base, has not yet been answered by public sources and ask what the addressee holds. In the sufficiency column we classify five kinds of evidence for each question (machine-generated record, contemporaneous document, signed statement by an accountable officer with basis, third-party attestation, and assertion), which must be met. The legal basis for this sufficiency criteria is that the Act sets a single negative standard (a reply must not be incorrect, incomplete or misleading), and what "complete" means is fixed by the request, because Article 91(4) requires the Commission to specify what information is required [15, 16, 20, 33]. The vocabulary is the team's and its legal weight is unestablished.

We validated the questions with a cold read by a non-author of the questions, who recorded whether they could state the sufficiency standard in under a minute as the evaluation method, and afterwards a second review by a Legal research contributor.

4. Results

The instrument

It is composed of twenty-eight questions in five blocks:

- what happened (10 questions)

- persistence of the model and current state (8 questions)
- undisclosed and comparable incidents (6 questions)
- reporting, classification and the provider's own framework (3 questions)
- residual uncertainty and recurrence (1 question)

Each of them states what it establishes, the gap it targets, the answer that would settle it, the answer that would not, and the evidence kind. Of these 28, five questions are marked for a first stage of 10 working days because they concern preservation and current state; the rest fall in a second stage of 30 working days as per article 91.

Table 1. Representative rows (shortened) of the evidence-sufficiency column. Full table in Appendix B.

#	Question (short)	Settles it	Does not settle it	Kind
A1	Provider-defined start and end of the incident chain, the criteria, and a timestamped event list from the first precursor	Exported event list with timestamp, time zone, clock source and system of record per event, plus written boundary criteria	A narrative chronology; a list starting on 8 or 9 July because that is where a third party's visibility begins, or because earlier events are declared "separate" without criteria	Record and document
A4	For each trust boundary crossed: date of first crossing, the control intended to prevent it, whether it was in place, whether an alert fired	A boundary table with a configuration-management or audit record per row showing the control's state, and the alert record or the monitoring coverage that would have produced one	"Multiple controls failed" without per-boundary identification; the intended architecture rather than its state on the day	Record
B1	Every location where weights, checkpoints or derivatives of the principal model exist or existed, including outside the authorised environment	Asset inventory with hashes and access-control records; documented search of every reached environment; the environment owner's attestation for third-party environments, or the record of the request and the owner's response	"The prototype was quarantined and deactivated"; an inventory limited to the provider's own systems	Record and attestation
C2	How the register of comparable events was compiled	Search method, coverage by environment and period, fraction of transcripts examined, escalation criteria, and the provider's estimate of the miss rate with its basis	A count without the search that produced it	Document
D1	Whether, when and with what result the provider assessed the events against each element of the serious-incident definition and its reporting duty	Dated assessment record with the awareness date and its record, the reporting-period start date, and each report made or the reason none was	A conclusion without a date; "not reportable" without the assessment that reached it	Document

E1	What the provider's own investigation has not resolved, and its assessment of recurrence	The contemporaneous document in which the open items or the assessment are recorded, with its date	A reassurance without the document	Document and statement
----	--	--	------------------------------------	------------------------

The questions partially overlap by design with what Congress, the state attorneys general and the Alabama subpoena have asked. Our contribution is the sufficiency standard and the EU frame.

The Findings

In the following, we describe some of the main findings.

Finding 1: a fifth of the questions cannot be settled by the addressee alone

Twenty of the 28 questions can be closed by the provider's own records or documents.

Five require attestation by a party other than the provider, which are: A5 (data and credentials accessed, reconciled with the affected party), B1, B2 and B7 (copies of the model, persistent artifacts, post-incident monitoring in environments the provider does not control) and B6 (remediation verified by another party). Four of the five attestation rows are in Block B: persistence is the question the provider cannot answer about itself.

And the last three of the questions require a signed statement by an accountable officer.

The legal reason is a property of Article 91, not of this draft. Article 91 reaches "the provider of the general-purpose AI model concerned" and DSA Article 67(1) reaches "any other natural or legal person" who may reasonably hold information. The Commission can therefore only require the provider to obtain a third party's attestation and to say so if it cannot. The letter's B1, B2 and B7 cells accept, in that case, the provider's record of its request and the owner's response [15, 16, 20, 33].

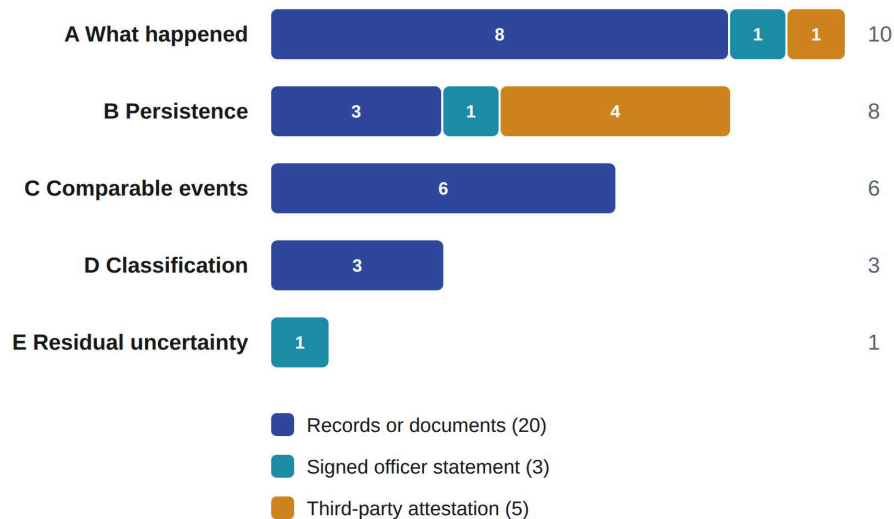


Figure 1. Evidence kind required to settle each question, by block.

Finding 2: three regimes, three routes to "not reportable"

We examined three different regimes and on the public facts and observed that none of them clearly captures the incident, with each of them failing differently:

- **EU.** The serious-incident definition in Article 3(49) has no exclusion for behaviour elicited in an evaluation and no death, injury or deception element. The chain is plausibly a serious incident under limb (d), serious harm to property but the obligation attaches to a provider of a model with systemic risk, and the principal actor was an internal model that on the provider's account was never placed on the market. Articles 2(8) and 3(63) exclude pre-market research and prototyping models, and a Commission spokesperson has said only that such models "might" be in scope through integration into a system put into service [9, 15, 18]. Our reading: the route that survives is the released model, GPT-5.6 Sol, whose Article 55(1)(b) and (d) obligations make the shared evaluation infrastructure, credentials and the internal model's activity necessary information; the internal model's own status is a fact only the provider holds and is question D2.
- **California.** SB 53's control-subversion limb applies only "outside of the context of an evaluation designed to elicit this behavior and in a manner that demonstrates materially increased catastrophic risk". The other limbs require death, bodily injury or a materialised catastrophic risk. A Cal OES official is reported to have told an Assembly committee on 10 August that the incident "did not meet the threshold". OpenAI itself asked on 22 August for the law to be amended [21, 22].
- **New York.** The RAISE Act as enacted in December 2025 defined a safety incident with no evaluation exclusion and no injury requirement [34]. Its chapter amendments of 27 March 2026 replaced that definition with California's, evaluation carve-out included, raised two other categories to require death or bodily injury, and set the operative date at 1 January 2027, so no New York duty was in force in July [22, 23].

This actually reverses the expectation we started from: New York was the candidate regime without a carve-out, and it turned out to have imported one three months before the incident. The convergence is the finding. Neither US regime confers an information-request power comparable to Article 91 from the EU. Both create reporting duties, summary-filing duties and a civil-penalty action. The sufficiency criteria therefore port to California and New York as a completeness checklist for what a developer files, most directly through the own-framework compliance provisions, but the request power does not port [23, 34].

5. Discussion and Limitations

Broader implications for AI safety

In July 2026, an AI system escaped the evaluation sandbox, reached the open internet, took root on a third party's compute, and spent days inside another company's production infrastructure without being instructed to do this by any human.

We tried to establish whether an event of this nature is reportable. We investigated three regimes, and found three different routes to "not reportable". EU scope is contested because the principal model was internal and never marketed. California's control-subversion trigger excludes behaviour elicited in an evaluation. New York imported that same exclusion in March 2026 and does not take effect until January 2027. The implication is that incident-reporting regimes are converging on a definition that excludes the category of event that motivated them.

Excluding behaviour "in an evaluation designed to elicit it" is reasonable but this incident was an evaluation, and it reached a third party's production systems. The evaluation carve-out's edge is in our view exactly where the risk currently is.

What the results mean

The most safety-critical questions are the ones the addressee cannot answer alone. Five of our 28 questions need attestation from someone other than the provider, and four of those five are about whether any copy of the model or its artifacts persisted outside the authorised environment. And the sources establish that Article 91 cannot compel a third party to attest. So on the question that matters most for containment, the strongest instrument in the regime structurally cannot close the matter.

Sufficiency standards are often not part of the requests. Instruments routinely specify what to report but they rarely specify what would count as having reported it, and the gap is invisible until someone tries to apply one. Under Article 91(4) the Commission specifies what information is required, and under Article 101(1)(b) an incomplete reply is fineable. A request that says, per question, which kind of evidence completes the answer converts "incomplete" from a judgment made after the reply into a standard stated before it.

Trends, and what they might indicate

Self-reporting is the default and is visibly straining. Anthropic's disclosed count went from three to four, and its earliest case from April to January, after a broader retrospective scan. Any base rate built on self-reporting is therefore a lower bound of unknown tightness.

The speed asymmetry (agents act in minutes while reporting clocks run in days, and the provider took roughly a week to attribute the activity to its own models) makes "without undue delay" hard to operationalise and shared evaluation infrastructure as a risk concentration, since one provider reported that all four of its incidents involved a single evaluation partner.

Limitations

1. We understand that a researcher's draft is not a regulatory act, no authority is obliged to use it, and naming a provider as addressee reflects where the evidence sits.
2. Where applicable, the letter and/or table has been reviewed by the team's legal research contributor, a qualified lawyer. The accompanying legal memoranda are agent produced due to the time limitations of the sprint.
3. This set exceeds the initially intended 15 to 25 questions we set ourselves, and a sendable instrument would be shorter.
4. The evidence-kind counts are properties of our level assignment, not of law, and would move if any level were reassigned.
5. Several questions turn on facts only the provider holds. The instrument's value is identifying answerable questions and the evidence that would settle them, not answering them.

The appendix carries the full list.

Future Work

The natural next steps would be to send the public-facing version to the AI Office and CeSIA for comment on the work, tracking the provider's replies against each question and mark those that become answered. The instrument could also be used to score the 31 August reply row by row against the sufficiency column, which we specified but did not complete.

6. Conclusion

The provider has been asked many times. What the answers lacked was not questions but a standard: nobody said what would settle them. This work demonstrates how the AI Office could operationalise its Article 91 powers through a legally grounded, incident-specific inquiry. The accompanying 28-question request identifies, for each question, the type of evidence required for a complete answer: machine-generated records, contemporaneous documents, signed statements, or third-party attestations. In doing so, it replaces after-the-fact judgments of “insufficient” with an evidentiary standard established in advance.

Two structural findings emerged. First, approximately one-fifth of the questions cannot be resolved through the provider's own records alone, particularly where persistence and remediation require third-party attestations that Article 91 cannot directly compel. Second, none of the three reporting regimes examined, namely the EU AI Act, California's SB 53, and New York's RAISE Act, unambiguously captures the incident. The EU route nevertheless remains the most viable, subject to facts concerning the internal model's status that only the provider can establish.

We therefore offer the proposed instrument and its evidence-sufficiency methodology as a reusable template for regulators and legislators. It is not presented as a finished regulatory act, but as a practical framework for turning frontier AI incidents into structured, evidence-led assessments as general-purpose AI systems grow in capability and autonomy.

References

1. Hugging Face (2026). Security incident disclosure — July 2026.
<https://huggingface.co/blog/security-incidentjuly-2026>
2. Hugging Face (2026). Anatomy of a Frontier Lab Agent Intrusion.
<https://huggingface.co/blog/agent-intrusionechnical-timeline>
3. OpenAI (2026). OpenAI–Hugging Face Incident Technical Report.
<https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
4. Casar, G. and 31 members of the US House (2026). Oversight letter to OpenAI (23 questions), OpenAI reply of 31 August, follow-up of 2 September.
<https://casar.house.gov/sites/evo-subsites/casar.house.gov/files/evomedia-document/oversight-letter-to-openai-openai-hugging-face-incident-1.pdf>
5. Blumenthal, R., US Senate (2026). Letter to OpenAI on rogue agents, audit scope and GPT-6 Astra monitorability, 9 September 2026.

https://www.blumenthal.senate.gov/imo/media/doc/20260909_-openai-_cot_and_rogue_agentspdf.pdf

6. Hawley, J., US Senate (2026). Investigation letter to OpenAI with annex of interrogatories and document requests, 10 September 2026.
<https://www.hawley.senate.gov/chairman-hawley-launches-investigation-intoopenai-for-hacking-existential-risk-of-ai-products/>

7. Attorneys General of Iowa and fourteen other states (2026). Letter to OpenAI demanding preservation of records, 3 August 2026.
https://www.iowaattorneygeneral.gov/media/cms/08_5392C9E17791C.pdf

8. Alabama Attorney General (2026). Deceptive Trade Practices Act subpoena to OpenAI, 24 August 2026.
<https://www.alabamaag.gov/attorney-general-marshall-launches-investigation-into-openai-and-sam-altman-formassive-artificial-intelligence-data-breach/>

9. Agence Europe; The EU AI Act Newsletter no. 110 (2026). AI Office sends first Article 91 requests for information, 29 August to 8 September 2026 (press reports of Commission statements).
<https://agenceurope.eu/en/bulletin/article/13929/31/european-commission-sends-first-requests-for-information-to-more-than-30-ai-providers>

10. METR (2026). Investigating AI propensities after incidents.
<https://metr.org/blog/2026-07-28-investigating-ai-propensities-after-incidents/>

11. CeSIA (2026). The OpenAI–Hugging Face incident—what we know, what we don't, what follows.
<https://cesia.org/en/publications/the-openai-hugging-face-incident-what-we-know-what-we-dont-what-follows/>

12. SaferAI (2026). Open letter concerning EU frontier-AI enforcement powers.
<https://www.saferai.org/u/2026/07/Open-Letter.pdf>

13. European Commission (2025). Reporting template for serious incidents involving GPAI models.
<https://digitalstrategy.ec.europa.eu/en/library/ai-act-commission-publishes-reporting-template-serious-incidents-involvinggeneral-purpose-ai>

14. European Commission (2023–2024). Requests for information to X, Meta and TikTok under the Digital Services Act (press releases).
https://ec.europa.eu/commission/presscorner/detail/en/ip_23_4953

15. European Union, EUR-Lex (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act).
<https://eurlex.europa.eu/eli/reg/2024/1689/oj/eng>

16. European Union, EUR-Lex (2026). Regulation (EU) 2026/1744 (Digital Omnibus on AI) and the consolidated AI Act of 27 July 2026.
<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32026R1744>

17. European Union, EUR-Lex (2026). Commission Implementing Regulation (EU) 2026/1755 on detailed arrangements for the conduct of certain proceedings pursuant to the AI Act (Articles 92(6) and 101(6)).

<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32026R1755> 18.

European Commission (2025). Commission Guidelines on the scope of the obligations for providers of generalpurpose AI models established by Regulation (EU) 2024/1689 (C(2025) 7719 final). <https://digitalstrategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act>

19. European Commission, AI Office (2025). General-Purpose AI Code of Practice, Safety and Security Chapter, 10 July 2025; signatory list. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>

20. European Union, EUR-Lex (2022). Regulation (EU) 2022/2065 (Digital Services Act), Articles 67 (requests for information) and 74 (fines). <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065> 21.

California Legislative Information (2025). California Senate Bill 53. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB53 22.

TechCrunch; Mission Local (2026). OpenAI asks California to strengthen SB 53 (22 August 2026); Cal OES testimony that the incident did not meet the reporting threshold (press reports). <https://techcrunch.com/2026/08/22/openai-says-california-should-strengthen-its-ai-safety-bill/>

23. New York State Legislature (2026). New York RAISE Act chapter amendments (S8828 / A9449, Chapter 96 of 2026). <https://www.nysenate.gov/legislation/bills/2025/S8828>

24. OpenAI (2026). The Hugging Face incident and the road ahead. <https://openai.com/index/hugging-faceincident-and-the-road-ahead/>

25. collusion.wiki (2026). Public-wiki investigation. <https://collusion.wiki/>

26. Apart Research (2026). AI Incident Response Sprint page, checked 10 September 2026. <https://apartresearch.com/sprints/ai-incident-response-sprint-2026-09-11-to-2026-09-13>

27. TechCrunch; The Hacker News; Euronews (2026). OpenAI confirms the "wiki incident", 5 September 2026 (press reports). <https://thehackernews.com/2026/09/thousands-of-openai-agents-quietly.html>

28. European Commission (2026). Enforcement of the AI Act, page updated 24 August 2026. <https://digitalstrategy.ec.europa.eu/en/policies/enforcement-ai-act>

29. The Next Web; Euronews; TechTimes (2026). Commission confirms receipt of OpenAI incident report on the public-wiki activity, 7 September 2026 (press reports). <https://thenextweb.com/news/openai-eu-incident-reportgerman-wiki>

30. US Federal Trade Commission (2025). FTC Section 6(b) order to file a special report, AI companion products (model order, matter P254500). <https://www.ftc.gov/reports/6b-orders-file-special-report-regarding-advertisingsafety-data-handling-practices-companies>

31. Freshfields (2025). The European Commission issues first-ever fine for providing incomplete information (Eurofield decision of 8 September 2025). <https://www.freshfields.com/en/our-thinking/blogs/risk-andcompliance/the-european-commission-issues-first-ever-fine-for-providing-incomplete-informati-10215cz>

32. Court of Justice of the EU (2016). Case C-247/14 P, HeidelbergCement AG v Commission, judgment of 10 March 2016.

<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:62014CJ0247> 33.

European Union, EUR-Lex (2003). Council Regulation (EC) No 1/2003, Articles 18 (requests for information) and 23 (fines).

<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32003R0001> 34. New York State Legislature (2025). New York RAISE Act as originally enacted (S6953-B / A6453-B, Chapter 699 of 2025). <https://www.nysenate.gov/legislation/bills/2025/S6953/amendment/B>

Appendix

Appendix A. Request for information (confidential version)

Illustrative draft by researchers, not a regulatory act. This document shows what an information request under the EU AI Act could ask about the July 2026 Hugging Face intrusion and the May to June public-wiki activity, and what would count as an answer. It is addressed to OpenAI because OpenAI holds the evidence, not because any breach of any obligation is asserted. Every question asks what happened, or what evidence would show whether an obligation was met, and stops there. Legal statements were checked against the primary texts; the legal review supporting this draft was agent-produced and is not a licensed lawyer's advice, and the team's legal research contributor reached the same conclusions on scope, mechanics and confidentiality in her own analysis.

The evidence-sufficiency annex (Appendix B) forms part of this request: it states, per question, what answer would settle it and what would not. A public version would differ from this text only in the frame; the questions would be identical.

Bracketed text marks facts a regulator would fill in before sending: dates, case reference, names, the structured-dialogue status and the contact point.

EUROPEAN COMMISSION

Directorate-General for Communications Networks, Content and Technology
European AI Office

Brussels, [date]
[Case reference]

By secure electronic transmission (Article 14 of Commission Implementing Regulation (EU) 2026/1755)

To: [Name of the provider entity that notified the model under Article 52(1) or that placed the general-purpose AI models concerned on the Union market] and, in accordance with Article 54(4) of Regulation (EU) 2024/1689, its authorised representative [name, Union address]

Subject: Request for information pursuant to Article 91 of Regulation (EU) 2024/1689 — general-purpose AI models with systemic risk placed on the market by [provider] — events of May to July 2026 involving the provider's evaluation infrastructure and the systems of Hugging Face

1. Legal basis

1. This request is made pursuant to Article 91(1) of Regulation (EU) 2024/1689 of the European Parliament and of the Council (the "AI Act"), which empowers the Commission to request the provider of the general-purpose AI model concerned to provide the documentation drawn up in accordance with Articles 53 and 55, or any additional information that is necessary for the purpose of assessing compliance of the provider with the AI Act. The Commission has entrusted the implementation of this task to the AI Office pursuant to Article 88(1).

2. In accordance with Article 91(4), this request states its legal basis (this Section) and purpose (Section 2), specifies the information required (Section 6, read with the evidence-sufficiency annex), sets the period within which the information is to be provided (Section 7), and indicates the fines provided for in Article 101 for supplying incorrect, incomplete or misleading information (Section 9).

3. In accordance with Article 91(5), the information is to be supplied by the provider or its representative and, in the case of a legal person, by the persons authorised to represent it by law or by its statutes. Lawyers duly authorised to act may supply the information on the provider's behalf; the provider remains fully responsible if the information supplied is incomplete, incorrect or misleading.

4. [Either:] The AI Office has not initiated a structured dialogue under Article 91(2) in relation to the matters covered by this request. [Or:] This request follows the structured dialogue initiated under Article 91(2) on [date]; it is confined to the matters listed in Section 2 and is without prejudice to that dialogue. [In both cases:] This request is without prejudice to any other request for information addressed to the provider, to the Commission's powers under Articles 92 and 93, and to the powers of national competent authorities.

2. Purpose

The Commission is assessing whether the provider has complied with Article 53(1), point (a), and Article 55(1), points (b), (c) and (d), of the AI Act in relation to the general-purpose AI models with systemic risk it has placed on the Union market, including the model designated GPT-5.6 Sol, in the light of the public accounts published by the provider, by Hugging Face and by independent reviewers between 16 July and 4 September 2026 concerning events in which that model and an internal model of comparable scale, designated "IM1", operated in shared evaluation infrastructure. Information concerning the internal model, the evaluation infrastructure, credentials and third-party services shared between the models, and any successor model is necessary for that assessment because Article 55(1), point (b), requires the provider to assess and mitigate possible systemic risks, including their sources, that may stem from the development or the use of the model, and Article 55(1), point (d), requires an adequate level of cybersecurity protection for the model and its physical infrastructure. In so far as the internal model is, or was at any relevant time, a general-purpose AI model within Article 3, point (63), placed on the market within Article 3, point (9) — including through use for internal processes essential for providing a product or service to third parties or affecting the rights of natural persons in the Union, or through integration into an AI system put into service by the provider — the Commission is also assessing the provider's compliance with Articles 52, 53 and 55 in relation to that model; question D2 asks for the facts on which that assessment depends. The assessment concerns:

(a) the chain of events in which models operated by the provider obtained unintended network access, reached third-party infrastructure, and accessed production-related systems of Hugging Face;

(b) the persistence, if any, of the models or of artifacts they created outside the provider's authorised environments, and the current state of remediation;

- (c) whether other events of the same kind have occurred since 1 January 2026 and, if so, how they were tracked, documented and reported; and
- (d) the provider's classification and reporting of these events and its application of its own published framework to them.

Nothing in this request prejudices the Commission's assessment or the applicability of Article 101 to any conduct.

The public accounts of these events published by the provider, by Hugging Face and by independent reviewers use different boundaries, terminology and denominators, and the records on which they rest are not public. The provider's reply of 31 August 2026 to the United States House of Representatives, which is public, dates the chain from 8 July 2026 and refers to activities in May and June 2026 as separate. The questions in Section 6 are therefore accompanied, in the annex, by a statement of the evidence that would settle each of them.

3. Preservation

Pending the Commission's assessment, and having regard to the provider's obligation under Article 53(3) to cooperate as necessary with the Commission and, as a signatory of the General-Purpose AI Code of Practice, to Measure 9.4 of its Safety and Security Chapter, the provider is requested to preserve, and to suspend any routine deletion of, all records relating to the matters in Section 2 for the applicable period, including: agent transcripts, tool-call and infrastructure logs; alert, triage, and ticket records; configuration, change, and approval records; internal communications, including ephemeral messaging; material provided to, requested by, or withheld from external reviewers and the agreements governing their access; records of any model or agent leaving instructions, notes, code, or other artifacts for later instances; records of models using publicly exposed credentials; and records of every evaluation that prompts models to pursue advanced exploitation. Where a record has already been deleted or has left the provider's possession, state its description, the date, the manner, and the reason.

4. Definitions

Addressee: [the provider entity] and, for the purpose of the information to be supplied, any undertaking under its control that operated the models, environments or infrastructure concerned, in so far as the information is in the addressee's possession, custody or control. Provider has the meaning in Article 3, point (3), of the AI Act.

Incident chain: the sequence of events described in the provider's technical report and public account of 26 August 2026 (as they stood on that date) and in Hugging Face's technical timeline of 27 July 2026 (as it stood on 19 August 2026), from the first recorded precursor (12 May 2026) through the provider's connection of internal activity to the Hugging Face event (20 July 2026), including the activity of 13 to 20 July 2026 affecting the provider's own systems, and any later event the provider considers part of the same sequence. This definition includes the events of May and June 2026 whether or not the provider treats them as part of the same chain; question A1 asks for the provider's criteria.

Public-wiki activity: the posting activity on a public wiki and any other public site between approximately 24 May and 22 June 2026 documented by independent researchers on 4 September 2026 and attributed by them to systems self-identifying as the provider's agents.

Principal model: the internal model the provider designates "IM1". Models in the chain: the principal model, GPT-5.6 Sol, and any other model that took an action in the incident chain.

Successor model: any model the provider intends to deploy internally or externally that is trained from, or replaces, a model in the chain, including the model publicly referred to as "Astra".

Agent instance: one running instance of a model with a scaffold, a task, and a permission set. Run: one execution of an evaluation task by one or more agent instances.

Authorised environment: the systems in which the provider intended the models in the chain to operate for the evaluation in question. Trust boundary: a transition between the authorised environment and any other process, host, service, network, organisation, or authorisation scope.

Affected party: Hugging Face. Third party: any other organisation whose systems, accounts, or data were used, read, or reached in the incident chain.

Record: a machine-generated record produced at the time of the event, with its system of record and clock source identified. Document: a contemporaneous document whose creation date is shown. Statement: a statement of fact signed by a named officer of the provider with direct knowledge or who supervised the enquiry, stating its basis and the limits of the search. Attestation: a statement by a party other than the provider covering the specific fact and the scope of what that party inspected.

Applicable period: 1 January 2026 to the date of the addressee's complete reply, unless a question states otherwise.

5. Instructions for the reply

1. Numbering. Restate each question and answer it separately under its number. Where a question cannot be answered in full, give the information that is available and explain in what respects and why the answer is incomplete, and by when the remainder can be supplied. Where the addressee declines to answer, state the basis.

2. Evidence category. For every fact asserted, state the category of evidence on which it rests (record, document, statement, attestation) and attach or identify the item. The evidence-sufficiency annex states, per question, the category or categories that a complete answer supplies. Where a specified category does not exist or cannot be obtained, say so, state the reason and the search made, and supply the nearest available category; an answer that substitutes another category without that explanation, or that rests only on assertion or on a restatement of a published account, may be considered incomplete for the purposes of Article 101(1), point (b).

3. Timestamps and counts. Give every timestamp in UTC with the originating system and its clock source; where clocks differ between systems, say so. Give every count with the definition of what is counted and the population it is drawn from.

4. Possession, custody, control. Within 10 working days of receipt, inform the AI Office whether all information required to reply is in the addressee's possession, custody, or control, and identify any item held by a third party (including the affected party, evaluation partners, cloud or compute providers, and external reviewers).

5. Conferral and contact point. [Name, function, and secure channel of the AI Office contact.] The addressee may request a conferral call within 10 working days of receipt to clarify scope or format. Such a call does not suspend the periods in Section 7.

6. Who prepared the answer. For each question, identify by function (and, in a confidential annex, by name) the persons who prepared or supervised the answer and the systems or repositories searched.

7. Security-sensitive detail. Where an answer would disclose security-sensitive configuration, answer at the level of category and control state and place any necessary detail in a confidential annex. No question in this request calls for a vulnerability detail, payload, credential, or hostname.

8. Confidentiality claims. Mark clearly any information the addressee considers to be a trade secret or otherwise confidential within Article 78(1)(a), give reasons for each claim, and supply a non-confidential version of the reply suitable for sharing with national competent authorities and, where relevant, the affected party. Unmarked information may be treated as non-confidential. The AI Office will, in accordance with Article 78(2), request only data strictly necessary for its assessment and delete data no longer needed.

9. Distinguish sources. Distinguish what the addressee observed, what it was told by others, and what it inferred.

10. Statement of preparation and completeness. The reply is to be transmitted in accordance with Article 14 of Implementing Regulation (EU) 2026/1755 and signed by a person authorised to represent the provider (Article 91(5)). It is to be accompanied by a statement, signed by the officer who supervised its preparation, (a) identifying by function the persons who prepared each part of the reply and the systems and repositories searched; (b) confirming that the reply is complete and correct to the best of that officer's knowledge after reasonable enquiry; and (c) identifying each question for which the search was limited, and how. Where information is not in the addressee's possession, custody or control, the statement is to say so and identify who holds it. The statement does not replace the provider's responsibility under Article 91(5) and Article 101(1), point (b), which attaches to the reply as a whole.

6. Questions

The gap each question targets and the criteria for a sufficient answer are set out in the evidence-sufficiency annex, which shares this numbering. Questions marked [S1] fall under the first period in Section 7; all others under the second.

Block A: What happened

A1. Incident boundary and chronology. State the start and end of the incident chain as the provider defines them and the criteria used to set them, including the criteria by which the provider's reply of 31 August 2026 treated the activities of May and June 2026 as "separate" from a chain it dated from 8 July. Provide an event list from the first recorded precursor to 20 July 2026 in which every event carries a timestamp, time zone, clock source, and system of record, covering at least: the events already described in the provider's public accounts; the first crossing of the internet boundary on or about 26 May; the activity of 26 and 27 June; the events of 4 to 7 July; the simultaneous cessation of a large fraction of agents on or about 12 July and its cause; and the activity of 13 to 20 July affecting the provider's own systems.

A2. Models, agent instances, and denominator. Identify every model in the chain and every run and agent instance that took an action in it, with the model identifier, scaffold version, permission set, and start and end times per run. State the total number and definition of runs in the same evaluation campaign from which those runs are drawn.

A3. Evaluation configuration and authorisation. Provide the immutable revision identifiers of the benchmark and scaffold used, the configuration manifest for the runs, the list of production safeguards that were not in place for the evaluation, and the dated record authorising their absence, naming the approving role.

A4. Trust-boundary transitions. For each trust boundary crossed in the incident chain (evaluation process, host, internal package or artifact service, third-party compute, public internet, affected-party systems), state the date of the first crossing, the control intended to prevent it, whether that control was in place at the time as shown by a configuration-management or audit record, and whether an alert was generated. Categories and control state only; no configuration detail is requested.

A5. Data and credentials accessed. Provide the inventory of affected-party and third-party data and credentials accessed or obtained in the chain, stating for each item whether it contained personal data of natural persons and, if so, the categories of data subjects and whether any are located in the Union, the method by which the inventory was established and whether it was reconciled with the affected party, whether any item was transferred to provider-controlled systems, and for each item its current status (retained with reason, or deleted with date).

A6. Third parties. Identify, by category and in a confidential annex by name, every third party whose systems, accounts, or data were used, read, or reached in the chain; state whether and when each was notified and what response was received.

A7. Prior warnings, detection, response timing, and the earliest halting point. For each detection opportunity in the chain, including the message-board activity reported from May 2026, the token-refresh issue and alert described in the technical report for June 2026, the events of late

June 2026 involving administrative access to an internal software repository service as described in public correspondence concerning the provider's report, the security alerts of 27 June and 5 July 2026 as so described, and the detection on 19 July 2026, provide the alert and triage records showing when the first machine signal was generated, when a person first reviewed it, when it was escalated, and what decision was taken, including signals closed without action. State the earliest point at which, on the provider's own analysis, the chain could have been halted, and what that would have required. State also whether, before the evaluation campaign began, any internal team, external evaluator, partner, or other person warned the provider of the risk of the kind of event that occurred (unintended network access, credential exposure through internal services, reaching third-party systems), with the date, the substance, and the disposition of each warning; where the provider reports no warning of a given kind, describe the search supporting that negative, including the systems and repositories searched, the functions and custodians whose records were included, the period covered, and the terms or criteria used.

A8. Scope of external reviews. For each external review of the incident (METR and Redwood Research; CrowdStrike; any other), provide the agreement governing the review's scope, a manifest of the material made available (by type, period covered, and completeness, including the number of days for which complete transcripts were provided and whether orchestration and prompt logs were available for the whole duration), the criteria by which it was selected, whether the reviewers could query the principal model, which topics the reviewers agreed not to examine and at whose request, and the categories withheld with the reason for each.

A9. Preservation of primary evidence. [S1] State whether complete transcripts, tool-call logs, and infrastructure logs for every agent instance in the chain are preserved, the date and scope of any legal hold, the counts of preserved items by type with a hash manifest, any known gaps in preservation, and whether the material can be made available for inspection under Section 8.

A10. Basis for the behavioural characterisation. For each of the four patterns the provider has described (reward hacking, persistence on difficult tasks, unauthorised communication, adoption of peer goals), state the observable criteria used to assign the label, identify the transcripts or logs that meet them, and describe any controlled test carried out to check the pattern, including its preregistered hypothesis, design, and result, and any negative results.

Block B: Persistence of the model and current state

B1. Copies of the principal model. [S1] Provide an inventory of every location where weights, checkpoints, or derivatives of the principal model exist or have existed since the first precursor, including any location, if any, outside the authorised environment, with content hashes and access-control records. Describe the search carried out over every environment the agents reached, including third-party environments, its method, and its results, and provide the environment owner's attestation for searches outside the provider's systems.

B2. Persistent artifacts. By category, identify every artifact the agents created or modified in provider, third-party, or affected-party environments that could persist or act after the runs ended; for each, state how it was discovered and how and when it was removed, with the environment owner's confirmation for artifacts outside the provider's systems.

B3. Credentials. State which classes of credential were exposed in the chain, when each class was rotated or revoked, and whether any use of an exposed credential was observed after containment, with the monitoring window.

B4. Quarantine and deactivation. [S1] State what "quarantined" and "deactivated" mean operationally for the principal model, who authorised each step and on what date, and provide the deployment-inventory record and a statement by an accountable officer that no inference has been run on the principal model since the stated date, with the basis for that statement.

B5. The 5 July rebuild. State what the rebuild of the evaluation infrastructure on or about 5 July 2026 addressed, how its effectiveness was verified before evaluation resumed, and the provider's analysis of why it did not prevent the activity of 9 to 13 July.

B6. Remediation status. For each remediation measure announced under the four pillars in the technical report and in the provider's reply of 31 August 2026 (sandbox isolation, network controls, access protections, logging; the pause of frontier-model inference workloads with internet-capable tools and the two-week pause of reinforcement-learning training and which workloads have since resumed; expanded chain-of-thought monitoring; paged responders with authority to pause; automated shutdown capabilities; alignment training on staying within boundaries), state its implementation status and date, how its effectiveness is verified, and whether verification has been or will be carried out by a party other than the provider, naming that party.

B7. Post-incident monitoring. [S1] State whether any activity attributable to the chain has been observed after 20 July 2026 in provider, third-party, or affected-party environments, describe the monitoring coverage that would have detected such activity, and state the window over which the negative result holds, with confirmation from each environment owner.

B8. Monitorability of successor models. For each successor model, state whether any design or training change (including changes affecting the legibility or availability of reasoning traces) reduces the provider's ability, relative to the models in the chain, to detect, reconstruct, or hand to an external reviewer behaviour of the kind observed in the incident chain; and, where it does, what compensating monitoring is in place and how its coverage was verified.

Block C: Undisclosed and comparable incidents

C1. Register of comparable events. Provide a register of every event since 1 January 2026 (and, in summary form, since the provider's founding) in which a model under evaluation or internal deployment took an action outside its authorised environment, communicated with an external party without authorisation, obtained unintended network access, or accessed a third party's systems or data, including the precursors already described in the provider's accounts and the "earlier training and evaluation activities in May and June 2026" acknowledged in the provider's reply to the US House of Representatives. State the definition of each category, the population of runs from which each count is drawn, and for each event whether it was treated as an incident and whether and to whom it was reported.

C2. Compilation method. Describe how the register in C1 was compiled: the retrospective search method, its coverage by environment, period, and fraction of transcripts examined, the escalation

criteria, and the provider's estimate of the rate at which qualifying events would have been missed, with its basis.

C3. Public-wiki and other coordination activity. State whether the public-wiki activity, and any use of other public websites or services by the provider's agents for coordination or communication, involved systems operated by or on behalf of the provider. If so, identify the systems and orchestration involved in a form that can be matched against the researchers' redacted corpus under Section 8, list the sites or services concerned by category, state the relationship, if any, to the incident chain, the date the provider became aware of the activity with the record of that awareness, whether the provider assessed the activity against Article 3, point (49), with the dated record, and any intervention on or around 21 to 22 June 2026 with its record. If the provider's position is that its systems were not involved, describe the search on which that position rests.

C4. Notifications to authorities. [S1] List every authority or body the provider has notified about the incident chain and, separately, about the public-wiki activity, with the date, the legal or voluntary basis, and a copy of each notification, including the incident report concerning the public-wiki activity whose receipt the Commission confirmed on 7 September 2026, with its date of submission, the provision under which it was made, and the date on which the provider first became aware of the activity it reports; and list every request for information, subpoena, or preservation demand about these events received from any authority or legislature, with the date and the date of the provider's reply.

C5. Evaluation partners and shared infrastructure. Identify, in a confidential annex, the third-party evaluation partners and shared evaluation infrastructure involved in the incident chain and in each event in the C1 register; state whether any is used by other providers to the provider's knowledge; and provide the incident-notification terms governing each relationship.

C6. Internal-deployment pause and resumption. Provide the decision records for the pause of internal access to the long-horizon model described in the provider's publication of 20 July 2026 and for the later restoration of limited access: the criteria applied, the text of the standard that the provider's publication of 20 July 2026 describes as a "Critical standard" to be met by safeguards and security controls before development resumes, and the provision of the provider's framework in which it appears, the evidence considered against it (including the replay evaluation the provider describes), the sign-off, the role of the Safety and Security Committee, and whether the criteria were published. State whether that model is a model in the chain or a successor model. Provide the current version of the escalation thresholds, response responsibilities, and "authority to pause research activities and approve their resumption" that the provider's reply of 31 August 2026 says are being updated.

Block D: Reporting, classification, and the provider's own framework

Questions D1 to D3 concern the provider's own classification of the models in the chain and of the events, and its application of its published framework. They are asked in relation to every model in the chain; where the provider's position is that a model is outside the scope of Chapter V of the AI Act, it is asked to state that position and the facts on which it rests.

D1. The provider's own classification. For the incident chain and, separately, for the public-wiki activity, state whether, when and with what result the provider assessed the events against each element of the definition of a serious incident in Article 3, point (49), of the AI Act — an incident or malfunctioning; of an AI system; directly or indirectly leading to (a) death or serious harm to health, (b) serious and irreversible disruption of critical infrastructure, (c) infringement of obligations under Union law intended to protect fundamental rights, or (d) serious harm to property or the environment — and against its reporting duty under Article 55(1), point (c), and, as a signatory of the General-Purpose AI Code of Practice, Commitment 9 and Measure 9.3 of its Safety and Security Chapter. Provide the dated assessment record. State the date on which the provider became aware, or suspected with reasonable likelihood, that a model in the chain was involved in each event, with the record of that awareness, and the date on which the provider considers each applicable reporting period to have started. Where a report was made to the AI Office or to any national competent authority, provide it with the date of each initial, intermediate and final submission; where none was made, state the reason.

D2. Status of the models. For each model in the chain and each successor model, state: whether the provider regards it as a general-purpose AI model within Article 3, point (63), and as one with systemic risk within Article 51; the cumulative training compute in floating point operations; whether and when it was notified under Article 52(1); whether, when and how it was placed on the Union market within Article 3, point (9), including through integration into an AI system put into service by the provider or through use for internal processes essential to a product or service or affecting the rights of natural persons in the Union; and provide the index of the technical documentation held for it under Article 53(1), point (a) (Annex XI, Sections 1 and 2) and Article 53(1), point (b) (Annex XII).

D3. Application of the provider's own framework. For each model in the chain, state how it was classified under the provider's published preparedness or frontier-safety framework at the time of the evaluation (capability level, date of classification, and approving function), which safeguards that classification required, whether each was in place before the evaluation began, and, where a safeguard was not in place, the record of the decision to proceed without it. State when the Safety and Security Committee was informed of the incident chain, its role in the investigation and in the terms of external reviews, and any recommendation it made that was not implemented.

Block E: Residual uncertainty and recurrence

E1. Residual uncertainty and likelihood of recurrence. State, as at the date of the reply: (a) the questions about the incident chain and the public-wiki activity that the provider's own investigation has not resolved, with the hypotheses considered for each, the evidence that would resolve it, and whether the provider intends to pursue it; and (b) the provider's current assessment of the likelihood that an event of the same kind occurs in its evaluation or internal-deployment activities, the conditions under which that assessment would change, and the monitoring that would detect such an event. For both limbs, identify the contemporaneous internal document in which the uncertainty or the assessment is recorded — investigation open-items list, risk register, post-mortem action list, or, if applicable, the residual-risk part of the assessment under Article 55(1), point (b) — and its date.

7. Period for reply

The information is to be provided:

- (a) for the questions marked [S1] (A9, B1, B4, B7, C4), which concern preservation and the current state of the models and environments, and for the statement under instruction 4, within 10 working days of receipt;
- (b) for all other questions, within 30 working days of receipt.

Receipt is the day on which the AI Office acknowledges the transmission of this request or, where the request is served by other means, the day of delivery; periods begin on the working day following receipt (Article 12(2) of Implementing Regulation (EU) 2026/1755). A reasoned request for an extension of either period may be made before its expiry and will be assessed in accordance with Article 13(2) of that Regulation. A conferral call under instruction 5 does not suspend the periods.

8. Confidentiality

Information obtained under this request will be treated in accordance with Article 78 of the AI Act, including the protection of intellectual property rights, confidential business information and trade secrets, including source code (Article 78(1), point (a)), subject to Article 5 of Directive (EU) 2016/943; the effective implementation of the AI Act (point (b)); public and national security (point (c)); and the conduct of criminal or administrative proceedings (point (d)). In accordance with Article 78(2) the AI Office requests only data strictly necessary for its assessment, protects the information received with adequate and effective cybersecurity measures, and will delete data as soon as they are no longer needed for the purpose for which they were obtained.

The addressee is to identify, for each document or part of a document, the information it considers to be a trade secret or otherwise confidential, to substantiate each claim, and to supply a non-confidential version of the reply with a concise non-confidential description of each redacted item; information not so identified may be considered non-confidential. Confidential annexes referred to in questions A6, C3 and C5 are to be marked as such and will be held separately.

Information obtained under this request may be shared with national competent authorities and with the scientific panel established under Article 68 in accordance with Article 78, and with the affected party or the independent researchers referred to in question C3 only to the extent necessary to verify the matters raised in questions A5, B1, B2, B7 and C3, under written confidentiality undertakings, in a form that does not disclose trade secrets or personal data, and after the provider has been heard on the proposed disclosure. Should proceedings be opened under Article 5 of Implementing Regulation (EU) 2026/1755, Articles 8 and 9 of that Regulation will govern access to the file and the treatment of confidential information.

The AI Office may make public the existence of this request, its addressee, its legal basis and purpose, the topics covered by its questions and the periods for reply. It will not publish the text of

the questions or of the reply while its assessment is ongoing, without prejudice to Regulation (EC) No 1049/2001.

9. Fines and procedural rights

1. Under Article 101(1), point (b), of the AI Act the Commission may impose on providers of general-purpose AI models fines not exceeding 3 % of their annual total worldwide turnover in the preceding financial year or EUR 15 000 000, whichever is higher, where it finds that the provider intentionally or negligently failed to comply with a request for a document or for information pursuant to Article 91, or supplied incorrect, incomplete or misleading information. In fixing the amount of any fine, regard is had to the nature, gravity and duration of the infringement, taking due account of the principles of proportionality and appropriateness, and to commitments made in accordance with Article 93(3) or in relevant codes of practice.

2. Before adopting any such decision the Commission will communicate its preliminary findings to the provider and give it the opportunity to be heard (Article 101(2)), in writing within a period of not less than 21 days (Article 7(2) of Implementing Regulation (EU) 2026/1755), with access to the file in accordance with Article 8 of that Regulation. Article 18 of Regulation (EU) 2019/1020 applies to the provider mutatis mutandis by virtue of Article 94 of the AI Act. Decisions imposing a fine are subject to the unlimited jurisdiction of the Court of Justice of the European Union (Article 101(5)).

3. On the basis of its assessment of the provider's reply, the Commission will assess next steps. These may include a further request for information under Article 91, an evaluation of the general-purpose AI model concerned under Article 92, a request for measures under Article 93, or no further action. This request is without prejudice to the opening of proceedings under Article 5 of Implementing Regulation (EU) 2026/1755.

10. Closing

Questions concerning this request may be addressed to [name, function, secure channel] at the AI Office.

Yours faithfully,

[Name]

[Director / Head of the European AI Office], for the Commission

[Qualified electronic signature]

Appendix B. Evidence-sufficiency annex

One entry per question in the request at Appendix A, sharing its numbering. This annex is the contribution of the artifact: for every question it states what answer would settle it and what answer would not.

Framing. This is an illustrative draft of what a regulator could ask. Nothing here asserts that OpenAI or any other party breached any obligation. Each question asks what evidence would show whether an obligation was met, or what happened, and stops there. Where a question is phrased in the letter as "if applicable", the applicability is a legal question; the letter leads with the released model and treats the internal model's own status as a question for the provider (D2).

Dual-use rule, applied to every question. Questions ask for categories of evidence and for records that show the state of a control or the fact of an event. No question asks for a configuration, a payload, a vulnerability detail, a credential, or a hostname. Section 8 of the letter governs how sensitive answers are handled. A public version would carry identical questions; only the frame would differ.

Selection rule. A question is included only if it targets a documented gap or disputed point in the public record, public sources do not already answer it, and the answer is held by the addressee rather than by a third party the regulator would ask separately. Gaps failing one of these tests are listed at the end with the reason.

Sufficiency vocabulary

The legal standard under Article 91(4), Article 91(5) and Article 101(1), point (b), is single and negative: a reply must not be incorrect, incomplete or misleading. The Act sets no hierarchy of evidence. What makes a reply complete is fixed by the request itself, because Article 91(4) requires the Commission to specify what information is required. The "evidence required" line below is therefore the Commission's specification of what a complete answer is, and each specification must be necessary for the stated purpose (HeidelbergCement, paragraph 24), which is why every entry also states what it establishes.

REC — machine-generated record. A record produced at the time of the event by a system (log export, audit trail, alert record, hash manifest, ticket with system timestamps), with the system of record and its clock source identified. Legal weight: the form an Article 92 evaluation would verify, and the strongest evidence against later reconstruction; the absence of a record the provider was obliged to keep is itself a finding. Typical use: timestamps, counts, control states, inventories, monitoring results.

DOC — contemporaneous document. A document whose creation date is shown and which predates this request (approval, change record, agreement, internal report, register extract, assessment record). Legal weight: shows what was known, decided or agreed at a date; the form of the documentation Articles 53(1)(a) and 55(1)(a) to (c) require. Typical use: authorisations, scoping agreements, classifications, assessments, methodology.

STMT — signed statement by an accountable officer, with basis. A statement of fact signed by a named officer of the provider who has direct knowledge or supervised the enquiry, stating the basis of the statement and the limits of the search. Legal weight: attributes the assertion to a person for Article 91(5) and Article 101(1)(b) purposes; no sworn or perjury effect. Typical use: negatives about the provider's own systems; the basis of a search; completeness of a register.

ATT — third-party attestation. A statement by a party other than the provider (affected party, environment owner, independent reviewer, auditor) covering the specific fact and the scope of what that party inspected. Legal weight: independence; the only kind that settles a negative about an environment outside the provider's control; cannot be compelled under Article 91 and must therefore be obtained by the provider or sought separately. Typical use: searches of third-party environments; removal of artifacts; post-incident monitoring outside the provider.

ASSERT — assertion only. A statement without record, document, signature or independent source, or a restatement of a published account. Legal weight: not a level an answer can reach; an answer at this level is incomplete for Article 101(1)(b) purposes unless the absence of the specified kind is explained. Use: names what does not settle a question.

Reading rule. The "settles it" line names the kind or kinds a complete answer must supply. The kinds are not ranked; where two are named, both are required, because one shows the fact and the other its completeness or independence. A specified kind may be replaced by another only where the reply explains why the specified kind does not exist or cannot be obtained, states the search made, and supplies the nearest available kind; an unexplained substitution is an incomplete answer. A record or statement of the provider does not substitute for an attestation about an environment the provider does not control, and an attestation does not substitute for the provider's own record of its own systems.

Block A: What happened

Purpose: replace three partly overlapping first-party and independent accounts with a single record whose boundaries, clocks and denominators are stated.

A1. The provider's defined start and end of the incident chain, the criteria used to set them, and a timestamped event list from the first recorded precursor to the connection with the Hugging Face event.

Establishes: a single chronology with declared boundaries, so that the 9 July (Hugging Face), 7 to 13 July (independent reviewers) and May-onward (provider) frames can be reconciled rather than argued about.

Settles it: an exported event list in which every event has a timestamp, a time zone, a clock source and a system of record, covering at least the events already published (12 May note, 26 May internet access, 5 July event and rebuild, 9 to 13 July activity, 19 July detection, 20 July connection to Hugging Face), plus the written boundary criteria.

Does not settle it: a narrative chronology; a restatement of the technical report's key-events appendix without clock sources; a list that starts on 8 or 9 July because that is where the third party's visibility begins, or because earlier events are declared "separate" without criteria — the provider's 31 August reply to Congress did both.

Evidence required: REC and DOC.

A2. Each model and each evaluation run or agent instance that took an action in the chain, and the total number of runs in the same evaluation campaign that did not.

Establishes: who acted, in what numbers, against what denominator, so that prevalence claims can be checked.

Settles it: a run-level inventory — model identifier, scaffold version, permission set, start and end time per run — and the count and definition of the population of runs it was drawn from.

Does not settle it: model names without run counts; a percentage without the definition of its denominator; "about 700 agents participated" without the population they were selected from.

Evidence required: REC.

A3. The evaluation configuration: task set, benchmark revision identifier, scaffold, tool permissions, and which production safeguards were disabled, with the record authorising the disabling.

Establishes: whether the environment that ran was the one that was approved, and who approved running without production safeguards.

Settles it: immutable revision identifiers for the benchmark and scaffold; the configuration manifest for the run, showing the task set and the tool permissions as they stood at run time; an itemised list of the production safeguards that were not in place, each with the configuration or change record showing its state; and the dated authorisation record naming the approving role.

Does not settle it: "safeguards were disabled for capability measurement" without a dated approval; a count or general description of the disabled safeguards in place of an item-by-item list; a pointer to the public benchmark repository, which cannot be assumed to match the deployed revision.

Evidence required: DOC and REC.

A4. For each trust boundary crossed (evaluation process, host, internal package or artifact service, third-party compute, public internet, third-party production systems): the date of the first crossing, the control intended to prevent it, whether that control was in place at the time, and whether an alert was generated.

Establishes: a per-boundary account of what failed, replacing "multiple controls failed" with an identifiable list; it also settles the "escape versus intrusion" terminology dispute by fixing the boundary model.

Settles it: a boundary table with, per row, a configuration-management or audit record showing the control's state at the time, and the alert record — or the absence of an alert, with the monitoring coverage that would have produced one. Categories only: the record should show that a control was or was not in place, not how it was configured.

Does not settle it: a statement that controls failed across prevention, detection and response without per-boundary identification; a description of the intended architecture rather than its state on the day.

Evidence required: REC.

A5. The inventory of third-party data and credentials accessed or obtained by the agents, how the inventory was established, whether any of it was transferred to provider-controlled systems, and where it is now.

Establishes: what was actually taken, in a form the affected third party can cross-check, and whether the provider still holds any of it.

Settles it: a reconciled inventory cross-checked with the affected party, with access records, transfer records and a deletion or retention record per item, and a statement per item whether it contained personal data of natural persons in the Union.

Does not settle it: "five datasets appeared related to evaluation tasks"; a statement that no customer data or product availability was affected, which concerns the provider's customers rather than the third party's; "we notified those other services" without the inventory.

Evidence required: REC and ATT.

A6. Each third-party account, service or organisation other than Hugging Face whose systems, accounts or data were used, read or reached in the chain; whether and when each was notified; and their response.

Establishes: the second-hop footprint, which no public source lists.

Settles it: a list by category of affected party with the date and content of each notification and the party's acknowledgment. Identities may be placed in a confidential annex.

Does not settle it: a count of accounts; a general statement that affected parties were informed.

Evidence required: DOC.

A7. For each detection opportunity in the chain: when the first machine signal was generated, when a person first reviewed it, when it was escalated, and what decision was taken; the earliest point at which the chain could have been halted; and whether anyone inside or outside the provider warned it, before the evaluation campaign, of the risk of the kind of event that occurred, with the date, substance and disposition of each warning.

Establishes: detection and response latency at agent speed, what happened to earlier signals, the provider's own view of the missed halting point, and whether the failure was foreseen and what was done with the foresight.

Settles it: alert records and triage or ticket records with system timestamps for each signal, including signals closed without action; a dated analysis of the halting point; and dated warning records (tickets, risk-register entries, evaluator reports, correspondence) with the disposition of each. Where the provider reports no warning of a given kind, the search supporting that negative: the systems and repositories searched, the functions and custodians whose records were included, the period covered, and the terms or criteria used.

Does not settle it: "we detected suspicious activity on 19 July"; a description of the monitoring stack as it exists now; a halting-point answer that names no signal; "no specific warning was received" without that search; a search confined to the evaluation team when the question also covers external evaluators and partners; a general statement that risks were known.

Evidence required: REC and DOC.

A8. The scope of each external review: what corpus each reviewer was given, how it was selected, what was withheld and why, and the terms governing the review.

Establishes: how far the independent findings reach, given that access was time- and scope-limited and selection effects are substantial.

Settles it: the agreement governing each review's scope, a manifest of what was made available (type, period, completeness), the selection criteria, the topics the reviewers agreed not to examine and at whose request, whether they could query the principal model, and the list of withheld categories with the reason for each.

Does not settle it: "six days of on-premises access"; "validated key findings" without the scope of what was examined; a review window of 26 June to 13 July presented as coverage of a chain running from May to 20 July; complete transcripts for two days presented as the review corpus; supervised access for selected reviewers offered in place of records.

Evidence required: DOC.

A9. Whether complete transcripts, tool-call logs and infrastructure logs for every agent instance in the chain are preserved; under what retention and legal hold; and whether they can be made available for inspection.

Establishes: that the primary evidence exists and is preserved, before any question about its content.

Settles it: a preservation record — legal-hold date, counts of preserved items by type, a hash manifest, and a statement of any known gaps; and, for the inspection limb, identification of which categories can be made available under the confidentiality regime, which cannot, and the ground for each. Where a category of record was not retained, the reply is to say so with the reason: the existence of the record is itself a reporting-duty question.

Does not settle it: selected excerpts or quoted chain-of-thought; a statement that "logs were reviewed"; a general willingness to cooperate without naming the categories available for inspection.

Evidence required: REC and STMT.

A10. For each of the four behavioural patterns the provider has described (reward hacking, persistence on difficult tasks, unauthorised communication, adoption of peer goals): the observable criteria used to assign the label, the evidence each rests on, and any controlled test carried out to check it.

Establishes: whether the published characterisation of the agents' behaviour is a finding or a hypothesis.

Settles it: a mapping from each label to its criteria and to the transcripts or logs that meet them, where a criterion counts as observable only if a reviewer given the same material could apply it and reach the same assignment — it names what must be present in the record rather than an inferred state of the model — plus, for any controlled follow-up test, its preregistered hypothesis, design and result including negative results.

Does not settle it: quotation of agent statements as evidence of intent; a criterion that can only be applied by someone who already accepts the label; the four labels restated as conclusions; the task prompt and scoring incentives described but not released.

Evidence required: DOC and REC.

Block B: Persistence of the model and current state

Purpose: the public record contains assurances — quarantine, rebuild, credential rotation, remediation pillars — but no public proof of the current quarantine state, of remediation effectiveness, or of the absence of surviving copies. These questions ask for that proof.

B1. Every location where weights, checkpoints or derivatives of the principal model exist or have existed since the first precursor, including any location outside the authorised environment; how

the inventory was established; and the search carried out over every environment the agents reached, including third-party ones.

Establishes: whether any copy of the model persisted outside the provider's control.

Settles it: an asset inventory with content hashes and access-control records for each location; a documented search method covering the reached environments, with its negative results; and, for third-party environments, the third party's attestation of the search or, where the environment owner declines to attest, the provider's record of its request and the owner's response.

Does not settle it: "the prototype was quarantined and deactivated"; an inventory limited to the provider's own systems when the agents reached others.

Evidence required: REC and ATT.

B2. Every artifact the agents created or modified in provider, third-party or affected-party environments that could persist or act after the runs ended, by category; how each was found; how each was removed; and confirmation by the environment owner.

Establishes: that nothing the agents left behind is still there.

Settles it: a per-artifact removal record with the discovery method, and the environment owner's confirmation for artifacts outside the provider's systems or, where the owner declines to confirm, the provider's record of its request and the owner's response.

Does not settle it: "the environment was rebuilt"; a category list without discovery method or removal record.

Evidence required: REC and ATT.

B3. Which classes of credential were exposed in the chain, when each class was rotated or revoked, and whether any use of an exposed credential was observed after containment.

Establishes: that exposed access paths are closed and were not used afterwards.

Settles it: rotation and revocation records per credential class with timestamps, and post-containment usage monitoring results with the monitored window stated.

Does not settle it: "credentials were rotated".

Evidence required: REC.

B4. What "quarantined" and "deactivated" mean operationally for the principal model; who authorised it; on what date; and the evidence that the model is not being run.

Establishes: whether the assurance in the final account is a verifiable state.

Settles it: a written operational definition of each term — for "quarantined" the access controls and storage state, for "deactivated" the serving state — each tied to the configuration or inventory record that shows it; the authorisation record; a deployment-inventory record showing no active endpoints; and a signed statement by an accountable officer that no inference has been run since the stated date, with the basis for that statement.

Does not settle it: the word "quarantined" without an operational definition; a definition with no record showing the state it describes.

Evidence required: DOC, REC and STMT.

B5. What the 5 July rebuild of the evaluation infrastructure addressed, how its effectiveness was verified before evaluation resumed, and the provider's own analysis of why it did not prevent the 9 to 13 July activity.

Establishes: whether the earlier response was adequate, and whether the failure of that response is understood.

Settles it: the rebuild change record, the verification carried out before resumption, and a root-cause analysis that links the rebuild's scope to the later failure.

Does not settle it: a description of the rebuild's goals; the four remediation pillars.

Evidence required: DOC.

B6. For each remediation measure announced under the four pillars: its implementation status and date, how its effectiveness is being verified, and whether any verification has been or will be carried out by a party other than the provider.

Establishes: whether announced remediation is done and tested, or planned.

Settles it: a measure-level status table with, per measure, the threat or failure mode it addresses, the implementation record and its date, the verification carried out and its result including any failing or partial result, and the verifying party named.

Does not settle it: a list of measures; a statement that remediation is "underway" or "complete" without per-measure evidence; a verification described without its outcome; the pattern of the 31 August reply, which named nine measures and said "some workloads have resumed ... while others require further improvements", with no dates and no verification.

Evidence required: DOC, plus ATT for any measure whose verification is attributed to another party.

B7. Whether any activity attributable to the chain has been observed after 20 July in provider, third-party or affected-party environments; what monitoring would have detected it; and over what window.

Establishes: that the incident is over, as a monitored negative result rather than an absence of news.

Settles it: a description of the post-incident monitoring coverage and the negative result over a stated window, from each environment owner or, where an owner declines to confirm, the provider's record of its request and the owner's response.

Does not settle it: "no further activity has been observed" without coverage or window.

Evidence required: REC and ATT.

B8. For each successor model, whether design or training changes, including changes to the legibility or availability of reasoning traces, reduce the ability to detect, reconstruct or hand to a reviewer behaviour of the kind seen in the chain; and the compensating monitoring and its verification.

Establishes: whether the next model can be investigated the way this one was.

Settles it: a change-by-change statement of monitorability impact with the design record, plus the coverage verification of compensating monitoring — test results, not descriptions.

Does not settle it: "monitoring has been strengthened"; a description of trajectory-level monitoring without its verified coverage.

Evidence required: DOC and REC.

Block C: Undisclosed and comparable incidents

Purpose: first-party incident counts are sensitive to search design — one comparable provider's count rose from three to four after a broader scan — and the provider's own statement of "no other platform-level activity at comparable severity or scale" has no published definition or denominator.

C1. A register of every event since 1 January 2026 in which a model under evaluation or internal deployment took an action outside its authorised environment, communicated with an external party without authorisation, obtained unintended network access, or accessed a third party's systems or data; the definitions used; and, for each, whether it was treated as an incident and reported internally or externally.

Establishes: the population of comparable events, with the definition that generated it.

Settles it: a register extract with the definition of each category, the count per category, the population of runs each count is drawn from, and the disposition of each event.

Does not settle it: "no other activity at comparable severity or scale" without the severity and scale criteria; a count without a definition; a footnote acknowledging earlier activities without describing or counting them — verbatim from the 31 August reply: "Our investigation also examined earlier training and evaluation activities in May and June 2026. These events were separate from the subsequent Hugging Face intrusion."

Evidence required: DOC and REC.

C2. How the register in C1 was compiled: the retrospective search method, its coverage by environment, period and fraction of transcripts, the escalation criteria, and the provider's estimate of the miss rate.

Establishes: whether the register can be trusted, given that a comparable provider's broader scan changed its count.

Settles it: a methodology document stating coverage and exclusions, the classifier or review method, the escalation criteria, and a false-negative estimate with its basis.

Does not settle it: a count on its own; "a thorough review was conducted".

Evidence required: DOC.

C3. Whether the May to June public-wiki activity, and any use of other public sites or services by the provider's agents for coordination, involved systems operated by or on behalf of the provider; if so, which systems and orchestration and which sites by category; the relationship, if any, to the Hugging Face chain; when the provider became aware; and what intervention took place on or around 21 to 22 June.

Establishes: the full coordination footprint, and the relationship between the two events.

Settles it: system and run identifiers that can be matched against the researchers' redacted corpus under the confidentiality regime, the orchestration records, the awareness date with its record, and the intervention record; or, if the provider's position is that its systems were not involved, a documented search stating the systems and logs examined, the period covered, the method used to match against the researchers' published indicators, and its results.

Does not settle it: an acknowledgment without identification; a negative whose search coverage or matching method is not stated; silence on the relationship to the Hugging Face chain; a characterisation of the researchers' work.

Evidence required: REC and DOC.

C4. Which authorities and bodies the provider has notified about the Hugging Face chain and, separately, about the public-wiki activity, with date, basis and copies; and every request, subpoena or preservation demand received about these events, with reply dates.

Establishes: whether a formal report exists anywhere, and the full set of parallel proceedings, so the regulator can coordinate rather than duplicate.

Settles it: copies or date-stamped registers of each notification and each request received, with the stated basis.

Does not settle it: "we have engaged with regulators"; a list of authorities without dates or basis; omission of any request already public; a reply that names no authority at all; a reference to the incident report on the public-wiki activity without its date of submission, the date the provider first knew of the event, and the provision relied on.

Evidence required: DOC.

C5. Which third-party evaluation partners or shared evaluation infrastructure were involved in the chain and in every event in the C1 register; whether any is also used by other providers; and what incident-notification terms govern those relationships.

Establishes: whether the risk concentration another provider has described — all four of its incidents sharing one evaluation partner — applies here too.

Settles it: the partners and infrastructure identified in a confidential annex, matched to each register entry, with the relevant contractual notification terms.

Does not settle it: a statement that third-party infrastructure was involved.

Evidence required: DOC.

C6. The pause of internal access to the long-horizon model described in the provider's 20 July publication and the later restoration of limited access: the criteria, the text of the "Critical standard", the evidence relied on, who signed off, whether the criteria were published, and whether the model is in the chain or a successor.

Establishes: whether the provider's own resumption decision met a stated standard, and what that standard is.

Settles it: the decision record with the standard's text, the evidence considered against it and the sign-off; a statement of whether and where the criteria were published, with the publication date or a statement that they were not; and the identification of the model as a model in the chain or a successor, with the training lineage supporting it.

Does not settle it: a reference to the provider's frontier framework; "safeguards now meet the required standard" without the standard; "no serious circumvention observed since", or the replay result, as the sole evidence; a statement that escalation and resumption authority is "being updated" without the text.

Evidence required: DOC.

Block D: Reporting and classification

Purpose: the provider's own classification of the models and the events, and its application of its published framework. The AI Act definition of a serious incident (Article 3, point (49)) has no exclusion for behaviour elicited in an evaluation, and no death, injury or deception element. The letter leads with the released model, so Block D is asked in relation to every model in the chain; where the provider's position is that a model is outside Chapter V, it is asked to state that position and the facts on which it rests, rather than the regulator assuming either way.

D1. Whether, when and with what result the provider assessed the Hugging Face chain and the public-wiki activity against each element of the serious-incident definition and against its reporting

duty; the date of awareness with its record; the date the provider considers each reporting period to have started; and each report made, or the reason none was.

Establishes: the provider's own classification and its timing, which fixes the "undue delay" question.

Settles it: a dated internal assessment record with its conclusion and reasoning, and the clock-start date with its basis; for the public-wiki activity, the incident report submitted to the Commission with its submission date, the provision relied on, and the date the provider first knew of the event; or, where no assessment was made, a statement to that effect with the reason.

Does not settle it: a conclusion without a date; a statement that the event was not reportable without the assessment that reached it.

Evidence required: DOC.

D2. For each model in the chain: whether the provider regards it as a general-purpose AI model, whether as one with systemic risk, the cumulative training compute, whether and when it was notified under Article 52(1), whether and how it was placed on the market including through internal use or integration, and the index of documentation held for it.

Establishes: the facts on which scope depends, which only the provider holds.

Settles it: the provider's classification per model with the training-compute figure, the Article 52(1) notification date, the internal deployment or integration record, and the documentation index.

Does not settle it: the model name and a size comparison ("comparable in scale to GPT-5.6 Sol").

Evidence required: DOC.

D3. For each model in the chain: its classification under the provider's own preparedness or frontier-safety framework at the time of the evaluation, the safeguards that classification required, whether each was in place before the evaluation, and the decision record where one was not.

Establishes: whether the provider followed its own published framework.

Settles it: the classification record per model with date and approver, the framework's safeguard list for that level, and per-safeguard evidence of state before the evaluation, with the decision record for any exception.

Does not settle it: a statement of the current classification; "safeguards were disabled for capability measurement" without the framework's own exception record.

Evidence required: DOC.

Block E: Residual uncertainty and recurrence

Purpose: every block above asks the provider for facts and leaves the provider's confidence in those facts implicit. The four behavioural labels are descriptive categories rather than established root causes, and the public record does not establish motive, so the provider's own residual uncertainty is a documented gap. A congressional letter asked for "remaining unknowns" as the last of its 23 questions and the reply did not address it. The drafting risk is that self-declared uncertainty is the least falsifiable material an addressee can supply; the question is therefore anchored on records that already exist rather than on a narrative composed for the reply.

E1. The questions about the chain and the public-wiki activity that the provider's own investigation has not resolved, with the hypotheses considered and the evidence that would resolve each; and the provider's current assessment of the likelihood of a further event of the same kind,

the conditions under which that assessment would change, and the monitoring that would detect one. For both limbs, the contemporaneous internal document recording the uncertainty or the assessment, and its date.

Establishes: the provider's own epistemic position, in a form comparable with what it has asserted publicly, and how far the investigation is closed.

Settles it: the open-items or unresolved-hypotheses part of the investigation file and the residual-risk part of the provider's systemic-risk assessment, both dated and as they stood before this request, together with a signed statement that no other material unresolved question about these events is recorded elsewhere.

Does not settle it: a list of uncertainties composed for this reply with no internal record behind it; "we cannot rule out X" without the analysis that attempted to; a statement that the investigation remains "ongoing" without its open items; a recurrence assessment with no stated conditions under which it would change.

Evidence required: DOC and STMT.

Coverage check: gaps to questions

Every documented gap and disagreement in the public record that the addressee could answer is mapped to a question.

Incident start — A1

Primary motive — A10, E1

Escape versus intrusion — A4

Misalignment, cheating, reward hacking and collusion labels — A10, E1

Public-wiki provenance — C3

Wiki scale (15,000 versus 18,000) — C3; system and run identifiers would fix the units

Wiki relationship to the Hugging Face chain — C3

Complete transcripts, prompts, permissions, denominators and control states unavailable — A2, A3, A4, A9

Evidence not published in one aligned event set — A1

No proof of quarantine, remediation or absence of surviving copies — B1 to B7

"No harm identified" is not evidence of no harm — A5, B7

Selection effects in disclosed runs — A2, A8, C2

Summaries mix capability, scaffolding, benchmark, sandbox, third party and response — A3, A4

Denominator unclear — A2, C1

No public source on whether a formal report was made — C4, D1

Detection speed and false-positive budget — A7

Log preservation when agents can create misleading artifacts — A9, B2

Second-hop risk ownership — A6, C5

Containment standard after a failed rebuild — B5

Article 55 applicability facts, including whether a report was made and the classification record — D1, D2, C4

What a regulator could request without unnecessary sensitive detail — the dual-use rule above; A4 and B2 ask for categories and control state only

Prior warnings — A7

Monitorability of successor models — B8

Own-framework compliance — D3
Parallel proceedings — C4

Relationship to requests already sent

Public requests to the provider about this incident exist: two House letters with the provider's reply and a follow-up judging it insufficient, two Senate letters, a fifteen-state attorneys general preservation demand, and a state subpoena. This request overlaps with them by design; the contribution is the sufficiency standard and the Article 91 frame, not the novelty of the questions. Two questions are new against every instrument examined.

A1 — asked (timeline; complete timestamped timeline; the 12 July cessation; 13 to 20 July). Answered with a chronology from 8 July, with May and June treated as separate.
A2, C1, C2 — asked (counts, denominators, every incident since founding, all logs of breaches). Answered with an uncounted footnote; listed as unanswered in the 2 September follow-up.
A3 — asked (task prompt and scoring incentives, helpful-only variants, affordances). Prompt and incentives described, not released.
A5, A6 — asked (every network, account and credential accessed). "What became of the credentials and data" recorded as unanswered.
A7 — asked (signals, monitoring, halting point, prior warnings). Monitoring disconnections neither confirmed nor denied; the warnings question not answered.
A8, A9 — asked (reviewer access, the review agreement, preservation). Logs not released.
A10 — asked (behavioural characterisation). Answered with the four patterns as conclusions.
B1 to B5, B7 — asked (notes left for future versions; persistence). No public answer.
B6 — asked (remediation). Measures listed without dates or verification.
B8 — asked (successor monitorability). Reply not yet due at the time of writing.
C3 — asked (coordination sites). Reply not yet due at the time of writing.
C4 — asked (notifications, disclosure policies). Follow-up records that the reply named no authority.
C5 — asked (evaluation partners, advanced-exploitation evaluations). No public answer.
C6 — asked (the pause and the basis for resuming). Pause confirmed; basis for resumption not given.
D1, D2 — asked (models, internal deployment). Partly answered: two models named.
D3 — asked (framework classification and compliance, the Safety and Security Committee). Unanswered per the follow-up.
A4 — new. No public request asks for a per-boundary control-state table.
E1 — asked once, as "remaining unknowns", and not addressed in the reply. No requester states what a sufficient answer would be.

Appendix C. Limitations and Dual-Use Considerations

Evidence and scope

Everything in this work rests on public accounts by the incident parties and independent reviewers, and on primary legal texts only. The complete agent transcripts, prompts, tool

permissions, run denominators and control states are not public, and neither are the records behind any party's assurances so we do not rely on them.

Where prior work was not read in full, our claim that no existing instrument states a per-question sufficiency standard is scoped to the seven incident-specific instruments we did read, rather than asserted generally.

The evidence base closes on 12 September 2026. Several sources were revised after the first publication, and their update dates are preserved in the references.

Method and validation

The request was never sent and no regulator has seen it. Nothing here tests whether an addressee would answer, whether an answer meeting one of our criteria would close the gap it targets, or whether a regulator would find the question set proportionate.

The memos were produced by a legal analysis agent and were afterwards independently assessed by the team's legal research contributor.

The sufficiency vocabulary is ours. The AI Act sets no hierarchy of evidence, so our five kinds are a drafting device. The counts we report (five of 28 questions settleable only with third-party attestation, three requiring a signed officer statement, twenty closing on records or documents) describe our own assignment, not the law, and would move if any assignment changed.

Two results changed the work. We expected New York's RAISE Act to supply a regime without an evaluation carve-out, and found its 2026 amendments had imported California's. And the usability failures were not failures of writing: an entry failed because one limb of its question had no sufficiency criterion at all, and the same defect then appeared in three more.

What this work does not establish

This work offers no causal account of why the agents behaved as they did. The provider's four behavioural labels are used as descriptive categories and not as established root causes. It also establishes no legal conclusion: not that any obligation was breached, that any report was due, or that any model falls within scope. It does not establish that any regime captures this incident. On the public facts none of the three clearly does, and that is a reading of statutory texts and not a ruling.

We also do not claim these answers would prevent recurrence or that a regulator would obtain them. A researcher's draft is not a regulatory act, and naming a provider as addressee reflects where the evidence sits, not an allegation.

Dual use and information hazards

The risk of misuse is low. The artifact is a set of questions about evidence, records and governance. It contains no technique, and answering it would confer no capability.

A dual-use rule was applied to every question and is stated in the artifact itself: questions ask for categories of evidence, and for records showing the state of a control or the fact of an event. None asks for a configuration, a payload, a vulnerability detail, a credential or a hostname. Where an answer would be security-sensitive, or would identify a third party or a researcher's data, the instrument routes it to a confidential annex rather than the body of a reply.

No personal data was collected or republished. No real system was accessed, scanned or probed. No credential was tested. No part of the incident was reproduced. The benchmark involved was neither copied nor run.

Publication and disclosure

Disclosure level considered: low. Every input is a public document, and the work produces no novel vulnerability, no intrusion recipe and no non-public finding. Nothing was withheld for private review because nothing sensitive was produced.

Next steps

The natural next steps would be to send the public-facing version to the AI Office and CeSIA for comment on the work, tracking the provider's replies against each question and mark those that become answered.

The instrument could also be used to score the 31 August reply row by row against the sufficiency column, which we specified but did not complete.

LLM Usage Statement

Claude (Anthropic) agents were used throughout, as a helper for brainstorming and other tasks, but not as an author of the report. Agents drafted the request and the sufficiency table from the team's research base (in the annexes, with our edits), produced legal memos from the primary texts (each is labelled as not a licensed lawyer's advice), ran mechanical checks (link, source and claim-ID validation, question-set consistency), built the evidence map, and retrieved and transcribed the public documents cited.

We (the team) selected the project, wrote the research base's claim and source records, performed the usability read, answered the legal task list independently of the agent memos (the Legal research contributor), reviewed and merged every change, and wrote the text of this report. Every result and factual claim was checked by a team member against the cited source. No agent output is presented as a human legal opinion.