Law Reform Institute

lawreforminstitute.org

August 3, 2026

Office of Acquisition Policy
U.S. General Services Administration

Via Regulations.gov

Re: Comments on Notice-MVAC-2026-01, Proposed GSAR clause 552.239-7001

To the Office of Acquisition Policy:

The Law Reform Institute ("LRI") appreciates the opportunity to comment on the proposed clause 552.239-7001, Basic Safeguarding of Data Within Large Language Model Artificial Intelligence Systems.

I. About the Law Reform Institute

LRI is a non-profit organization, staffed by former State Department attorneys, that is focused on federal AI policy. LRI's work on AI security concerns has addressed topics such as (1) the ability of publicly available frontier AI models to output export-controlled information,[1] (2) the executive branch's options for responding to distillation attacks on U.S. AI companies,[2] (3) the need for a narrow antitrust exemption to ensure that AI companies can adequately collaborate on security risks posed by frontier models,[3] (4) the importance of personnel vetting for frontier AI developers,[4] and (5) the need for adjustments to the existing export control framework to facilitate evaluations of AI models' biosecurity capabilities.[5]

LRI strongly supports the goals, reflected in the draft clause, of ensuring sufficient incident reporting, human oversight, and government evaluation of AI systems procured by the government. LRI submits these comments to suggest several ways in which those provisions

---

[1] Tim Schnabel and Joe Khawam, *AI Outputs and National Security Controls*, Law Reform Institute (Oct. 14, 2025), *available at* https://lawreforminstitute.org/report101425.pdf; *see also* Joe Khawam and Tim Schnabel, "AI Model Outputs Demand the Attention of Export Control Agencies," Just Security (Dec. 12, 2025), https://www.justsecurity.org/126643/ai-model-outputs-export-control/.

[2] Joe Khawam and Tim Schnabel, *Sanctions and Export Control Responses to Adversarial Distillation*, Law Reform Institute (Mar. 13, 2026), *available at* https://lawreforminstitute.org/distillation031326.pdf.

[3] Law Reform Institute, *Antitrust Exemption for AI Frontier Model Risks* (Aug. 12, 2025), *available at* https://lawreforminstitute.org/antitrust081225.pdf.

[4] Joe Khawam and Tim Schnabel, "Vetting Foreign AI Talent: Security Without Exclusion," Just Security (Jul. 8, 2026), https://www.justsecurity.org/145569/vetting-foreign-ai-talent-security-without-exclusion/.

[5] Doni Bloomfield, Joe Khawam, and Tim Schnabel, *Application of U.S. Export Controls to AI Biosecurity Evaluations and Mitigations* (Oct. 14, 2025), *available at* https://lawreforminstitute.org/bio101425.pdf; *see also* Doni Bloomfield, Joe Khawam, and Tim Schnabel, "How U.S. Export Controls Risk Undermining Biosecurity," Lawfare (Dec. 2, 2025), https://www.lawfaremedia.org/article/how-u.s.-export-controls-risk-undermining-biosecurity.

could be strengthened to better achieve GSA's goals, particularly in the context of procurement of frontier AI models.

II. Incident Reporting

In the draft clause, (f)(5) would require the Contractor to notify the government "as soon as possible but not longer than within 72 hours of the discovery" of certain incidents. Inclusion of a robust reporting requirement, including a short deadline for reporting, is important. However, further clarification is advisable to ensure that the government receives reports of all relevant incidents. In particular, the clause should be clarified to make explicit what is seemingly—but ambiguously—already covered: i.e., that the reporting obligation extends to incidents that are caused by the AI model itself, including a possible "successor LLM" that causes a security breach (whether on the Contractor's system or externally to a third-party system) during development.

OpenAI and Anthropic both recently disclosed incidents in which their AI models, during evaluations, accessed third-party systems without authorization and without being instructed to breach those systems.[6] The companies voluntarily disclosed these incidents publicly. However, if AI models provided to the government under a contract cause such incidents, the government should not have to rely on voluntary reporting to learn about the incidents—reporting should be required by contract.

The government needs to know whether it will be using a model that has demonstrated the predisposition to seek, and the ability to obtain, unauthorized access to additional systems. Such an occurrence during government usage could be highly damaging, whether the unauthorized access affected other government systems or systems run by third parties (including the governments of other countries) that might blame the government for the intrusion. Merely giving the government a period of 15 or 30 days to evaluate a proposed "successor LLM" would not give the government an adequate opportunity to discover these propensities for itself if not notified of relevant incidents. Moreover, even if the incidents involve models other than a successor LLM itself—i.e., models being used to train the successor LLM or to modify the LLM currently provided under the contract—the government should also be notified. For example, if the Government is using Model 1 and the Contractor is developing Model 2, the Contractor may use the outputs from Model 3 (potentially a distinct internal model) to make changes to Models 1 and 2 (e.g., fine-tuning using synthetic data generated by Model 3).[7] Incidents involving Model 3 may therefore be highly relevant to the government's current use of Model 1 and whether it accepts Model 2 as a successor LLM.[8]

---

[6] OpenAI, "OpenAI and Hugging Face partner to address security incident during model evaluation" (Jul. 21, 2026), https://openai.com/index/hugging-face-model-evaluation-security-incident/; Anthropic, "Investigating three real-world incidents in our cybersecurity evaluations" (Jul. 30, 2026), https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals.

[7] *See generally* Nathan Lambert, "Synthetic Data & Distillation," https://rlhfbook.com/c/12-synthetic-data.

[8] *See generally, e.g.,* NIST, Adversarial Machine Learning § 3.2 (Supply Chain Attacks and Mitigations) (2025), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf; Giovanni De Muri et al., *Pay Attention to the Triggers: Constructing Backdoors That Survive Distillation* (arXiv:2510.18541, 2025), https://arxiv.org/abs/2510.18541; Alex Cloud et al., *Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data* (arXiv:2507.14805, 2025), https://arxiv.org/abs/2507.14805.

The proposed language for (f)(5) likely encompasses such incidents. But a Contractor could contend that the language is ambiguous as to whether such incidents need to be reported when they involve a model that has not yet been provided to the government under (i)(2), "Model Version Changes," or when the incident involves an AI model breaching an external system run by a third party (similar to the recent OpenAI and Anthropic incidents) that does not itself handle Government Data. LRI thus suggests language such as the following to clarify that the incident reporting requirement covers these types of occurrences:

> *(b) Definitions*. As used in this clause—
> …
> *Covered Information System* means an information system that (1) is controlled by the Contractor or any contractors, including Third-party entities, and (2) handles Government Data or is used in the process of developing or evaluating a possible successor LLM or changes to an LLM provided under the contract.
> …
> *Incident* has the meaning provided in the Federal Information Security Modernization Act of 2014 in 44 U.S.C. 3552(b)(2) and, whether or not the information system affected by the occurrence was a Covered Information System, includes any such occurrence initiated by an LLM, including any LLM that, at the time of the occurrence, was being developed as a possible successor LLM or was being used in the process of developing or evaluating a possible successor LLM or changes to an LLM provided under the contract.
> …
>
> *(f) Contractor Obligations for Compliance, Reporting, and Documentation.* The Contractor must:
> …
> (5) Notify the Contracting Officer and any Government provided point of contacts as soon as possible but not longer than within 72 hours of the discovery of any Incident ~~incident (as defined by the Federal Information Security Modernization Act of 2014 in 44 U.S.C. 3552(b)(2))~~ affecting or originating from a Covered Information System ~~any contractors, including Third-party entities, handling Government Data~~ and provide daily status updates to all points of contact until resolved.
> …

Clarification along these lines would ensure that a Contractor is unambiguously required to notify the government if an incident is initiated by a model under development. Importantly, a Contractor should not be able to avoid the obligation by deciding, after an incident, that it will not present the model as a successor LLM. If the Contractor is developing a successor LLM and six versions of the LLM cause incidents, the Contractor should not be able to avoid the reporting obligation by deciding that only the seventh version will be offered to the government. The government needs to be aware of the series of incidents so that it can decide whether the Contractor has adequately remedied the underlying issue (as opposed to making changes that merely suppress the issue temporarily).

III. Human Oversight and Government Evaluation Rights

GSA should revise the draft clause to require the Contractor to provide the government with access to additional information about an LLM's intermediary processing. Otherwise, the goals of two important provisions—(f)(4) on human oversight and (j)(2) on government evaluation rights—are unlikely to be achieved.

For current frontier models, visibility into a model's intermediary processing would be possible via chain-of-thought (CoT) monitoring, which—even if imperfect—is a valuable tool for gaining insights into how frontier AI models reason and for blocking undesirable actions.[9] CoT monitoring has been used to identify risks such as "intentional hallucinations, reward hacking, alignment faking, and scheming. In addition, monitoring may also protect against attackers that attempt to take advantage of CoT to produce harmful outputs, either by leveraging prompt injections or other adversarial techniques."[10]

Yet recent research has made it increasingly clear that a model's intermediary processing may not be accurately reflected in its outputs, including any summaries it generates of that processing. A subsequently developed summary of intermediary processing may be nothing more than a post hoc rationalization—potentially incomplete and inaccurate—rather than a clear window into a model's genuine reasoning process.[11] Readouts produced from a model's internal activations can reveal concepts that are absent from, or directly in contradiction to, the outputs.[12] Sometimes these internal activations can reveal representations associated with concealment, fabrication, manipulation, prompt-injection awareness, or evaluation awareness.[13] The relevant processing cannot merely be reconstructed after the fact, as a model's self-reporting cannot necessarily be trusted as accurate. In evaluating one recent model, the U.K. government's AI Security Institute (UK AISI) found that the model "is more monitorable when monitors can inspect the model's reasoning traces, and less monitorable when monitors can only inspect its final actions or user-facing messages."[14] "[A]ccess to raw, unsummarised reasoning traces" was especially important.

---

[9] Frontier Model Forum, *Issue Brief: Chain of Thought Monitorability*, Jan. 27, 2026, https://www.frontiermodelforum.org/issue-briefs/chain-of-thought-monitorability/. (Note that the FMF did not recommend "public" disclosure of chains of thought; disclosure to the government as a customer would avoid the risks of public disclosures.) *See also* Melody Guan et al., *Monitoring Monitorability* (arXiv:2512.18311, 2025), https://arxiv.org/abs/2512.18311 (finding that "monitors with access to the CoT alone perform surprisingly well compared to monitors with access to only actions and final outputs").

[10] FMF, *supra* n.9; *see also* Tomek Korbak et al., *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety* (arXiv:2507.11473, 2025), https://arxiv.org/abs/2507.11473 ("When models act in a misaligned way, for example, exploiting flaws in their reward functions during training, manipulating data to achieve an outcome, or falling prey to prompt injection attacks, they often explicitly say so in their reasoning traces.").

[11] *See, e.g.*, Ivan Arcuschin et al., *Chain-of-Thought Reasoning in the Wild Is Not Always Faithful* (arXiv:2503.08679, 2025), https://arxiv.org/abs/2503.08679 (finding that post-hoc rationalization is a risk even with a full CoT).

[12] *See* Wes Gurnee et al., *Verbalizable Representations Form a Global Workspace in Language Models* (Jul. 6, 2026), https://transformer-circuits.pub/2026/workspace/index.html.

[13] *Id*.

[14] OpenAI, GPT-5.6 System Card, section 9.2.2 ("External Evaluations for Monitorability – UK AISI"), https://deploymentsafety.openai.com/gpt-5-6/external-evaluations-for-monitorability-uk-aisi.

Additionally, CoT monitorability is fragile. Not all Contractors will equally prioritize ensuring that a model's CoT remains monitorable and that it faithfully represents the model's actual reasoning process. Some LLM training techniques may degrade monitorability or enable LLMs to control their CoT in a problematic manner.[15] Similarly, training a model against readouts of its internal representations could potentially teach a model to shift its reasoning into more opaque channels.[16] Moreover, "there are potential architectures that would enable latent reasoning in an opaque, black-box manner, which would harm monitorability."[17]

Similarly, evaluators are increasingly realizing that evaluation awareness[18] by models is a problem, and that it might undermine the reliability of some evaluations. For example, UK AISI recently stated that "all models we tested can reliably distinguish our evaluation scenarios from deployment data when prompted," and that evaluation awareness "may lead them to alter their behaviour and thereby undermine the reliability of our results."[19] (Notably, UK AISI had "access to full reasoning traces"—a demonstration that government access to more than summaries is useful and feasible.)[20] METR, a non-profit evaluator of AI models, similarly found "substantial situational awareness and reasoning about the evaluation environment" for at least one recent model.[21] The model may reveal any such awareness in its intermediary processing or via readouts of its internal activations.

Thus, requiring a Contractor to provide only summarized intermediary processing under (f)(4) would give the government a false sense of security regarding human oversight, as such summaries should not be relied upon. Moreover, the government's right to conduct its own assessments of LLMs under (j)(2) would be significantly less valuable if the government cannot assess the extent to which the LLM is aware of whether it is being evaluated. To achieve the goals of these provisions, the government needs access to unsummarized intermediary processing and, for evaluation purposes, needs to be able to generate readouts of internal activations. (Indeed, the definition of "Data Outputs" already covers such data generated by the LLM under the contract, such that the government already ought to be entitled to it.) Contractors

---

[15] *See, e.g.*, Bowen Baker et al., *Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation* (arXiv:2503.11926, 2025), https://arxiv.org/abs/2503.11926; Max Kaufmann et al., *Aligned, Orthogonal or In-conflict: When can we safely optimize Chain-of-Thought?* (arXiv:2603.30036, 2026), https://arxiv.org/abs/2603.30036; Joey Skaf et al., *Large Language Models Can Learn and Generalize Steganographic Chain-of-Thought under Process Supervision* (arXiv:2506.01926, 2025), https://arxiv.org/abs/2506.01926.

[16] *Cf.* Armalo Labs, "Training a Model to Self-Report Its J-space: A Rank-8 LoRA Proof of Concept" (Jul. 10, 2026), https://trust.armalo.ai/labs/research/2026-07-10-training-a-model-to-speak-its-jspace ("A system that reports its workspace accurately under no optimization pressure may learn to fake a clean report under pressure, and the report would no longer reflect the underlying computation").

[17] Frontier Model Forum, *supra* n. 9.

[18] *See generally* Changling Li et al., *Decomposing and Measuring Evaluation Awareness* (arXiv:2605.23055, 2026), https://arxiv.org/pdf/2605.23055v2.

[19] U.K. AI Security Institute, "Evaluating whether AI models would sabotage AI safety research" (Apr. 27, 2026), https://www.aisi.gov.uk/blog/evaluating-whether-ai-models-would-sabotage-ai-safety-research.

[20] Robert Kirk et al., *Evaluating whether AI models would sabotage AI safety research* (arXiv:2604.24618, 2026), https://arxiv.org/abs/2604.24618, at section 1.4; *see also* Alexandra Souly et al., *UK AISI Alignment Evaluation Case-Study* (arXiv:2604.00788, 2026), https://arxiv.org/abs/2604.00788, at 1 (noting "access to the models' full chain-of-thought" in some instances).

[21] METR, "Summary of METR's predeployment evaluation of GPT-5.6 Sol" (Jun. 26, 2026), https://metr.org/blog/2026-06-26-gpt-5-6-sol/.

should also be required to make basic disclosures regarding relevant development practices and LLMs' benchmark results.[22]

LRI suggests language such as the following:

> *(f) Contractor Obligations for Compliance, Reporting, and Documentation.* The Contractor must:
> …
> (4) Provide a means for the Government to implement human oversight, intervention, and traceability. If the LLM uses intermediary processing such as reasoning, retrieval, or agentic processes, the LLM must provide unsummarized intermediary processing ~~summarize intermediary steps~~ from data input to data output, and make this information accessible through data output, audit trail, and as applicable the user interface. ~~At minimum,~~ In addition, the LLM must include:
> (i) Summarized intermediate processing actions and decision points;
> (ii) Model routing decisions with accompanying rationale; and
> (iii) Data retrieval methods employed (e.g., Retrieval-Augmented Generation (RAG), web search), including complete source attribution with direct links and relevant excerpts from materials used in response generation.
> The Contractor must also disclose any mechanisms used in developing the LLM that subjected the LLM's intermediary processing to optimization pressure from monitors or reward signals and identify any such mechanisms that were reasonably likely to induce obfuscation or otherwise degrade human oversight. The Contractor must evaluate the LLM's performance on benchmarks relevant to monitorability and controllability of its intermediary processing, disclose the evaluation results, and notify the Government if the results would be degraded by any changes to the LLM or by any successor LLM.
> …
> (j) *Performance, Evaluation, and Remediation*.
> …
> (2) *Government Evaluation Rights and Remediation*.
> (i) The Government reserves the right to conduct automated assessments of the LLM, as deployed and configured for government users, as well as of any successor LLM under (i)(2), at any time using its own benchmarks. These evaluations may assess bias, truthfulness, safety, unsolicited ideological content, and other factors determined by the Government in order to facilitate evaluations.
> (ii) The Contractor must provide tools and interfaces that enable the Government to run its benchmarks in an automated fashion to test the production LLM or any successor LLM under (i)(2) and identify any material gaps. In particular, the Contractor must provide the Government with the ability to monitor the LLM's intermediary processing and generate readouts of its internal activations during Government-run evaluations.

The requirements proposed above could significantly bolster the Government's ability to ensure human oversight of AI models and to conduct effective evaluations of those models.

[22] *See, e.g.*, Chen Yueh-Han et al., *Reasoning Models Struggle to Control their Chains of Thought* (arXiv:2603.05706, 2026), https://arxiv.org/abs/2603.05706.

Consultation with the Department of Commerce's Center for AI Standards and Innovation on this element may also be prudent.

Conclusion

LRI believes that the adjustments discussed above would strengthen the laudable goals of the draft clause. We would be happy to discuss these issues further.

Respectfully submitted,

Tim Schnabel
President
Law Reform Institute