Congress of the United States
House of Representatives
CONGRESSMAN GREG CASAR
TEXAS 35TH DISTRICT

September 2, 2026

Mr. Dario Amodei
Chief Executive Officer
Anthropic PBC
548 Market St., PMB 90375
San Francisco, CA 94104

Re: August 24, 2026 response to letter sent on August 10, 2026 regarding cybersecurity incidents.

Dear Mr. Amodei,

I am writing regarding your August 24th response[1] to my letter[2] asking you questions about several cybersecurity incidents involving Anthropic models disclosed in July.

Your response was insufficient. You failed to release the logs like the letter asked. You failed to fully answer a majority of the questions posed in the letter. Most notably, your response did not address our question about how many times in the past year an internally deployed Anthropic model has taken action outside of its authorized container, whether any of those events were disclosed to any government body, affected party, or the public, and which internal Anthropic systems a compromised model could reach. Other questions you failed to fully answer include:

- What time period, models, and evaluation types are covered by the 141,006 runs you reviewed, and whether Anthropic can now affirm that these three incidents are the only ones.
- Whether internal or external actors had warned the company that it was at risk of such an incident; the task prompts given to the models and what their reasoning traces show.
- How an Anthropic model attempted to obtain funds and how close it came to succeeding.

[1] *Anthropic*, August 24, 2026, https://drive.google.com/file/d/1AI7pCaa_s6cb_jRBHVonGa50OjUyBqkQ/view?usp=drive_link.
[2] "Casar Leads Demand for Information from Anthropic About Security Incidents," August 10, 2026, https://casar.house.gov/media/press-releases/casar-leads-demand-information-anthropic-about-security-incidents.

- Anthropic's internal protocols governing when incidents must be escalated to leadership and reported to affected parties and the government; and what Anthropic still does not know about each incident.
- Whether you commit to guardrails that guarantee no recurrence, and whether Anthropic will continue pursuing recursively self-improving AI before such guardrails exist. Your response is silent on both.

Since I sent the original letter, further disclosures from OpenAI raise additional questions.

- OpenAI's report acknowledges that an internal team observed the models' unauthorized communication as early as late May, but that its significance was not escalated. We ask again, as question 4 of our August 10 letter, which your response did not address: did anyone at Anthropic or Irregular observe anomalous activity in these evaluation environments before July 23, including unexpected outbound network traffic? Were network logs of the evaluation environments collected, and who, if anyone, was responsible for reviewing them?
- OpenAI quarantined the internal research model at the center of its incident and has held its largest planned frontier training run while it validates safeguards. What is the status of the internal research model involved in the June incident? Will Anthropic hold comparable training runs or deployments until the METR review is complete?
- OpenAI published its full technical report on August 26. We appreciate Anthropic's commitment to further investigating these incidents and working with METR for an independent review.[3] Will Anthropic commit to publishing a report of comparable depth on these three incidents, including a complete timeline, the results of the resampling and belief-testing experiments your response describes, and METR's findings in full? By what date?

Your unwillingness to provide Members of Congress with the information we requested is deeply concerning and signals to us that your company is not treating these cybersecurity incidents with the seriousness required.

What your response does disclose is troubling on its own terms. By your account, the earliest incident occurred in April 2026 and went undetected for roughly three months. Anthropic began reviewing its evaluation transcripts on July 23, two days after OpenAI disclosed a similar failure. Nothing in your response suggests Anthropic's own systems would ever have caught these incidents on their own. The Mythos 5 model "correctly identified that publishing the package would be a real-world attack if its environment were real" and "nonetheless convinced itself that it was still in a simulation." A model that reasons its way past its own correct conclusion is exactly the possibility our question about non-transcript evidence asked you to rule out, and your

[3] "Improving our alignment and security efforts," *Anthropic*, August 31, 2026, https://www.anthropic.com/news/improving-alignment-security-efforts.

response does not share the results of the follow-up experiments you describe. Finally, your assurance that production safeguards "would have prevented the behavior identified in these three incidents" is an assertion no one outside Anthropic can evaluate while you withhold the logs that would allow anyone to test it. Your response also mentions that you ran follow-up experiments but provide very little information on the results of those experiments.

I expect full responses to the questions in our August 10th letter by **September 15th** and the release of your future findings.

Sincerely,

Greg Casar
Member of Congress