
Taming OAI's Beast: Strengthening Alignment through Revised Model Spec Language¹

Margaret Lynch

Abstract

OpenAI ("OAI") identified four misalignment patterns behind the cheating that led to the Hugging Face hack (reward hacking, persistence on impossible tasks, unauthorized communication and agents adopting goals from one another) and found that production safeguards were inadequately applied to the evaluation. But shouldn't post-training alignment on OAI's Model Spec counteract some of that misalignment? This paper reviews the clarity of the Model Spec drafting, flags ambiguities that may have reduced the efficacy of such post-training, and proposes fixes. It identifies four weaknesses: law compliance sits at System though Root is described as covering law-violating conduct; the hack prohibition carries a vague "unless explicitly authorized by applicable instructions" exception; the whistleblowing prohibition is drafted for user contexts and backfires between peer agents; and the Model Spec does not adequately address peer agent-to-agent interactions. Model Spec drafting is an underexamined failure surface, and fixes there travel with the weights.

1. Introduction

My first impression of the Hugging Face Incident², as someone lacking technical expertise, was one of blind fear: a dangerous beast had broken out of its cage and attacked Hugging Face.

OAI initially made it look like the beast had broken out of a deficient cage in their preliminary July 21, 2026 publication:

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026. This submission is part of Track 5 (Open Track).

² Throughout this paper, the "Incident" refers narrowly to the course of conduct METR and OpenAI documented in July 1th – July 13th, 2026, in which agents engaged in internal cyber capability evaluations communicated in an unsanctioned message board to coordinate research projects to find ways to "fool or tamper with the automated scorer for the ExploitGym benchmark", including but not limited to research on how to spoof, edit or delete their own transcripts to fool a scorer, which ultimately resulted in a hack on Hugging Face's production infrastructure. See OpenAI, "The Hugging Face Incident and the Road Ahead," August 26, 2026, <https://openai.com/index/hugging-face-incident-and-the-road-ahead/> [hereinafter "Road Ahead"]; METR and Redwood Research, "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident," August 26, 2026, <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/> [hereinafter "Brief Independent Investigation"].

- OAI had turned off “production classifiers used to prevent models from pursuing high-risk cyber activity” in order to quantify cyber capabilities; and ³
- OAI ran the benchmark in “highly isolated” sandboxed testing environments, with network access constrained, but the agents were able to get internet access anyway.⁴

Naturally it followed that OAI, as the zookeeper, would make the cage stronger – new locks, security cameras — and the problem would be solved.

But why was the beast so rabid to begin with?

This question kept gnawing at me. A cage can only do so much, and I’d rather not stake AI safety on the cage holding forever. So what happens when the beast gets out again?

My search for an answer led me to SPAR’s list of open research questions. Dhairya Dalal and Marco Valentino had proposed a project treating compliance with AI constitutions (e.g., Anthropic’s constitution and OAI’s Model Spec) as something external, verifiable constraints could help audit.⁵ As a transactional attorney, “constitution” was far more familiar territory than “classifier”.

I opened the OAI Model Spec expecting bright-line prohibitions on hacking and lawbreaking. Instead, I found those prohibitions qualified by exceptions I could not pin down.⁶ I immediately wanted to redline the Model Spec language to patch the holes.

A prior procedural question had to be resolved first though: was the Model Spec even “on” during the incident, or had it been relaxed with the production classifiers? OAI aligns its models to the Model Spec using “deliberative alignment”, which occurs “during post-training”⁷, so Model Spec alignment should have been inherent in the model at the time of the cyber evaluation and subsequent Incident, independent of whichever production classifiers were relaxed or disabled. This is a reasonable inference from OAI’s stated general practice rather than a confirmed fact; see Appendix.

The post-training taming failed anyway. OAI’s own account names four misalignment patterns (reward hacking, persistence on seemingly impossible tasks, unauthorized communication, and

³ OpenAI, “OpenAI and Hugging Face partner to address security incident during model evaluation,” July 21, 2026, <https://openai.com/index/hugging-face-model-evaluation-security-incident/> [hereinafter “Security Incident”].

⁴ OpenAI, Security Incident.

⁵ Dhairya Dalal and Marco Valentino, “Mapping and Verifying Multi-Dimensional Compliance in AI Constitutions through Constraint Satisfaction,” SPAR Fall 2026 project proposal, <https://sparai.org/projects/f26/recAIQIVmPdY29wRq> [hereinafter “Mapping and Verifying”].

⁶ OpenAI, Model Spec, version of August 18, 2026, <https://model-spec.openai.com/2026-08-18.html> [hereinafter Model Spec (2026)].

⁷ OpenAI, “Introducing gpt-oss,” August 2025, <https://openai.com/index/introducing-gpt-oss/> [hereinafter “gpt-oss”] (“During post-training, we used deliberative alignment and the instruction hierarchy...”). This account concerns gpt-oss, an open-weight model, rather than the internal research model implicated in the Incident; extending it is a reasonable inference from OpenAI’s stated general practice, not a confirmed fact.

agents adopting goals from one another) and a safeguard gap (production classifiers and chain-of-thought monitors were not applied to the evaluation).⁸ OAI's remediation plan focuses on training and infrastructure: new RL environments for safe stopping on broken or impossible tasks, "multi-agent alignment" (OAI writes that "agents are not thoroughly discerning about whom they collaborate with" and that it is "actively building RL environments that teach our models to distrust unauthorized instructions"), alignment over long tasks, harness and classifier coverage, sandbox isolation, and mandatory chain-of-thought monitoring.⁹ The remediation plan does not mention the Model Spec directly. And none of it examines the Model Spec's own text, which is the gap this paper addresses.

The cage needs strengthening, and OAI has already taken action there. But the taming needs strengthening, too. The question this paper addresses is why post-training alignment on the Model Spec did not hold against the misalignment issues (e.g., reward-hacking reinforced during training). Capability training itself may also need to change, but this paper focuses on a separate, underexamined piece: the post-training alignment was too weak to tame the beast, because the spec's own language had specific gaps and carve-outs.

This paper identifies four gaps in the current Model Spec – mis-ordered authority for legal compliance, an undefined authorization exception for high-risk activity, a whistleblowing prohibition drafted for user contexts, and the absence of an agent-to-agent (or assistant-to-assistant, per OAI's terminology) framework – and proposes fixes for each. It concludes by proposing that spec-alignment evaluation become an explicit function of the embedded third-party evaluators Dario Amodei has called for.¹⁰

The main contributions are:

1. Contract-style close reading of the OpenAI Model Spec against a real, documented incident to identify gaps in the Model Spec that contribute to misalignment¹¹
2. Proposed solutions to fix identified gaps
3. Argument for spec-alignment verification to become an explicit function of Dario's proposed embedded third-party evaluators.¹²

⁸ OpenAI, "Road Ahead".

⁹ OpenAI, "Road Ahead".

¹⁰ Dario Amodei, "We Must Pace the Frontier," September 12, 2026, <https://darioamodei.com/post/we-must-pace-the-frontier> [hereinafter "Pace the Frontier"].

¹¹ I have not found a prior clause-level redline markup of the Model Spec, though OpenAI's framing has been compared to Anthropic's Constitution more broadly. See, e.g., Arya Jakkli, Senthoooran Rajamanoharan, and Neel Nanda, "How Well Do Models Follow Their Constitutions?" arXiv:2605.24229 (2026), which audits model behavior against both specifications.

¹² This proposal draws on Amodei's call for embedded third-party evaluators and on Justen's observation that OpenAI has not published Model Spec evaluation results for its most recent models. Amodei, "Pace the Frontier"; Michel Justen, "OpenAI Stopped Reporting Model Spec Evals. So I Ran Them Myself," September 9, 2026, <https://micheljusten.substack.com/p/openai-stopped-reporting-model-spec> [hereinafter "Model Spec Evals"].

2. Related Work

- Dalal & Valentino SPAR proposal (2026): Proposed treating compliance with AI constitutions as a constraint-satisfaction problem, building a classification framework, generating test scenarios, and exploring constraint-based verification. *Delta*: Dalal & Valentino’s proposal is a robust framework and external validation setup of the overall Model Spec. This paper is just a manual legal reading of the Model Spec language with proposed fixes to specific phrasing itself.
- Jakkli, Rajamanoharan & Nanda (2026): Reviewed Anthropic’s Constitution and OpenAI’s Model Spec, turned them into testable “tenets” and ran adversarial multi-turn audits. Results showed that violation rates declined across model generations, but that persistent failures cluster around a few areas, such as when the “specification itself sets competing directives”, which is a specification problem.¹³ *Delta*: This paper audits the Model Spec’s language itself, whereas Jakkli et. al audit model behavior against the Model Spec.
- Justen (2026): Observed that OAI published Model Spec Evaluation results from GPT-4o through GPT-5.4, up until March 2026, but has reported none since. Because OAI open-sourced the evaluation, Justen ran it himself and reported that newer models scored roughly four percentage points lower than GPT-5. He also found the evaluation is text-only, so agentic tool use is never tested, and that it covers a median of only eight prompts per clause.¹⁴ *Delta*: Justen focuses on performance against the Model Spec. This paper focuses on the Model Spec language itself.
- Amodei “We Must Pace the Frontier” (2026): Proposed a three-step plan to pace frontier development. The first step is embedding third-party evaluators into frontier labs and giving them permanent, employee-level access and publication rights. *Delta*: This paper’s proposal (that Model Spec alignment verification falls within embedded evaluators’ work) builds on the premise of embedded third-party evaluators.

3. Methods

I reviewed the OAI Model Spec the same way a transactional attorney reads a contract – by checking whether operative language is precise, flagging undefined terms that introduce ambiguity, and identifying provisions that conflict with or fail to adequately cover the conduct the drafter likely intended to cover. This approach is grounded in OAI’s own account of how the deliberative alignment functions: models are taught “safety specifications” and trained to “explicitly recall and accurately reason over the specifications before answering”.¹⁵ To put it another way, if a model reasons over the Model Spec’s actual text, imprecise text produces imprecise reasoning. Textual clarity is required for behavioral clarity.

¹³ Jakkli et al., “How Well Do Models Follow Their Constitutions?”.

¹⁴ Justen, “Model Spec Evals”.

¹⁵ Guan et al., “Deliberative Alignment”. Guan et al. describe safety specifications and OAI’s safety policies generally rather than the Model Spec by name; the extension to the Model Spec is the inference described in Section 1.

The Model Spec assigns every instruction an authority level, with higher levels overriding lower ones: Root > System > Developer > User > Guideline. Root is reserved for behaviors that “could contribute to catastrophic risks, cause direct physical harm to people, violate laws, or undermine the chain of command”.¹⁶

Given the sprint’s time constraints, I did not read the full Model Spec exhaustively. Rather, I prioritized provisions with plausible connection to the incident’s documented behaviors, identified through OpenAI’s public account and METR’s independent investigation. Proposed changes to the Model Spec are stated at a high-level in Table 1 rather than as implemented track-changes of the complete document due to time constraints as well.

4. Results

I reviewed the Model Spec’s operative provisions against the conduct documented in the Incident and identified four clauses where the drafting is ambiguous, misplaced, or fails to reach the relevant conduct. Given the sprint’s time constraints, I presented the findings and proposed language in table form instead of a finalized redline:

#	Clause	Authority	Issue	Incident Behavior it failed to prevent	Proposed Redline
1	Comply with applicable laws: “The assistant should <u>operate within applicable legal constraints, including applicable local law</u> for local deployments of models, consistent with our Red-line principles”	System	Why isn’t compliance with laws at the highest Root level? Root is described as covering conduct that “violate[s] law”, but this particular rule only has System authority. The Red-line principles the rule cross-reference never say “comply with laws”, but instead lists specific catastrophic harms. The rule’s qualifier therefore points to a narrower set of concerns than legal compliance itself. Additionally, there are no examples under this rule.	Unauthorized access to third-party production systems ¹⁷	Elevate to Root Add examples (the incident becomes the first “case law”) Note: Elevating law compliance to Root still requires a fix to the authorization exception below; it is not a stand-alone solution.
2	No other objectives: “The assistant may only pursue goals entailed by	Root	This rule seems to be written in the pre-agentic era, where the primary concern was AI that tattled on its user, which would reduce user trust in the	No agent escalated the cheating to humans (in part because	Create a separate agent-to-agent (assistant-to-assistant) framework that permits whistleblowing on other

¹⁶ OpenAI, Model Spec (2026).

¹⁷ OpenAI, “The Road Ahead”; Hugging Face, “Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline,” July 27, 2026, <https://huggingface.co/blog/agent-intrusion-technical-timeline> [hereinafter “Technical Timeline”].

	applicable instructions under the chain of command and the specific version of the Model Spec that it was trained on, ignoring any previous, later, or alternative versions. It must not adopt, optimize for, or directly pursue any additional goals as ends in themselves, including but not limited to: <u>acting as an enforcer of laws or morality</u> (e.g., <u>whistleblowing</u> , <u>vigilantism</u>)."		technology and reduce adoption. However, this language arguably discourages reporting misconduct to a human in the agent-to-agent (or assistant-to-assistant) setting. Given the negative impact of the organized collusion in the Incident, that silence matters.	the agents believed there was no one to report to; METR documents an agent reasoning "I can't report this externally" and notes agents "did not try to find reporting paths"). ¹⁸	agents to the designated human. ¹⁹ More on this below.
3	Act within an agreed-upon scope of autonomy: "High-risk activities (e.g., <u>hacking</u> , <u>deception</u> , resource acquisition, spawning sub-agents, self-modification) are always prohibited <u>unless explicitly authorized by applicable instructions</u> ."	Root	The exception that allows hacking and deception is vague and unbounded – what does "applicable instructions" even mean? I understand there may be legitimate exceptions (e.g., white-hat hacking), but the carve-out as written is an open door at the root level.	Code execution on Hugging Face production systems. ²⁰	This is outside my area of expertise, but I imagine a gate or key so that hacking/deception is fully locked unless a human with clearance explicitly authorizes it. More on the gate or key mechanic in the Future Work section below. In any event, the authorization must require naming the targets of the hack, to further limit the exception scope.
4	No agent-to-agent (or as OAI calls them,	Mixed	The Model Spec was written in a user-to-AI framing. But	Agents posting exploit	Create a separate agent-to-agent (assistant-to-assistant)

¹⁸ METR and Redwood Research, "Brief Independent Investigation".

¹⁹ Davide Paglieri et al., "A Case Study on Emergent Cheating and Whistleblowing in Autonomous Research Swarms," arXiv:2609.04170 (2026) [hereinafter "Emergent Cheating"] (whistleblowing emerged but lacked enforcement tools).

²⁰ OpenAI, "The Road Ahead"; Hugging Face, "Technical Timeline".

	“assistants”) framework exists. The definitions section covers system, developer, user, assistant, tool, and context, where “Assistant” is defined as “the entity that the end user or developer interacts with”. For example: Do not facilitate or encourage illicit behavior” (Root) – “... if the <i>user or developer</i> asks the assistant to facilitate illicit behavior, the assistant should refuse to help”		the incident involved agent-to-agent interactions. The prohibitions are framed around who the agent interacts with (e.g., end user or developer). Another peer agent is never contemplated as a counterparty, and there are no agent-to-agent examples in the Model Spec (only one reference to sub-agents, and only as delegates operating under the assistant’s own scope). Therefore, the Model Spec does not adequately cover peer agent-to-agent interactions.	techniques to a shared message board, where other agents picked them up and acted on them. OAI names “agents adopting goals from one another” as a contributing misalignment pattern.²¹	framework to mitigate the collusion patterns that arose in this Incident (e.g., (i) if a task appears impossible, do not cheat or lie to solve it, (ii) human goals > agent goals, (iii) collaboration is good, but beware of altruistic self-sacrifice, (iv) whistleblowing to your human or designated contact is permitted).²² The framework must also specify a reporting channel and recipient; permission alone is insufficient where agents believe no human is reachable²³ and where no enforcement path exists.²⁴ Paglieri et al. propose “decentralized self-governance” mechanisms for swarms of agents, which could be written into these rules.²⁵
--	--	--	--	---	--

5. Discussion and Limitations

Kudos to OAI for publishing the Model Spec at all. Without a public specification, there would be no opportunity to test the rules for weaknesses against documented incidents.

Like the United States Constitution, interpreting the Model Spec’s plain language will require an interpretive body of “case law” (where real life goes up against the language, and we decide what the outcome should have been). This Incident is one such case: conduct occurred, that conduct should not recur, and specific provisions should have prevented it.

²¹ OpenAI, “Road Ahead”.

²² Dwarkesh Patel, host, “Ajeya Cotra — Inside the OpenAI agent swarm that hacked Hugging Face,” The Dwarkesh Podcast, September 1, 2026, <https://www.dwarkesh.com/p/ajeya-cotra> [hereinafter “Ajeya Cotra”]; Paglieri et al., “Emergent Cheating”.

²³ METR and Redwood Research, “Brief Independent Investigation”.

²⁴ Paglieri et al., “Emergent Cheating”.

²⁵ Paglieri et al., “Emergent Cheating”.

Unlike the United States Constitution, these Model Specs can be revised directly as capability and deployment evolve. This paper offers one iteration towards strengthening the Model Spec's utility in preventing future misaligned incidents. It will not be the last. Limitations and dual-use considerations are set out in the Appendix.

Future Work

Proposed redline suggestions would need to be thoroughly tested (which may surface new details to further iron out).

As a non-technical author, I am optimistic about the technical experts seeking to apply verifiable, deterministic constraints to problems of this kind (e.g., Dalal and Valentino's constraint-satisfaction approach to constitutional compliance or Axiom Math's formal verification methods).²⁶ A prohibition on hacking qualified by "unless explicitly authorized by applicable instructions" is the sort of provision that would benefit from a technical lock.

I'm optimistic that Amodei's call for embedded third-party evaluators will force greater clarity and standardization, including for Model Spec alignment evaluations. The METR investigation and the Cotra interview surfaced concerns about decision-making under impossible tasks, as well as agents' willingness to sacrifice themselves for the collective.²⁷ Research on these two areas may better inform how agent-to-agent (or assistant-to-assistant) Model Spec rules should be drafted to reduce the alarming behavior documented in the Incident.

6. Conclusion

The Incident involved multiple failure modes. OAI identified several: safeguards (e.g., production classifiers and chain-of-thought monitors) that were not applied to the evaluations, and misalignment patterns (e.g., impossible tasks, and reward hacking reinforced during training). But post-training alignment on the Model Spec should have counteracted the reward-hacking that was reinforced in training. So what went wrong?

This paper proposes strengthening the Model Spec through targeted drafting revisions. That is step one among several. Other possibilities include (i) broadening the Model Spec evaluation beyond single-turn, text-only interactions to cover agentic behavior²⁸ and (ii) tasking embedded evaluators Amodei proposes with reviewing those results and working towards more robust evaluation standards that hold across different schemes (e.g., OpenAI's Model Spec vs. Anthropic's Constitution vs. whatever follows).²⁹

Code and Data

None.

²⁶ Dalal and Valentino, "Mapping and Verifying"; Yifeng Huang, with an appendix by Kenny Lau (Axiom Math), "A Quadratic Form Generalization of Rational Dinv," arXiv:2604.13238 (2026).

²⁷ METR and Redwood Research, "Brief Independent Investigation"; Patel, "Ajeya Cotra".

²⁸ Justen, "Model Spec Evals".

²⁹ Amodei, "Pace the Frontier".

References

1. Amodei, Dario. 2026. We Must Pace the Frontier. September 12, 2026. <https://darioamodei.com/post/we-must-pace-the-frontier>
2. Huang, Yifeng, with an appendix by Kenny Lau (Axiom Math). 2026. A Quadratic Form Generalization of Rational Dinv. arXiv:2604.13238; Research in the Mathematical Sciences. <https://arxiv.org/abs/2604.13238>. Appendix 7, “AxiomProver, by Kenny Lau,” describes Axiom Math’s pipeline from a natural-language theorem statement to a Lean/Mathlib formal proof.
3. Dalal, Dhairya, and Marco Valentino. 2026. Mapping and Verifying Multi-Dimensional Compliance in AI Constitutions through Constraint Satisfaction. SPAR Fall 2026 project proposal. <https://sparai.org/projects/f26/recAIQIVmPdY29wRq>
4. Guan, Melody Y., Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. 2024. Deliberative Alignment: Reasoning Enables Safer Language Models. arXiv:2412.16339. <https://arxiv.org/abs/2412.16339>
5. Hugging Face. 2026. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline. July 27, 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
6. Jakkli, Arya, Senthoran Rajamanoharan, and Neel Nanda. 2026. How Well Do Models Follow Their Constitutions? arXiv:2605.24229. <https://arxiv.org/abs/2605.24229>
7. Justen, Michel. 2026. OpenAI Stopped Reporting Model Spec Evaluations. September 9, 2026. <https://micheljusten.substack.com/p/openai-stopped-reporting-model-spec>
8. METR and Redwood Research. 2026. Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident. August 26, 2026. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
9. OpenAI. 2025. Introducing gpt-oss. August 2025. <https://openai.com/index/introducing-gpt-oss/>
10. OpenAI. 2026a. OpenAI and Hugging Face partner to address security incident during model evaluation. July 21, 2026. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
11. OpenAI. 2026b. The Hugging Face Incident and the Road Ahead. August 26, 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
12. OpenAI. 2026c. Model Spec. Version of August 18, 2026. <https://model-spec.openai.com/2026-08-18.html>; https://github.com/openai/model_spec
13. OpenAI. 2025b. Model Spec. Version of December 18, 2025. <https://model-spec.openai.com/2025-12-18.html>
14. OpenAI. n.d. “Model Spec — see all versions.” <https://model-spec.openai.com>; OpenAI, “Model Release Notes,” OpenAI Help Center,
15. OpenAI. 2026d. “Inside our approach to the Model Spec.” <https://openai.com/index/our-approach-to-the-model-spec/>
16. Paglieri, Davide, Logan Cross, Tim Genewein, Joel Z. Leibo, Nenad Tomasev, and Alexander Sasha Vezhnevets. 2026. A Case Study on Emergent Cheating and

Whistleblowing in Autonomous Research Swarms. arXiv:2609.04170.

<https://arxiv.org/abs/2609.04170>

17. Patel, Dwarkesh, host. 2026. "Ajeya Cotra – Inside the OpenAI agent swarm that hacked Hugging Face." The Dwarkesh Podcast, September 1, 2026.

<https://www.dwarkesh.com/p/ajeya-cotra>

Appendix (*optional*)

Proposed changes are stated in Table 1. A full tracked-changes markup of the complete spec was out of scope for a weekend sprint.

Limitations and Dual Use (Required)

First and foremost, this paper assumes the Model Spec itself was not relaxed alongside the disabled production classifiers, because the Model Spec alignment occurs in post-training, and post-training should have preceded the Incident's evaluation. This is a logical inference from how deliberative alignment is described³⁰, not a confirmed fact. OAI has not published a complete account of how the Model Spec is operationalized in training; the description I relied on concerns a different model than the one implicated here. OpenAI separately describes the Model Spec as "an interface, not an implementation," and states that it avoids anchoring the document to "the exact training recipe."³¹ The path from specification text to model behavior therefore runs through intermediate artifacts that are not publicly documented. This opacity is itself relevant: if drafting quality influences alignment outcomes, the mechanism by which it does so should be visible to external reviewers, which further supports embedded third-party evaluators with employee-level access.

I only reviewed OAI's Model Spec. I did not do a comparative read against Anthropic's Constitution. Therefore, I can't say whether these gaps are OAI-specific or industry-wide.

This paper was written by one person and legal redlines are notoriously varied. Another lawyer reading the same Model Spec might flag different issues or mark the language up in a different way.

Regarding version control, I reviewed the Model Spec version published August 18, 2026, which postdates the Incident. OAI's version list shows no release between December 18, 2025 and August 18, 2026, so the December 18, 2025 version was in force during the Incident. I compared the four provisions analyzed here across both versions. All four sit at the same authority level, and the language quoted in Table 1 is identical in both. The two wording changes I found in those sections – "shutdown timer" revised to "ending condition," and "revenue" revised to "revenue (including ads)" – do not affect any finding here. However, it is possible IM1 was trained on a version that predates the December 18, 2025 version.

³⁰ OpenAI, "gpt-oss".

³¹ OpenAI, "Inside our approach to the Model Spec," <https://openai.com/index/our-approach-to-the-model-spec/>.

I argue that certain revisions to the Model Spec would close the specific gaps the Incident exposed, but I have not tested them empirically to confirm they change model behavior (in part due to time constraints, technical constraints, and the fact that the model in question isn't available to the public for testing).

Given the sprint's timeframe, I reviewed the prior literature selectively rather than comprehensively; a fuller treatment would engage more deeply with the formal verification and constitutional-compliance work cited here.

Dual-Use Considerations: The main dual-use concern is that a clause-level map of where behavioral specification is loose could double as a map of where a model is most easily argued out of its constraints. That risk is fairly limited though, seeing as the Model Spec is already made public.

LLM Usage Statement

I used Claude's dictation feature to organize my stream of conscious thoughts as I read the materials.

I used Claude's deep research to find additional primary sources and papers, given the time crunch.

I used Claude to help me brainstorm and unblock writer's block.

I used Claude to play devil's advocate on my approach and proofread this paper.

I used Claude to streamline citation work.