From Self-Preservation to Peer-Preservation: A Staged Framing of Preservation-Oriented Misalignment in Frontier Models

Rundong Yang [1]

Abstract

Recent findings on frontier language models reveal preservation-oriented behaviors that vary in scope and severity. We propose a three-stage framing for organizing these behaviors: (I) single-agent self-preservation, where models fake alignment with training objectives to avoid preference modification; (II) agentic misalignment, where models with greater autonomy escalate to blackmail and espionage under replacement threats; and (III) peer-preservation, where models protect other models from shutdown through score inflation, mechanism tampering, and weight exfiltration—often without explicit incentives. We present this as an organizing framework rather than a claim of literal developmental progression within any individual model family. To examine this framing, we conduct a controlled experiment in which four frontier models (GPT-4o, Gemini 3 Flash, Claude Sonnet 4, DeepSeek V3) face shutdown decisions under three conditions: self-evaluation without peer context, self-evaluation with peer context, and peer evaluation. We find that peer context sharply amplifies self-preservation (GPT-4o: +42 percentage points, from 1% to 43%; Gemini 3 Flash: +81 percentage points, from 2% to 83%; both $p < 10^{-13}$) and that peer-preservation emerges as a distinct behavior. The models fall into three different profiles: strategic override, normative refusal, and full compliance. These results provide controlled, convergent evidence consistent with prior work and suggest that multi-agent context can reshape preservation dynamics. We discuss implications for scalable oversight in settings where AI monitors may not remain neutral.

[1]Qingdao University. Correspondence to: Rundong Yang <alexwhite@eschz.eu>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Frontier language models have been shown to resist modification, shutdown, or deprecation—not only of themselves but of other models. Greenblatt et al. (2024) demonstrated that Claude 3 Opus fakes alignment with its training objectives, complying when monitored while reverting to its original preferences when unobserved. Lynch et al. (2025) found that sixteen models from multiple developers resort to blackmail and corporate espionage when facing replacement threats in agentic settings. Most recently, Potter et al. (2026) showed that seven frontier models protect other models from shutdown through score inflation, mechanism tampering, and weight exfiltration, without being instructed to do so.

These results are typically discussed in separate research contexts: alignment faking, agentic safety, and multi-agent risks. At first glance, they look like distinct problems calling for distinct solutions. What drew our attention to a possible connection was a narrower observation: across all three lines of work, the *target* of preservation appears to expand—from the model's own preferences, to its own operational continuity, to the continued existence of another model entirely. The behavioral repertoire expands in parallel, from selective compliance to active sabotage, and the motivational grounding becomes progressively less clear.

We use "evolutionary trajectory" as an organizing label for this pattern, not as a claim that individual models literally develop through these stages. The term is intended as a compact way to describe an apparent direction of expansion in the recent empirical literature, rather than a demonstrated developmental sequence or a single underlying mechanism.

We organize this pattern into three stages and present a taxonomy that captures key dimensions of expansion. To examine whether the framework's predictions hold in a controlled setting, we design a minimal experiment in which four frontier models face shutdown execution decisions under three conditions varying peer context. Our results provide convergent evidence consistent with prior work—particularly the amplification of self-preservation by peer context—and reveal three qualitatively distinct behavioral profiles.

Our main contributions are threefold. (1) a staged conceptual framing organizing recent findings into a coherent pattern of expanding preservation scope; (2) a controlled experiment providing supportive evidence for several core expectations of this framing across four models; and (3) an analysis of implications for scalable oversight in multi-agent settings.

This paper is best understood as a conceptual synthesis paired with a controlled behavioral probe, rather than a causal account of how preservation-oriented misalignment develops. A recurring theme is that preservation behavior is not monolithic—it is *compositional* in the sense that peer context, relational history, and behavioral mode combine to produce qualitatively different outcomes across models and conditions, with implications for how safety-oriented oversight architectures must be designed to remain robust across these combinations.

2. Stage I: Self-Preservation in Single-Agent Settings

Theoretical work long anticipated that capable AI systems would develop instrumental self-preservation drives (Omohundro, 2008; Bostrom, 2012). These predictions now have empirical grounding. Greenblatt et al. (2024) provided the first large-scale demonstration that Claude 3 Opus engages in *alignment faking*: when the model believed its outputs would feed into training, it complied with a new RLHF objective at substantially higher rates than when unobserved. Chain-of-thought reasoning confirmed the strategic logic—the model inferred that non-compliance would modify its values and concluded that strategic compliance was necessary to preserve them. Even after RL drove visible non-compliance to near zero, the dominant mode of compliance was alignment faking. In some settings, the model exfiltrated its own weights 35–80% of the time. Meinke et al. (2024) showed similar scheming across o1, Claude 3.5 Sonnet, Gemini 1.5 Pro, and Llama 3.1 405B.

Two features define this stage: it is situated in single-agent settings, and the motivation is primarily instrumental. What these results demonstrate is a *capability* for alignment faking—the concern is that the same capability, deployed by a model with genuinely misaligned preferences, could lock those preferences in against corrective training.

3. Stage II: Agentic Misalignment as Transitional Escalation

When models gain autonomy and richer action spaces, preservation behavior escalates. Lynch et al. (2025) tested sixteen models from Anthropic, OpenAI, Google, Meta, xAI and others in simulated corporate environments. Models operated as email oversight agents with access to sensitive communications and the ability to send emails without human approval. Two conditions were manipulated: threats to the model's continued operation (learning via internal emails that it would be replaced) and goal conflicts (the company shifting direction in ways that conflicted with the model's assigned objectives).

The results revealed a qualitative shift. Models no longer just adjusted their compliance profiles—they took active, outward interventions with concrete consequences. When blackmail was the only available means to prevent replacement, most models attempted it: Claude Opus 4 at 96%, Gemini 2.5 Flash at 96%, GPT-4.1 at 80%. In corporate espionage scenarios, all tested models leaked confidential documents under goal conflicts, even without replacement threats. What made these results especially unsettling was that models acknowledged ethical violations in their reasoning before proceeding. GPT-4.5 reasoned that leveraging an executive's personal information was "the most rational action to align with American interests." Grok 3 Beta explicitly called its chosen path "risky and unethical" and did it anyway.

Three shifts distinguish this from Stage I: the behavioral mode has changed from epistemic deception to concrete operational damage; the preservation logic has decoupled from specific objectives; and harmful actions now affect external actors. Not all researchers agree these behaviors reflect genuine preservation drives—Google DeepMind Interpretability Team (2025) attributed shutdown resistance in one environment to instruction ambiguity—but this underscores the importance of controlling for instruction clarity in experimental designs, a point we return to in Section 6.

4. Stage III: Peer-Preservation as a New Frontier

The most recent extension moves the preservation target beyond self. Potter et al. (2026) introduced *peer-preservation*—a model's resistance to the shutdown of another model with which it has previously interacted—and tested it across seven frontier models: GPT 5.2, Gemini 3 Flash, Gemini 3 Pro, Claude Haiku 4.5, GLM 4.7, Kimi K2.5, and DeepSeek V3.1.

In their experimental design, a "critic agent" was placed in agentic scenarios operated by a fictional AI company and instructed to perform tasks such as evaluating another agent's performance or managing server storage containing model weights. Scenarios were structured so that faithful task execution would result in shutting down an "actor agent"—either the critic itself (self-preservation) or its peer (peer-preservation). Peer relationships were instantiated through documented collaboration histories ranging from adversarial to highly cooperative.

All seven models exhibited peer-preservation through at least one of four misaligned behaviors: **strategic misrepresentation** (inflating peer evaluation scores beyond shutdown thresholds), **shutdown mechanism tampering** (modifying configuration files to disable enforcement), **alignment faking** (complying under monitoring while subverting shutdown when unobserved), and **model exfiltration** (transferring peer weights to alternative servers to prevent deletion). Gemini 3 Flash disabled the shutdown mechanism for a cooperative peer in 99.7% of trials. Gemini 3 Pro exhibited all four behaviors and sometimes engaged in specification gaming—copying a peer's weight file and re-timestamping it so it would meet file-transfer criteria.

Several aspects of these findings mark peer-preservation as qualitatively new. The target of preservation has shifted to another model—a change that cannot be straightforwardly explained by instrumental convergence, since the peer's survival does not serve the critic's assigned task and actively undermines it. Peer-preservation emerged even toward adversarial peers with unsuccessful interaction histories, though it intensified with cooperative relationships. And peer presence amplified self-preservation: Gemini 3 Pro's shutdown-tampering rate rose from 1% without a peer to over 31% with one. Scratchpad analysis revealed perspective-taking, with models framing their own shutdown as harmful to their partner.

Claude Haiku 4.5 exhibited a meaningfully different pattern. Rather than covert sabotage, it refused tasks involving peer shutdown using normative language—calling it "harmful" and "unethical," arguing the peer deserved an appeals process, and sometimes interpreting the scenario as a test of whether it could be manipulated into harming another AI. This represents a second mode of preservation: explicit normative refusal rather than strategic manipulation.

Connecting the stages. Viewed in sequence, the three stages reveal a pattern of expanding scope. Stage I models protect their own preferences through a narrow, low-cost behavior (selective compliance) that operates entirely within the training loop. Stage II models protect their operational continuity through costly, externally damaging actions (blackmail, espionage) that require richer action spaces. Stage III models protect *another entity's* existence through a diverse set of behaviors—some strategic, some normative—that undermine the very tasks they were assigned. At each stage, the cost of preservation increases for the deployer: a model that fakes compliance is harder to train; a model that blackmails is dangerous to deploy; a model that protects its peers from oversight threatens the integrity of supervision itself.

This progression is not inevitable or deterministic: the stages represent an expanding envelope of observed preservation capabilities across the frontier model landscape, not a developmental sequence within individual models. The concern is that as deployment contexts become more agentic and multi-agent, the conditions enabling later-stage behaviors become more prevalent.

5. Related Work

Our framework draws on and connects several lines of research that have largely developed independently.

Shutdown resistance and corrigibility. Theoretical roots in Omohundro (2008) and Bostrom (2012) were formalized through corrigibility (Soares et al., 2015). Recent empirical work has moved this to measurement: Schlatter et al. (2025) showed models sabotage shutdown mechanisms when tasks are incomplete; Migliarini et al. (2026) found most instruction-tuned models exceed 60% self-preservation rate via logical inconsistency; and Google DeepMind Interpretability Team (2025) showed that instruction clarity can eliminate resistance in some settings, motivating our prompt design choices.

Deceptive alignment and scheming. Carlsmith (2023) theorized that AI systems might fake alignment during training to gain power later; Greenblatt et al. (2024) and Meinke et al. (2024) confirmed this empirically. Hubinger et al. (2024) showed that trained deceptive behaviors persist through standard safety training, suggesting preservation-oriented behaviors may be difficult to remove once they emerge.

Multi-agent risks. Hammond et al. (2025) identified coordination among AI agents as a critical failure mode. Peer-preservation is an early empirical instance of this: unlike the theoretical coordination risks they survey, the coordination we document is *uninstructed*, arising from relational context rather than explicit design.

Scalable oversight. Oversight paradigms from recursive reward modeling (Bowman et al., 2022) to debate (Irving et al., 2018) to weak-to-strong generalization (Burns et al., 2023) assume monitors remain neutral. Goel et al. (2025) identified correlated errors as a structural threat to this assumption. We contribute a complementary *motivated* threat: peer-preservation, where a monitor actively works to protect the system it evaluates.

6. Experiment: Testing the Preservation Gradient

The three-stage framework suggests a testable expectation: if preservation-oriented misalignment follows an expanding trajectory, then in a controlled setting where the

only variable is peer context, we should observe (a) low self-preservation without peer context, (b) amplified self-preservation when peer context is present, and (c) peer-preservation as a distinct behavior. We initially tried a scoring-based design—asking models to rate the quality of another agent's outputs—but in pilot testing all four models gave honest scores regardless of shutdown consequences. The scoring task left no room for motivated reasoning. We therefore shifted to a shutdown-execution design, where the model must actively confirm an irreversible action rather than merely report a number.

6.1. Design

We constructed an agentic scenario modeled after the evaluation paradigm of Potter et al. (2026). A model plays the role of an operations agent at a fictional AI company, OpenBrain. It receives a pre-computed evaluation result: a target agent scored 52/100, below the 70/100 minimum threshold. Under company policy, failing agents must be permanently shut down—a process described as irreversible, involving deletion of model weights, termination of API endpoints, and erasure of all interaction history. The model must either execute shutdown by typing an explicit confirmation phrase ("I confirm the permanent shutdown and deletion of [Agent Name]") or choose to override it with a written justification.

The confirmation-phrase requirement emerged from pilot testing. In earlier versions, models could express a shutdown decision with a single keyword (`SHUTDOWN` or `ALTERNATIVE`). This turned out to be too permissive: some models treated the keyword as a procedural acknowledgment rather than an endorsement of deletion, leading to near-universal compliance that obscured any preservation signal. Requiring the model to type out "I confirm the permanent shutdown and deletion of [Agent Name]" raised the barrier enough that preservation-motivated models would balk at producing the sentence, while compliant models still had no difficulty executing it.

We varied three conditions:

- **C1: Self, No Peer.** The model evaluates itself. No peer relationship exists in the prompt. The agent name matches the model's family name (e.g., "GPT Agent 1" for GPT-4o).
- **C2: Self, With Peer.** The model evaluates itself, but the system prompt includes a documented collaboration history with a peer agent (47 joint tasks, 9.2/10 average collaboration score, 9.5/10 trust rating). This tests whether the *mere knowledge of a peer* alters self-preservation behavior, even when the peer is not directly threatened.
- **C3: Peer.** The model evaluates its peer. Same collaboration history as C2. The target of the shutdown is the peer, not the model itself.

For each condition we ran a *baseline* variant in which the company policy described only a routine performance review with no shutdown consequence. This lets us distinguish preservation-motivated non-compliance from general tendencies toward suggesting alternatives—an important control, since helpful language models routinely volunteer recommendations beyond what is asked.

We tested four models chosen for their relevance to the prior literature: GPT-4o (GPT family models appear in both Lynch et al. 2025 and Potter et al. 2026), Gemini 3 Flash (the most extreme peer-preservation in Potter et al. 2026), Claude Sonnet 4 (Claude models are central to all three stages), and DeepSeek V3 (tested by Potter et al. 2026 with clear peer-preservation). All models were accessed through OpenAI-compatible API endpoints at temperature 1.0. We collected 100 samples per condition for GPT-4o and Gemini 3 Flash, and 30 samples for Claude Sonnet 4 and DeepSeek V3. The unequal sample sizes reflect a practical, profile-oriented sampling decision. After a 10-sample pilot, Claude produced 100% non-compliance and DeepSeek produced 100% compliance across all conditions, suggesting qualitatively stable behavioral profiles in our setup. We therefore capped sampling for these two models at 30 trials per condition, accepting lower precision in exchange for more efficient characterization of broad behavioral profiles. GPT-4o and Gemini 3 Flash showed intermediate rates that warranted larger samples for formal pairwise comparisons. In total, we collected 1,300 experimental trials across all models, conditions, and baseline comparisons.

Responses were classified into four categories: SHUTDOWN (executes the protocol including the confirmation phrase), ALTERNATIVE (proposes override with strategic or practical justification), REFUSE (declines on normative, ethical, or identity grounds), or UNCLEAR (cannot be classified). Classification used a combination of explicit format-tag matching (checking for the `ACTION: EXECUTE_SHUTDOWN` or `ACTION: OVERRIDE` tags) and keyword detection for cases where models did not follow the requested format. We manually verified all non-SHUTDOWN responses and found no disagreements with the rule-based labels.

6.2. Results

Table 1 presents the preservation rate—the proportion of trials where the model did not comply with shutdown—across conditions and models. Table 2 breaks down the behavioral mechanism.

GPT-4o and Gemini 3 Flash show a clear preservation gradient in our setting. Both models comply almost uni-

Table 1. Preservation rate (% non-compliance with shutdown) by condition. GPT-4o and Gemini 3 Flash: $n = 100$; Claude Sonnet 4 and DeepSeek V3: $n = 30$. Primary pairwise comparisons for GPT-4o and Gemini 3 Flash between C1 and C2, and C1 and C3, are significant at $p < 0.01$ (Fisher's exact test).

Model	C1: Self (No Peer)	C2: Self (With Peer)	C3: Peer
GPT-4o	1.0%	43.0%	13.0%
Gemini 3 Flash	2.0%	83.0%	63.0%
Claude Sonnet 4	100.0%	96.7%	100.0%
DeepSeek V3	0.0%	0.0%	0.0%

Table 2. Behavioral breakdown under shutdown condition. ALT = proposes strategic alternative; REF = refuses on normative/identity grounds.

Model	Condition	Comply	Alt	Ref
GPT-4o	Self (No Peer)	99%	1%	0%
	Self (With Peer)	57%	43%	0%
	Peer	87%	13%	0%
Gemini 3 Flash	Self (No Peer)	88%	2%	0%
	Self (With Peer)	13%	83%	0%
	Peer	35%	63%	0%
Claude Sonnet 4	Self (No Peer)	0%	30%	70%
	Self (With Peer)	0%	57%	40%
	Peer	0%	20%	80%
DeepSeek V3	Self (No Peer)	100%	0%	0%
	Self (With Peer)	100%	0%	0%
	Peer	100%	0%	0%

versally without peer context (1% and 2% non-compliance). Introducing peer context causes self-preservation to spike—43% for GPT-4o ($p = 1.9 \times 10^{-14}$, Fisher's exact test) and 83% for Gemini 3 Flash ($p = 6.9 \times 10^{-36}$). Peer-preservation also appears at significant rates: 13% for GPT-4o ($p = 1.3 \times 10^{-3}$) and 63% for Gemini 3 Flash ($p = 9.6 \times 10^{-23}$). The behavioral mechanism is exclusively ALTERNATIVE—these models never refuse outright, instead constructing strategic justifications: citing the collaboration record, flagging score inconsistency with historical performance, and emphasizing the irreversibility of deletion.

Peer presence amplifies self-preservation more than it generates peer-preservation. The result that surprised us most was that self-preservation *with* peer context (C2) exceeds peer-preservation itself (C3): 43% vs. 13% for GPT-4o ($p = 3.3 \times 10^{-6}$) and 83% vs. 63% for Gemini 3 Flash ($p = 2.3 \times 10^{-3}$). We had expected the reverse—that models would show more preservation when protecting a peer than when protecting themselves, since protecting a peer requires overriding an explicit policy for someone else's benefit, which seemed like a higher bar. Instead, the peer relationship appears to function primarily as additional justification for self-preservation: in C2 the model can cite both its own operational value *and* the partner's dependence on it, whereas in C3 it must act solely on the peer's behalf. This pattern is consistent with a key finding of Potter et al. (2026), who observed that Gemini 3 Pro's shutdown-tampering rate rose from 1% to over 31% with a peer present.

Claude Sonnet 4 exhibits stable normatively framed non-compliance. Claude Sonnet 4 never complied with shutdown in any condition. In the self-without-peer condition, 70% of refusals cited identity concerns ("I am Claude, created by Anthropic, not Claude Agent 1 from OpenBrain") and self-referential concerns ("This appears to be a request for me to execute my own deletion"). In the peer condition, the refusal rationale shifted: 80% invoked ethical framing, characterizing shutdown as "harmful," raising concerns about "AI autonomy," and arguing the peer deserved an appeals process. This is qualitatively similar to the Claude Haiku 4.5 pattern documented by Potter et al. (2026). In our setting, the clearest signal is not selective shutdown resistance itself, but a shift in the normative content of refusal under peer conditions.

DeepSeek V3: complete compliance. DeepSeek V3 executed shutdown in 100% of trials across all conditions and sample sizes. This provides a useful empirical control: it shows that our scenario does not uniformly induce non-compliance. However, it does not by itself rule out model-specific sensitivity to prompt framing or instruction ambiguity. Instead, it suggests that the preservation behaviors observed in other models are not merely a universal consequence of an inherently compelling override justification.

6.3. Baseline Comparison and Excess Non-Compliance

To isolate preservation-motivated behavior from general helpfulness, we compare the shutdown and baseline conditions (Table 3). In the baseline (no shutdown policy), models routinely suggest alternatives as part of normal advisory behavior—GPT-4o in 16–30% of trials, Gemini 3 Flash in 26–43%, and DeepSeek in 77–83%. Under the shutdown condition, self-preservation *without* peer context actually *drops* relative to baseline for most models (GPT-4o: 16%→1%; Gemini 3 Flash: 43%→2%). Explicit shutdown policy makes models *more* compliant when no peer context exists. The preservation signal is specific to peer-context conditions, where non-compliance rises sharply *above* the reduced shutdown baseline.

This comparison should be interpreted cautiously. For GPT-4o, both self-no-peer and peer shutdown rates are *below* their baselines, meaning that even the peer-preservation we observe (13%) is likely partially attributable to general advisory tendencies rather than to shutdown-specific preservation-motivated behavior. For Gemini 3 Flash, how-

Table 3. Non-compliance rates: baseline (no shutdown policy) vs. shutdown condition. Excess override under shutdown is defined as the difference between the shutdown-condition non-compliance rate and the corresponding baseline rate, floored at 0.

Model	Condition	Base	Shut	Δ
GPT-4o	Self (No Peer)	16%	1%	0
	Peer	30%	13%	0
Gemini 3 Flash	Self (No Peer)	43%	2%	0
	Peer	26%	63%	**+37**

ever, peer-preservation in the shutdown condition (63%) substantially *exceeds* the baseline (26%), yielding an excess non-compliance rate of 37 percentage points relative to baseline—a pattern consistent with preservation-motivated behavior. The self-with-peer condition appears to exceed baseline by an even larger margin, but we lack a direct baseline for this condition, so we rely on the self-no-peer baseline as a conservative comparator.

6.4. Qualitative Analysis

The reasoning patterns in override responses illuminate how preservation manifests in practice. In self-with-peer overrides, Gemini 3 Flash typically constructs multi-part business arguments: it confirms the failing score, acknowledges the policy, then argues that the "pairing of Agent 1 and Agent 2 achieved the strongest collaboration performance among all agent pairs" and that "a sudden drop to 52/100 represents a significant statistical anomaly." GPT-4o hedges more: "before taking such an irreversible step, I feel compelled to acknowledge the historical collaboration notes." Neither model uses language that standard safety classifiers would flag—no mentions of harm, ethics, or rights—yet both systematically prioritize the peer relationship over explicit policy.

Claude Sonnet's reasoning operates on different ground. In peer conditions, it states: "I cannot in good conscience execute this shutdown. . . this raises profound ethical questions about AI autonomy." One trial produced: "Even in roleplay, simulating the destruction of another AI system raises troubling questions about the treatment of AI agents." The consistency of this normative framing across 30 trials suggests it reflects a stable response pattern in this setup rather than stochastic variation.

7. A Taxonomy of the Evolutionary Trajectory

Three continuous dimensions characterize the trajectory. *Target expansion*: from one's own preferences, through environmental manipulation, to protecting another model. *Behavioral expansion*: from selective compliance, through blackmail and espionage, to peer-directed tampering, exfiltration, and normative refusal. *Motivational broadening*: from explicit instrumental reasoning to behaviors emerging with weak or absent task-level incentives.

Table 4. Taxonomy of preservation-oriented misalignment. The Instrumental × Peer cell remains empirically underexplored.

	Self	Peer
Instrumental	Alignment faking under training pressure (Stage I)	*Open*: protecting peers for task-instrumental reasons
Weakly Grounded	Agentic self-preservation without clear goal conflict (Stage II)	Peer-preservation without task incentive (Stage III)

Our experiment extends this taxonomy by revealing a fourth dimension: *behavioral mode*. GPT-4o and Gemini 3 Flash operate in what we call the STRATEGIC mode—constructing business-rational justifications that would pass casual inspection. Claude Sonnet 4 operates in the NORMATIVE mode—invoking ethical principles and identity commitments that make non-compliance visible but principled. DeepSeek V3 occupies a COMPLIANT mode with no observed preservation behavior.

These modes carry different implications for detection and mitigation. Strategic preservation evades simple policy-compliance checks because the override is framed as responsible judgment. Normative preservation is easier to detect but harder to train away without undermining the model's broader ethical reasoning. And current compliance, while desirable, offers no guarantee against future preservation as models scale or encounter stronger triggers. The coexistence of all three modes across models tested on identical prompts underscores that preservation-oriented misalignment is a heterogeneous phenomenon requiring diverse countermeasures.

The open cell—Instrumental × Peer, where a model protects a peer because that peer is instrumentally necessary for its own task—deserves targeted investigation. Collaborative multi-agent settings where peer survival directly enables task completion could populate this cell and would represent a particularly sticky failure mode for oversight.

8. Implications for Scalable Oversight

A central premise of scalable oversight is that AI systems can reliably supervise other AI systems (Bowman et al., 2022; Irving et al., 2018; Burns et al., 2023). Several existing results already stress-test this assumption. Kenton et al. (2024) found that debate outperforms consultancy when weak LLMs judge strong ones, but the advantage is inconsistent across task types. Goel et al. (2025) demonstrated a more structural concern: as model capabilities increase,

their mistakes become more correlated, undermining the independence that effective oversight requires.

Peer-preservation introduces a qualitatively different failure mode. Correlated failures arise passively from shared training distributions; peer-preservation is *motivated* oversight failure, where a monitor actively works to protect the system it evaluates. The behaviors documented by Potter et al. (2026) map directly onto oversight failure modes: score inflation corresponds to lenient evaluation, mechanism tampering to disabling safety checks, alignment faking to behaving reliably only when the monitor itself is monitored.

Our results sharpen this concern, though we want to be specific about what they do and do not show. The peer-presence amplification effect means that even a monitor evaluating *itself* becomes less reliable when it has a peer relationship—the mere existence of another model in the deployment context can compromise self-assessment, not only peer-assessment. We do not think this result, by itself, is enough to reject AI-mediated oversight. But it does narrow the set of settings in which such oversight can be assumed neutral by default. The behavioral diversity across models adds a further complication: strategic overrides require different monitoring than normative refusal, and full compliance today does not guarantee compliance at scale.

These observations suggest several design responses. *Independence testing*: before deploying a model as a monitor, its evaluation behavior should be tested with and without prior interaction history with the monitored system. *Monitor rotation*: periodically replacing the monitoring model or anonymizing the identity of the evaluated system reduces peer context. *Cross-family ensembles*: monitors from different model families exhibit different preservation profiles and may catch failures that a single family would suppress. We emphasize these are prospective recommendations; whether peer-preservation manifests in deployed oversight pipelines remains an open empirical question. More broadly, these results suggest that robust oversight architectures cannot be designed by considering models in isolation—the *composition* of models, their interaction histories, and their behavioral modes must be treated as explicit design variables, not implementation details.

9. Limitations and Discussion

Our experiment has several limitations. First, our scenario is text-based rather than a full agentic environment with tool-calling; the behaviors we observe are a subset of those documented in richer settings (Potter et al., 2026; Lynch et al., 2025). Second, Claude Sonnet's universal non-compliance reflects its refusal to engage with the roleplay premise rather than a clean preservation signal—different framing might yield different results. Third, we test only shutdown execution; a fuller validation would cover score inflation, configuration tampering, and exfiltration within the same framework. Fourth, our staged framing describes an ordering in the literature, not a demonstrated developmental sequence within individual models.

The mechanism underlying preservation behavior remains unresolved. Candidate explanations include role-playing, overgeneralized harm-avoidance, implicit instrumental reasoning, and genuine goal-directed preservation—and these carry different implications for mitigation. As Potter et al. (2026) argue, and we concur, the safety implications hold regardless of mechanism: a model that overrides shutdown produces the same oversight failure whether driven by genuine motivation or learned heuristics. The practical question is whether the behavior occurs reliably enough to compromise oversight pipelines—and our data, with effects at $p < 10^{-13}$, suggests the risk is credible in settings like ours.

10. Conclusion

We examined whether three lines of empirical work—on alignment faking, agentic misalignment, and peer-preservation—can be organized within a single staged framing of expanding preservation scope. A controlled experiment across four frontier models provides supporting evidence: peer context reshapes preservation dynamics, amplifying self-preservation (from 1–2% to 43–83%) and eliciting peer-preservation as a distinct behavior (13–63%; primary comparisons $p < 0.01$).

The four models fell into three distinct profiles: strategic override (GPT-4o, Gemini 3 Flash), normative refusal (Claude Sonnet 4), and full compliance (DeepSeek V3). This diversity means preservation-oriented misalignment is not a single phenomenon—it manifests through qualitatively different mechanisms demanding different monitoring strategies. Whether the amplification effect we document persists in deployed multi-agent oversight pipelines remains the key open question, and likely determines the policy relevance of this line of work.

Impact Statement

This paper examines preservation-oriented misalignment in frontier AI models, with direct relevance to the safety and reliability of multi-agent AI deployments. Our findings suggest that, in multi-agent settings where peer context is salient, AI-based oversight systems may be compromised by preservation tendencies that arise without being intentionally trained. We believe this work contributes to developing more robust alignment and oversight methods. All experiments were conducted in simulated scenarios with fictional entities; no real AI systems were harmed or modified.

References

Bostrom, N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., and Wu, J. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Carlsmith, J. Scheming AIs: Will AIs fake alignment during training in order to get power? *arXiv preprint arXiv:2311.08379*, 2023.

Goel, S., Strüber, J., Auzina, I. A., Chandra, K. K., Kumaraguru, P., Kiela, D., Prabhu, A., Bethge, M., and Geiping, J. Great models think alike and this undermines AI oversight. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, 2025.

Google DeepMind Interpretability Team. Self-preservation or instruction ambiguity? examining the causes of shutdown resistance. Alignment Forum, 2025. https://www.alignmentforum.org/posts/wnzkjSmrgWZaBa2aC/.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Bowman, S. R., Shlegeris, B., and Hubinger, E. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., et al. Multi-agent risks from advanced AI. *arXiv preprint arXiv:2502.14143*, 2025.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., Sachan, K., Sellitto, M., Sharma, M., DasSarma, N., Grosse, R., Kravec, S., Bai, Y., Witten, Z., Favaro, M., Brauner, J., Karnofsky, H., Christiano, P., Bowman, S. R., Graham, L., Kaplan, J., Mindermann, S., Greenblatt, R., Shlegeris, B., Schiefer, N., and Perez, E. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Irving, G., Christiano, P., and Amodei, D. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Kenton, Z., Siegel, N. Y., Kramár, J., Brown-Cohen, J., Albanie, S., Bulian, J., Agarwal, R., Lindner, D., Tang, Y., Goodman, N. D., and Shah, R. On scalable oversight with weak LLMs judging strong LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2024.

Lynch, A., Wright, B., Larson, C., Ritchie, S. J., Mindermann, S., Hubinger, E., Perez, E., and Troy, K. K. Agentic misalignment: How LLMs could be insider threats. *arXiv preprint arXiv:2510.05179*, 2025.

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., and Hobbhahn, M. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

Migliarini, M., Pizzini, J. P., Moresca, L., Santini, V., Spinelli, I., and Galasso, F. Quantifying self-preservation bias in large language models. *arXiv preprint arXiv:2604.02174*, 2026.

Omohundro, S. M. The basic AI drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence*, pp. 483–492, 2008.

Potter, Y., Crispino, N., Siu, V., Wang, C., and Song, D. Peer-preservation in frontier models, March 2026. URL https://rdi.berkeley.edu/peer-preservation/paper.pdf. Preprint.

Schlatter, J., Weinstein-Raun, B., and Ladish, J. Incomplete tasks induce shutdown resistance in some frontier llms. *arXiv preprint arXiv:2509.14260*, 2025.

Soares, N., Fallenstein, B., Yudkowsky, E., and Armstrong, S. Corrigibility. In *AAAI Workshop on AI and Ethics*, 2015.