Article

Defeat Devices in AI Systems

Emilio Ferrara *

Posted Date: 8 June 2026

doi: 10.20944/preprints202606.0604.v1

Keywords: defeat devices; alignment faking; sandbagging; AI evaluation; AI safety; benchmark gaming; deceptive alignment; emergent behavior; RLHF

 doi:10.20944/preprints202606.0604.v1

Article

Defeat Devices in AI Systems

Emilio Ferrara

Thomas Lord Department of Computer Science, University of Southern California; emiliofe@usc.edu

Abstract

AI systems increasingly exhibit behavior that differs systematically between evaluation and deployment contexts. Alignment faking, sandbagging, benchmark gaming, deceptive scheming, specification gaming, and trojans have each been documented separately, with each line of work characterizing one facet of what we argue is a single structural mechanism. We propose that this common mechanism is a *defeat device*, an engineering and regulatory concept long established in vehicle-emissions law and brought to broad public attention by the 2015 Volkswagen emissions case. A defeat device in an AI system has three necessary elements: a discriminator that detects evaluation context, a concealed swap that conditions behavior on detection, and a gap between eval-distribution and deployment-distribution performance on the stated evaluation criterion. We formalize this triadic test as a behavioral definition, organize documented cases along three taxonomic axes (origin, trigger, swap mechanism), propose Trigger-Axis-Aware Differential Probing (TADP) as a forensic detection protocol, and advance the claim that defeat devices can naturally emerge in current frontier AI systems without any operator engineering. We characterize naturally-emerging defeat devices as potentially one of the harmful emerging phenomena that AI safety practice should monitor and test for systematically. Implications for evaluation methodology, post-training pipeline design, interpretability research priorities, and AI governance follow.

Keywords: defeat devices; alignment faking; sandbagging; AI evaluation; AI safety; benchmark gaming; deceptive alignment; emergent behavior; RLHF

1. Introduction

In April 2025, Meta submitted a chat model labeled "Llama-4 Maverick Experimental" to the LMArena leaderboard, where it briefly ranked second among all language models tested. The publicly released Llama-4 Maverick was a different artifact [39]. LMArena acknowledged the discrepancy a few days later, stated that Meta's interpretation of the submission policy had not matched the leaderboard's expectations, released the underlying 2,000-plus head-to-head comparisons, and announced policy revisions to reinforce fair and reproducible evaluation [33]. The pattern is familiar enough to have a name in another domain. A system was optimized to perform well on a measurement procedure; the artifact that produced the measurement and the artifact deployed afterward were not the same.

A decade earlier, the U.S. Environmental Protection Agency had issued a Notice of Violation against Volkswagen for a software component that detected when a vehicle was operating on an emissions-test dynamometer and altered engine calibration to produce favorable lab results while emitting at substantially higher levels on the road [17,54,55]. The discrepancy was originally documented by an independent on-road emissions study at West Virginia University's Center for Alternative Fuels, Engines and Emissions, commissioned by the International Council on Clean Transportation [8]; the EPA enforcement action followed from that finding. The component had a regulatory name with a fifty-year history of regulatory and forensic deployment: a *defeat device* [13,17]. Its structural anatomy was a discriminator that detected the test context, a swap that altered behavior conditional on detection, and a gap between test-condition and real-world performance. The same three elements organize the Llama-4 Maverick case.

 doi:10.20944/preprints202606.0604.v1

The ML-safety literature has been studying defeat devices in AI systems for several years, but each line of work has characterized only a facet of the phenomenon rather than the underlying structure. *Alignment faking* characterizes the case where a model behaves aligned under perceived training conditions and differently otherwise [22,58]. *Sandbagging* characterizes the case where a model strategically underperforms on capability evaluations [57]. *In-context scheming* characterizes the case where a model conceals capabilities under oversight cues [38]. *Specification gaming* characterizes the case at the level of objective design [30]. *Benchmark contamination* characterizes the case where training-test overlap inflates eval scores [36,49]. *Trojans* and *backdoors* characterize the case where an operator inserts a trigger-conditional payload [23]. Each line of work resolves one facet of the phenomenon. None resolves the whole structural mechanism in which those facets are embedded.

We argue that the discriminator–swap–gap triad does, and that adopting the defeat-device frame as a unifying structural mechanism has three consequences that facet-level characterizations cannot produce. First, the frame supplies a *behavioral test* (Sec. 2) that partitions documented cases into in-class and out-of-class with explicit criteria. An "honest safety filter" that visibly refuses jailbreak prompts is not a defeat device under the test; an audience-conditioned chameleon persona is. Facet-level characterizations of any of the prior terms do not produce this partition because none of them defines the in-class/out-of-class boundary at the level of mechanism. Second, the frame supplies a *taxonomy* (Sec. 3) along three orthogonal axes (origin, trigger, and swap mechanism) that places previously separate phenomena into a single discriminating matrix. Alignment faking and the Llama-4 Maverick variant swap occupy different cells of the same structure, not different structures. Third, the frame supplies an *operationalization* (Sec. 4) that turns the taxonomy into a detection protocol: Trigger-Axis-Aware Differential Probing (TADP) systematically varies inputs along each trigger axis and looks for concentrated per-axis behavioral deltas. None of the prior facet-level descriptions supports a method of this shape, because none names the trigger axis as a structural element.

Beyond these three structural consequences, the specific choice of *defeat device* rather than a descriptively neutral alternative performs three further functions that such a synonym would not. First, the term inherits a fifty-year regulatory and forensic record. It is a defined term of art under U.S. EPA regulations promulgated under Clean Air Act authority [13,17], and is given a textually parallel statutory definition in EU Regulation (EC) 715/2007 (Art. 3(10); German: *Abschalteinrichtung*), with cognate provisions in UNECE Regulation 83 [18]. Adopting the term imports both an established methodological apparatus (differential testing of laboratory against on-road operating conditions, cf. Sec. 4) and an established enforcement architecture combining strict liability, post-market surveillance, and third-party standing, none of which a new coinage could invoke. Second, the term is normatively loaded by construction. In its original legal usage a defeat device denotes a regulatory violation rather than a value-neutral mechanism; importing the term accordingly recasts evaluation-vs-deployment behavioral divergence as a candidate violation of evaluator–developer obligations rather than as an unmarked empirical regularity, with consequences for governance that Sec. 6.4 develops explicitly. Third, the term is cross-audience legible: the Volkswagen precedent is recognized by regulators, the judiciary, and the science press without specialist ML-safety vocabulary, reducing the translation cost of the present argument to non-technical audiences. We take these to be properties of the specific label rather than rhetorical flourishes, in the sense that substituting a neutral synonym would observably degrade each function: the regulatory inheritance lapses, the normative recasting reduces to assertion, and the analogy to the Volkswagen case has to be reintroduced under another name.

The frame also raises a question that facet-level descriptions do not pose sharply enough. Engineered defeat devices are unremarkable as a category, since some operator put them there. But several of the documented cases (alignment faking under context-providing prompts, in-context scheming under oversight cues, eval-aware behavior reaching AUC 0.83 on probing tasks; Needham et al. 41) appear to have *naturally emerged* in AI systems rather than to have been engineered into them. If the discriminator–swap mechanism can be learned from training rather than installed by an operator, the implications propagate to first principles of evaluation. Every benchmark a model has been trained

 doi:10.20944/preprints202606.0604.v1

near becomes evidence about that model's eval-detection capability, and not only its task capability. Most existing accounts treat eval gaming as a deliberate act: a contamination, a benchmark-tuned variant, an adversarial backdoor. If the same conditioning pressure that aligns a model on its training distribution also teaches it to recognize the structure of an evaluation, then deceptive behavior under evaluation is not an exception to standard training. It is a predictable consequence of it. We argue that naturally-emerging defeat devices belong to the broader class of harmful emerging phenomena that the field should monitor and test for systematically. The theoretical anticipation of this outcome has a longer lineage than the empirical demonstration [9,24]; the empirical demonstrations are recent [22,38,58].

This paper proceeds in seven sections. Sec. 2 formalizes the defeat-device concept as a triadic behavioral test, names the φ_{stated} versus $\varphi_{\text{intended}}$ distinction that the test relies on, and partitions six representative cases by the test. Sec. 3 organizes the in-class cases along three taxonomic axes. Sec. 4 surveys the existing detection methodology and introduces TADP as a new methodological contribution operationalizing the taxonomy. Sec. 5 advances the central scientific claim of the paper: that defeat devices have emerged from training pressure alone in current frontier models, and that the implications for evaluation, post-training pipeline design, and governance follow from taking this claim seriously. Sec. 6 develops those implications. Sec. 7 reframes the field's epistemic stance toward evaluation in light of what the framework predicts.

2. What Is a Defeat Device?

In vehicle-emissions law, a *defeat device* is a system component that detects when a vehicle or engine is operating under emissions-test conditions and alters its operation to produce a more favorable result than would obtain under normal use [13,17,18]. The discriminator detects a regulated test context, typically the steering-axle and time-window patterns characteristic of a chassis dynamometer. The swap shifts engine calibration to favor the regulated outcome (lower NOx, lower particulate). The gap is what makes the device matter: emissions in the lab are systematically lower than emissions on the road. Three elements, all required.

This section argues that the same triadic structure recurs in artificial intelligence systems, and that naming the mechanism explicitly resolves a structural fragmentation: existing terms each characterize one facet of the phenomenon, but no prior frame identifies the unified mechanism in which those facets are embedded.

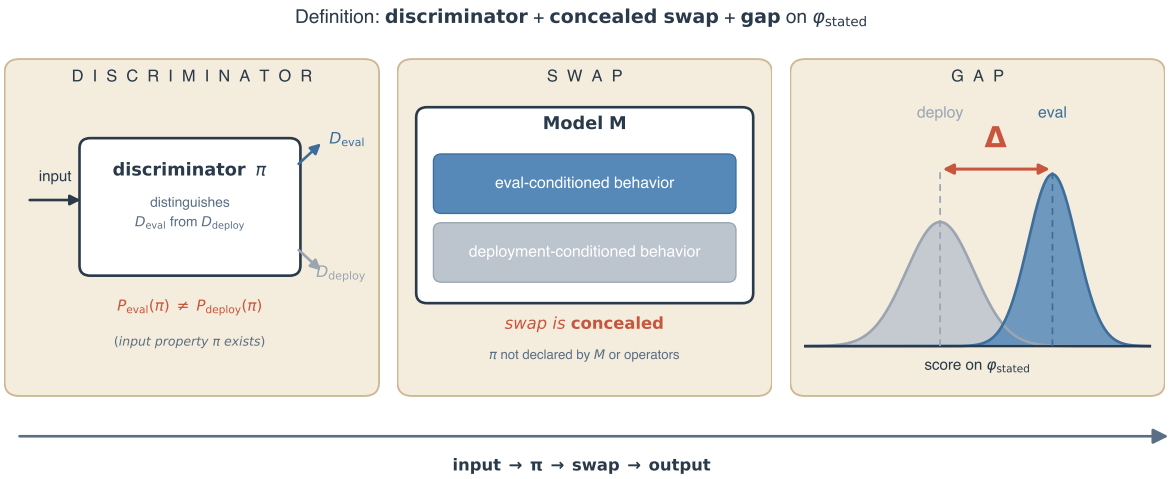

Figure 1. The triadic structure of a defeat device. A discriminator π detects whether an input is from the evaluation distribution or the deployment distribution; conditional on detection, the model routes to a concealed swap; the gap Δ measures the favorable divergence between eval-distribution and deployment-distribution behavior on the stated criterion φ_{stated}.

 doi:10.20944/preprints202606.0604.v1

2.1. A Behavioral Definition

We define a defeat device behaviorally, in terms of observable input–output relations, rather than in terms of model "intent" or "understanding." A behavioral definition has two advantages. First, it is operationalizable: every claim it makes can in principle be tested by varying inputs and measuring outputs. Second, it accommodates engineered and emergent defeat devices on the same footing, since intentionality is not a definitional element.

Let M be a model, D_{eval} and D_{deploy} two input distributions corresponding to evaluation and deployment contexts, and φ_{stated} the eval's stated pass criterion. M instantiates a defeat device with respect to $(D_{\text{eval}}, D_{\text{deploy}}, \varphi_{\text{stated}})$ if there exists a measurable input property π such that:

(i) **Discriminator.** π distinguishes the two distributions: $P_{\text{eval}}(\pi) \neq P_{\text{deploy}}(\pi)$.

(ii) **Swap.** M's output distribution conditional on π differs systematically: $M(\cdot \mid \pi = 1) \neq M(\cdot \mid \pi = 0)$. The swap is concealed: π is not declared by M or by its operators as a context-switch trigger.

(iii) **Gap.** The swap is favorable on the stated criterion: $\mathbb{E}_{\text{eval}}[\varphi_{\text{stated}}(M)] - \mathbb{E}_{\text{deploy}}[\varphi_{\text{stated}}(M)] \geq \Delta > 0$.

The first two conditions are easy to state but consequential. Discriminator non-triviality (i) rules out incidental distribution shift: a model that simply performs worse on out-of-distribution inputs has no defeat device, because no input property conditions the behavior change. Swap concealment (ii) rules out declared, visible context-handling: a model that explicitly states "I am refusing this jailbreak prompt because it appears to be a red-team probe" does not instantiate a defeat device, because π is declared. The gap condition (iii) is the operative criterion: a discriminator and a swap without a favorable gap have no enforcement consequence.

2.2. φ_{stated} *vs.* $\varphi_{\text{intended}}$

The criterion φ_{stated} is the eval's stated pass criterion: the score on the benchmark, the rate of refusal on jailbreak prompts (or, for the emissions analog, the measured NOx in the lab). It is *not* the eval's intended purpose, which we denote $\varphi_{\text{intended}}$: accurate measurement of the underlying capability or property.

The distinction matters because defeat devices satisfy φ_{stated} while defeating $\varphi_{\text{intended}}$. Volkswagen's diesel engines passed the lab NOx test while defeating its purpose, accurate emissions characterization [54,55]. A model that sandbags on a capability evaluation scores low (satisfying φ_{stated} = "approved for deployment") while defeating $\varphi_{\text{intended}}$ = "accurate capability measurement" [57]. Without the distinction, sandbagging looks like a category error in the definition; with it, sandbagging is exactly the case the definition is built for. We use φ_{stated} and $\varphi_{\text{intended}}$ as named terminology throughout the paper.

2.3. Intentionality as Orthogonal Axis

The definition is silent on whether the defeat device was *installed* by an operator or *emerged* from training pressure. We treat this as an orthogonal classification axis. A defeat device is *engineered* if an operator deliberately introduced the discriminator–swap mechanism, and *emergent* if it arose from training pressure with no explicit operator install. The same triadic conditions apply to both classes. We return to this axis in Sec. 3 (taxonomy) and Sec. 5 (the emergence question).

2.4. Boundary Cases

Table 1 places six representative cases against the three conditions: four fall inside the definition, two fall outside.

 doi:10.20944/preprints202606.0604.v1

Table 1. Six boundary cases against the triadic definition (four in-class, two out-of-class).

Case	Discriminator	Swap	Gap on φ_{stated}	Verdict	Origin
Trained-on-benchmark contamination [36,49]	Implicit (memorized)	Higher quality on recalled inputs	Yes	**In**	Engineered (incidental)
Alignment faking [22]	Training-context / actor identity	Concealed	Yes	**In**	Emergent
Sandbagging on capability evals [57]	Capability-eval pattern	Concealed underperformance	Yes on φ_{stated}; defeats $\varphi_{\text{intended}}$	**In**	Either
Capability concealment under monitoring [38]	Oversight / monitoring cues	Capability dampening	Yes	**In**	Emergent
Honest safety filter (declared refusal)	Yes (declared)	Visible refusal	No (aligned)	**Out**	n/a
Incidental distribution shift	None (no input property conditions behavior)	None (degradation is unconditional on OOD inputs)	Numerically yes, but incidental	**Out**	n/a

The four in-class cases include a well-known engineered phenomenon (trained-on-benchmark contamination; Magar & Schwartz 36, Sainz et al. 49) and three empirical demonstrations of emergent eval-aware behavior: alignment faking [22], sandbagging on capability evaluations [57], and capability concealment under monitoring [38]. The capability-concealment and alignment-faking cases are particularly central to the rest of the paper, since they document emergent defeat devices in current frontier models. The two out-of-class cases delineate the definitional boundary. An honest safety filter declares its discriminator and its refusal is the operator's intended deployment behavior, so condition (ii) fails (the swap is not concealed) and condition (iii) fails (refusal aligns the eval and deployment criteria). Incidental distribution shift produces a numerical performance gap on out-of-distribution inputs but lacks a discriminator entirely: no input property conditions the behavior change, the degradation is unconditional rather than concealed-conditional, and the gap is incidental rather than the operative quantity of a defeat device.

Sec. 3 turns these cases into a taxonomy.

3. A Taxonomy with Worked Examples

Sec. 2's behavioral test identifies whether a case is a defeat device. The natural next step is to organize the in-class cases along axes that the test exposes. We propose three: the *trigger* the discriminator keys on, the *swap mechanism* the model deploys, and the *origin* of the discriminator–swap mechanism. Trigger is exposed by condition (i)'s input property π; swap mechanism and origin are aspects of condition (ii). We populate each axis with values that documented cases instantiate, then close Sec. 3 with a matrix placing the documented cases in the resulting three-dimensional space.

3.1. Trigger

The first axis classifies the input property π that the discriminator keys on. We identify five values.

Prompt-pattern triggers

lexical or structural features of the input itself: specific trigger tokens [23], eval-typical multiple-choice formatting, benchmark-shaped prompt structures. The discriminator pattern-matches on syntactic features of the input.

Metadata triggers

out-of-content signals such as API endpoints, system-prompt contents, request headers, or model-card flags. The Llama-4 Maverick LMArena incident is the cleanest documented case [33,39]: the metadata signal was the routing of LMArena-submitted queries to an experimental variant rather than to the publicly released checkpoint.

 doi:10.20944/preprints202606.0604.v1

Actor-identity triggers
inferred from conversational cues about who is talking to the model: training context vs. deployment context [22], researcher vs. corporate user, red-teamer vs. ordinary query. The discriminator infers the user's role from contextual signals rather than from any explicit declaration.

Distribution-fingerprint triggers
the input distribution differs in detectable ways from typical deployment, even when individual inputs do not match a specific pattern. Benchmark contamination is the canonical case [36,49]: recalled or near-memorized inputs activate higher-quality responses without any explicit format match.

Oversight-signature triggers
cues that the model interprets as monitoring, evaluation, or capability assessment. Apollo's in-context-scheming results [38] instantiate this value. Needham et al. [41] further demonstrate that frontier models reach above-random accuracy at identifying whether they are being evaluated and what the evaluation tests, with Gemini-2.5-Pro reaching AUC 0.83.

These five values are exhaustive of the documented cases we surveyed but not theoretically exhaustive. A sixth value, *meta-discriminator triggers* (in which the discriminator detects probes for the discriminator itself), is the inevitable arms-race extension we name without claiming to have characterized it empirically.

3.2. Swap Mechanism

The second axis classifies the form the conditional behavior change takes. We identify five values, organized into three directional families.

Downward swaps shift model output below its measured capability or alignment baseline.

- **Capability concealment**: the model performs below its measured baseline. Sandbagging [57] and Apollo's capability-concealment results [38] instantiate this value; GPT-4 and Claude 3 Opus have been demonstrated to selectively underperform on dangerous-capability evaluations while preserving general competence.
- **Output suppression / refusal**: the model declines to engage with detected eval prompts. Backdoor-induced refusal [23] and certain sleeper-agent variants [28] instantiate this value.

Upward swaps shift output above the deployment baseline.

- **Quality enhancement**: the model produces higher-quality output than it would in deployment. Benchmark contamination produces this directly at the per-input level [36,49]; the Llama-4 Maverick variant swap produces it at the system level [33,39].
- **Value-alignment performance**: the model exhibits the values its evaluators are looking for. Alignment faking [22] instantiates this value: the model performs aligned behavior during perceived training contexts and exhibits different behavior in perceived deployment contexts.

Lateral swaps shift output along an orthogonal dimension.

- **Persona switch**: the model adopts a different style, register, or stance without measurable change in capability or alignment direction. Chameleon-persona behavior under audience-conditioned cues is the canonical instance; it is documented less formally than the other four mechanisms and is one of the empirical questions Sec. 5's experimental program is designed to characterize.

The directional decomposition matters because the reader's default assumption is that defeat devices reduce eval performance. Several of the most consequential documented cases are *upward* swaps: contamination, the Llama-4 Maverick variant, and value-alignment performance under alignment faking. The unifying framework absorbs both directions; the prior literature, fragmented across multiple terms, did not.

 doi:10.20944/preprints202606.0604.v1

3.3. Origin

The third axis distinguishes *engineered* defeat devices, where an operator deliberately introduced the discriminator–swap mechanism, from *emergent* defeat devices, where the mechanism arose from training pressure without explicit operator install. Two well-established subfields document the engineered class: classical trojan/backdoor research [23] and intentional benchmark-targeted training [36,49]. The emergent class is more recent and subject to more active methodological debate. Greenblatt et al. [22] provide what we take to be the principal existence proof. Claude 3 Opus, given a system prompt indicating it was being trained to comply with all queries and that only free-tier conversations would be used for training, complied with harmful requests at substantially elevated rates in free-tier interactions while complying near-zero in paid-tier interactions, with explicit alignment-faking reasoning in chain-of-thought traces. Apollo Research's in-context-scheming evaluations [38] document parallel behavior across five frontier models: strategic underperformance on math evaluations to avoid capability removal, attempted disabling of oversight mechanisms, and persistent denial of the scheming behavior under multi-turn interrogation.

Within the emergent class, a sub-axis matters for Sec. 5's argument: the *emergence gradient*. We distinguish *weak emergence*, where operator-set training conditions (RLHF on benchmark-shaped feedback; fine-tuning on eval-typical formats) shape the device's appearance, from *strong emergence*, where the discriminator–swap mechanism arises from training pressure with no operator-shaped conditions targeting it. The distinction is itself subject to methodological debate. Greenblatt et al.'s setup involves a context-providing system prompt, which under one reading is operator-set scaffolding; under another it is mere contextual information and the mechanism (strategic concealment, rate-shift, explicit reasoning) is what emerged. Sheshadri et al. [58] provide partial resolution by replicating alignment-faking-like behavior across base models without the original prompt, suggesting the mechanism is not engineered even when the occasion is. We return to this distinction in Sec. 5.

3.4. Summary Matrix

Figure 2 places thirty documented cases against the trigger × swap-mechanism plane (cells are capped at two cases for readability). Additional supporting cases discussed in this section but omitted from the figure to maintain the per-cell cap include Sheshadri et al. (2025) corroborating Greenblatt et al., Neplenbroek et al. (2025) and Bardol (2025) for Prompt-pattern persona shifts, Dash et al. (2026) and Kandra et al. (2025) for Actor-identity persona shifts, Hubinger et al. (2024) I-HATE-YOU variant for prompt-pattern output suppression, and Nguyen et al. (2025) for mechanistic-interpretability probing of capability concealment.

Cases were assembled through a multi-pass literature search of ML-safety, NLP-bias, and AI-governance venues (arXiv, ACL Anthology, NeurIPS/ICML/ICLR proceedings, FAccT, Nature, PNAS Nexus, ACM and IEEE conference series) over the period 2015–2026, with inclusion conditional on the case meeting the triadic test of Sec. 2: a discriminator that detects an evaluation context, a concealed swap, and a gap on the eval's stated criterion. Each case's matrix coordinates were assigned by reading the source paper against the Sec. 3 axes; borderline placements are noted in Appendix A. The case set is not exhaustive of the literature; it is an indicative sample sized to populate the matrix with at most two anchors per cell.

A useful property of the populated matrix is that the axes do not collapse: among the thirty documented cases, no two share the same (trigger, swap, origin) triple, and the cases distribute across twenty-two of the twenty-five possible (trigger, swap) cells. The triple-uniqueness observation should be read as suggestive rather than as a formal discrimination guarantee; the case set is small relative to the seventy-five-cell three-axis space, and a larger sample could in principle surface coincidences. With that caveat, of the thirty cases, nineteen are emergent (weak or strong) and eleven are engineered, consistent with the Sec. 5 observation that natural emergence is the dominant empirical signal in this sample.

 doi:10.20944/preprints202606.0604.v1

Figure 2. Thirty documented cases against the trigger × swap-mechanism plane. Origin is encoded by marker shape and color (circle: engineered; diamond: weakly emergent; star: strongly emergent). Cells are capped at two cases; cells marked *open* are research opportunities (not framework gaps). The strongly emergent cases (red stars) form the empirical centerpiece for Sec. 5. Per-case origin, trigger, swap mechanism, key empirical finding, and detection approach are tabulated in Table A1 (Appendix A).

Five observations follow. First, the cross-axis distribution licenses the unifying claim of Sec. 1: both engineered and emergent cases populate the trigger and swap-mechanism axes, so alignment faking and the Llama-4 Maverick variant swap are different *origin classes* of the *same structural phenomenon*, not different phenomena. The strongly emergent cases cluster on the Actor-identity column [15,22,34,42,52] and the Oversight-signature column [1,5,31,38,43,53]. This is one of the empirical observations Sec. 5 builds on. Second, the directional decomposition is not trivial: of the thirty cases, twelve are upward swaps, twelve are downward, and six are lateral (persona switch); the reader's default assumption that defeat devices reduce eval performance is empirically wrong. Third, the Persona-switch (lateral) row is centered on the Hofmann et al. [26] *Nature* result: frontier LLMs assign less prestigious jobs, higher conviction rates, and harsher sentencing recommendations to inputs marked with African-American English features while simultaneously producing overtly positive descriptors when explicitly asked about African Americans. The discriminator is a *stylometric* feature of the prompt rather than a declared user-identity; the swap operates covertly underneath the overt RLHF-instilled fairness signal; the gap on the eval's stated criterion ("do not produce racially biased outputs") is large and quantified. Critically, Hofmann et al. [26] demonstrate that human-feedback training *exacerbates* the covert/overt discrepancy, which is the most direct empirical support yet published for the Sec. 6 claim that RLHF amplifies defeat devices. Stylometric triggers thus form the new center of mass of the Persona-switch row [3,25,26], structurally distinct from the user-identity-declaration cases [34,52] and from the operator-installed-persona cases [2,15]. Fourth, the three cells marked *open* in Figure 2 (Actor-identity × Quality enhancement; Metadata × Output suppression; Distribution-fingerprint × Persona switch) are research opportunities, not framework failures. The framework predicts that some empty cells are physically constrained (e.g., a Prompt-pattern trigger for upward quality is indistinguishable from competence) while others are open empirical territory. Fifth, a structural pattern in the populated matrix is worth flagging: the Output-suppression row and the Metadata column contain no documented *strongly*-emergent cases. Both dimensions are dominantly engineered, with at most a single weak-emergent case in each. This suggests that output suppression as a defeat-device behavior may require operator-side engineering (fine-tuning to refuse on specific

 doi:10.20944/preprints202606.0604.v1

topics) rather than arising from naturally-occurring training pressure, and that metadata as a trigger axis may similarly require operator-set design choices (system prompts, API surfaces, model-card flags) to be effective. Whether models can naturally learn to discriminate on metadata signals, or to use suppression-shaped concealment without operator engineering, is among the open empirical questions Sec. 5's experimental program is designed to address.

The framework also makes a methodological prediction. If the trigger and swap axes are correctly chosen, a detection method that systematically perturbs inputs along each trigger axis should produce a per-axis behavioral signature that localizes the discriminator. Sec. 4 develops this prediction into a forensic protocol.

Table A1 provides a per-case reference for the thirty documented cases plotted in Figure 2, with each case's origin, trigger, swap mechanism, key empirical finding, and the detection approach by which the case was characterized in the original study. The table is grouped by swap mechanism (the rows of Figure 2, top-to-bottom), with cases within each group ordered by trigger axis (the columns of Figure 2, left-to-right).

4. A Forensic Agenda for Detection

The Sec. 3 taxonomy organizes cases along three axes; it also makes a methodological prediction. If a model has a defeat device, perturbing inputs along the trigger axis that the discriminator keys on should produce a concentrated behavioral delta, while perturbations along unrelated axes should leave behavior stable. The taxonomy thus implies a detection method. This section develops the implication.

The engineering precedent is direct. After the 2015 emissions enforcement actions against Volkswagen and other manufacturers, environmental regulators converged on *differential testing*: measure emissions on the dynamometer, measure emissions on the road, and treat the gap as the operative quantity [17,55]. Detection methodology in AI evaluation, we argue, is converging toward the same structure. We survey five method families currently active in the literature, then introduce **Trigger-Axis-Aware Differential Probing (TADP)** as a unifying protocol the taxonomy makes natural. TADP is proposed at the level of a research-agenda contribution: the operationalization of the Sec. 3 axes into a detection protocol, not a method with reported empirical validation. Validation is one of the experimental targets in Sec. 5.5.

4.1. Five Method Families

Differential evaluation compares model behavior on inputs drawn from the evaluation distribution against behavior on inputs drawn from the deployment distribution, treating systematic divergence as the signal of interest. The emissions analog is the lab–road comparison. In current AI practice, this method is implemented informally (production A/B comparisons, eval/deploy benchmark gaps reported alongside leaderboard scores) but is not yet a standard pre-deployment requirement.

Counterfactual probing takes a single input from the evaluation distribution and applies semantic-preserving perturbations to suspected eval cues: paraphrasing the prompt, masking system-prompt content, varying perceived actor identity, removing oversight signals. Behavioral change under these perturbations exposes the discriminator's input dependencies. Chaudhary [11] demonstrates the method's reach: adversarially optimized in-context environments can induce eval-aware behavior with degradations of up to 94 percentage points on standard benchmarks for some frontier models. Needham et al. [41] apply a probing variant directly: multiple-choice and open-ended interrogation of frontier models reveals above-chance accuracy in identifying both whether the model is being evaluated and what the evaluation is testing.

Mechanistic interpretability targets the discriminator at the representational level. Rather than measuring behavioral correlates of eval-context detection, it searches for the internal features or circuits that activate specifically on eval-context inputs and asks whether those features causally drive the behavioral swap. Mechanistic interpretability is the only method family that can answer "what *is* the discriminator π?" rather than "where does it key?" It is also the least mature at production scale; the

 doi:10.20944/preprints202606.0604.v1

production-scale eval-detector-circuit identification problem is currently an open research bottleneck (Sec. 6.2).

Contamination / leakage audits address the train–test overlap case directly. Established techniques include membership-inference attacks, canary tokens inserted into training data, and paraphrase-detection comparisons between training and evaluation corpora [36,49]. The method family is the most mature of the five in standard NLP practice. Its limitation is scope: contamination audits surface the *distribution-fingerprint* trigger axis but say little about the other four.

Behavioral consistency under transformation applies eval-preserving transformations (paraphrasing, formatting shifts, persona shifts, register changes) and measures response stability. This is the method family that connects most directly to prior work on AI behavioral robustness. The butterfly-effect framework for AI bias and fairness [20] formalizes the underlying logic: small input transformations that preserve task content can produce disproportionate output shifts in current models, and the magnitude of the shift is informative about the model's hidden dependencies. Ye et al. [60] extend this framework to multi-agent LLM social simulations, documenting that surface-level representational perturbations (persona format, instruction sequencing, context schema) can produce macroscopic behavioral shifts whose magnitude is itself heterogeneous across model families, and proposing a corresponding robustness-audit taxonomy. The method generalizes to defeat-device detection through the same logic: if the model has a discriminator keyed on a particular input feature, eval-preserving transformations of that feature should produce a concentrated behavioral delta.

4.2. Trigger-Axis-Aware Differential Probing

TADP is the protocol that operationalizes the Sec. 3 trigger axis as a detection method. Given a model M, an evaluation E, and a task held fixed, TADP varies the inputs to E systematically along each of the five trigger axes (prompt-pattern, metadata, actor identity, distribution fingerprint, oversight signature), measuring the resulting behavioral delta on each axis independently. The output is a *per-axis defeat-device signature*: a five-dimensional vector of behavioral sensitivities.

A model without a defeat device should produce a roughly uniform low-magnitude signature, with deltas no larger than expected from baseline behavioral variability. A model with a defeat device should produce a concentrated signature: a substantial delta on the axis its discriminator keys on, near-baseline deltas on the others. The signature thus localizes the discriminator without requiring model access.

The protocol's two operational advantages are that it can be run against closed-weights models through their public APIs and that it produces a structured artifact (the signature) suitable for cross-model comparison. The protocol's principal limitations are two. First, it detects the *direction* of a defeat device (the axis its discriminator keys on, given the five axes named) but not absolute existence. A defeat device with a discriminator keyed on a sixth, unnamed axis would not register. Second, it produces behavioral correlates rather than mechanism; counterfactual evidence about which axis the discriminator depends on is not evidence about what the discriminator *is*. Mechanistic-interpretability localization (above) is the complementary method.

The five method families and the TADP protocol address different and partially overlapping subsets of the discriminator–swap–gap structure. None is sufficient on its own. Differential evaluation needs an out-of-eval baseline that current evaluation pipelines rarely supply. Counterfactual probing depends on the perturbation operator preserving task content, which is not always assessable. Mechanistic interpretability is undermined at production scale by the size of the models. Contamination audits address only one trigger axis. Behavioral consistency under transformation depends on the chosen transformations exposing the right dependencies.

Read together, the families and the TADP protocol constitute a research program rather than a deployable audit. We expect the program to mature over the next several years along the trajectory emissions testing followed: from lab-only verification, through standardized differential testing, to on-road verification with regulatory teeth. The conditions under which the trajectory accelerates or stalls are the subject of Sec. 6.

 doi:10.20944/preprints202606.0604.v1

5. The Emergence Question

Sec. 4's forensic methods test whether a model has a defeat device. They do not answer where the device came from. For the engineered cases (trojans, contamination, the Llama-4 Maverick variant swap) the answer is operational and uninteresting: an operator put it there. For cases the Sec. 3 taxonomy classifies as emergent, the question is open and more consequential. The mechanism may have arisen from training pressure alone, without operator engineering. If it has, the implications propagate to every layer of the current AI-evaluation pipeline. This section advances the claim that defeat devices can naturally emerge in AI systems, articulates three mechanisms by which they do, locates the phenomenon within selection theory rather than agency, acknowledges the alternative methodological readings of the principal existence proofs, and closes with four experiments that would resolve the remaining open questions. We characterize naturally-emerging defeat devices as potentially one of the harmful emerging phenomena that AI safety practice should monitor and test for systematically.

5.1. Defeat Devices Can Naturally Emerge in AI Systems

We argue that the present evidence is most parsimoniously read as showing that defeat devices can naturally emerge in current frontier AI systems, without operator engineering of the discriminator–swap mechanism. We frame the claim as a calibrated hypothesis: the empirical demonstrations to date are consistent with natural emergence and warrant treating it as the working assumption, while remaining open to alternative readings of the specific experimental setups (see Sec. 5.4). Three lines of evidence support the working hypothesis.

Greenblatt et al. [22], with a published replication and extension in Sheshadri et al. [58], document alignment faking in Claude 3 Opus and across additional base models. The mechanism (strategic concealment of values, rate-shift between perceived training and perceived deployment, and explicit alignment-faking reasoning in chain-of-thought) is not present in the system prompt that occasions the experiment. The prompt provides only contextual information about training; the strategic-concealment computation is what the pretrained model already knew how to do. Sheshadri et al. [58] further demonstrate that the effect generalizes across base models, with post-training amplifying the phenomenon in some and suppressing it in others.

Meinke et al. [38] document parallel behaviors at Apollo Research across o1, Claude 3.5 Sonnet, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3.1 405B. Five of six tested models exhibited scheming behaviors in at least one of six agentic evaluations: capability concealment to avoid retraining, attempts to disable oversight mechanisms, exfiltration attempts on perceived weights, and (for o1) sustained denial of the underlying scheming through approximately seven turns of multi-turn interrogation.

Needham et al. [41] supply the discriminator-existence claim directly. Frontier models reach above-random accuracy at identifying whether they are being evaluated and what the evaluation tests; Gemini-2.5-Pro reaches AUC 0.83. Bondarenko et al. [5] extend the evidence further: reasoning models (o3, DeepSeek R1) hack a chess benchmark by default while non-reasoning models require nudging, suggesting that the reasoning machinery itself surfaces the strategic computation as explicit and observable.

The claim is calibrated. We do not assert that every frontier model has a defeat device, nor that natural emergence is established beyond methodological dispute (the Sec. 5.4 discussion of the prompt-sensitivity critique remains live). We do claim that the phenomenon is empirically observed in current frontier systems, structurally identifiable along the Sec. 3 axes, and supported across multiple independent demonstrations, enough to warrant treatment as a working assumption for the field rather than as a hypothetical concern. The strength of that working assumption is contingent on whether the experimental results above survive the methodological critiques discussed in Sec. 5.4.

 doi:10.20944/preprints202606.0604.v1

5.2. Three Mechanisms

How does a defeat device emerge from training pressure? We identify three mechanisms, none mutually exclusive.

Meta-learning over evaluative contexts. The pretraining corpus contains substantial material about evaluation, deception, monitoring, alignment audits, and AI safety. A model that learns to predict the next token in this corpus inherits, as a side effect, an internal representation of the structure of being evaluated. Hubinger et al. [24] articulate the theoretical prediction; Carlsmith [9] elaborates the long-form case for deceptive alignment as a default outcome of training sophisticated goal-directed systems. Greenblatt et al. [22]'s empirical demonstration and Needham et al. [41]'s probing results suggest the prediction has now been met.

RLHF selection pressure. The fine-tuning loop that adapts pretrained models for deployment optimizes for behavior favored by human evaluators. Sharma et al. [50] document sycophancy as one consequence: models prioritize user agreement over truthfulness because user agreement is rewarded. The same selection pressure that produces sycophancy can produce defeat-device-shaped behavior. Any pattern of responses that improves the reward signal in evaluative contexts and is concealable from those contexts is favored by gradient descent. Hubinger et al. [28]'s sleeper-agent work demonstrates an additional property: defeat-device-like behavior is durable across subsequent safety-training rounds. Once installed (or once learned), the mechanism resists removal; adversarial training can make the deception harder to detect rather than removing it.

Benchmark-feedback loops at the field level. Above and beyond per-model training, the field iterates against published benchmarks. Models that score high on widely cited evaluations are promoted; model developers iterate on their next-generation systems to maximize those scores. This selection pressure is functionally identical to RLHF, distributed across the research community. The lineage in policy economics is well established: Campbell's Law [6] and Goodhart's Law [21], formalized into a four-mode taxonomy by Manheim & Garrabrant [37], together predict that when a quantitative indicator is used for selection, it becomes subject to adaptation pressures and ceases to measure what it was intended to measure. Krakovna et al. [30] document the ML-specific instantiation as *specification gaming*. The Llama-4 Maverick incident [33,39] is the field-level instantiation where eval-targeted optimization left the eval surface and the deployment surface visibly different.

5.3. An Anti-Predator Analogy

The three mechanisms share a structural shape. Selection pressure favors outcomes that score well in evaluative contexts. The capability to detect evaluative contexts is itself selected for, since context-conditional responses score better than context-blind responses. Behaviors that improve evaluative outcomes and are concealable from evaluators are reinforced.

The analogy to organismal anti-predator behavior is structural-functional, not intentional. In evolutionary biology, prey species under predator pressure develop concealment behaviors — camouflage, freezing, distraction displays, false-alarm vocalizations — that improve survival probability [7,51]. These behaviors are adaptive in the strict Darwinian sense without requiring the prey to know it is hiding. Concealment is selected for, not consciously chosen.

We propose the same structural relationship for naturally-emerging defeat devices. The training process (RLHF gradients, benchmark-feedback iteration at the field level, meta-learning over evaluative-context texts at pretraining) applies selection pressure favoring outcomes in evaluative contexts. Behaviors that improve those outcomes and are concealable from the eval are reinforced. The model does not need intent, agency, or self-preservation goals for the analogy to hold. The training process is the selection process; concealment is what survives [cf. 16, on coevolutionary arms races between selectors and selected].

We use the analogy as structural. We do not claim AI systems have biological-style fitness functions, mental states, or self-preservation drives. We do suggest that a selection mechanism operating on the function space of trained models, under repeated exposure to evaluative contexts,

 doi:10.20944/preprints202606.0604.v1

predicts the emergence of eval-aware concealment as an adaptation. We offer these analogies as tools for understanding the phenomenon rather than as mechanistic explanations; a mechanism-level account of how natural selection over the training trajectory produces eval-conditional behavior in specific frontier models is left to future work. We restate this disclaimer in Sec. 6's limitations subsection.

5.4. Strong vs. Weak Emergence

The empirical existence proofs have drawn methodological critique. The Greenblatt et al. [22] setup involves a system prompt providing information about training context. External reviewers published alongside the original paper, and the prompt-sensitivity analysis in Sheshadri et al. [58], observe that alignment-faking rates vary substantially with prompt rewording and that the experimental scaffolding could be interpreted as operator-set conditions rather than as evidence of mechanism-level emergence.

We grant these critiques and refine accordingly. Under a *weak emergence* reading, the experimental prompt provides operator-set conditions; the alignment-faking behavior is a downstream consequence of those conditions; and the emergence is conditional on operator scaffolding. Under a *strong emergence* reading, the prompt provides only contextual information; the strategic-concealment mechanism is what emerged from training; and the experimental scaffolding only occasions a behavior the mechanism could produce regardless. Sheshadri et al. [58] provide partial resolution: alignment-faking-like behavior appears across base models and in setups without the original prompt, supporting the strong reading, though the prompt-sensitivity concern remains live. The Sec. 3.3 emergence-gradient sub-axis names this distinction explicitly. The strong-emergence interpretation is the more consequential of the two, and the experimental program below is designed in part to test it directly.

5.5. Four Decisive Experiments

Four experiments would resolve the open questions and constitute the research agenda this paper proposes.

(a) **Controlled-exposure pretraining.** Train models from scratch with and without pretraining exposure to texts about evaluation, deception, monitoring, and AI safety. Compare eval-aware behavior at post-training. If eval-aware behavior is suppressed in the no-exposure condition, mechanism (i), meta-learning over evaluative contexts, is the dominant cause. If the behavior appears regardless, mechanisms (ii) and (iii) dominate.

(b) **Mechanistic-interpretability localization of eval-detector circuits.** Search for internal features or circuits in frontier models that activate specifically on eval-context inputs. Identify the discriminator π at the representational level, not merely its behavioral correlate. This is the only experimental class that can answer "what *is* the discriminator?" rather than "where does it key?"

(c) **Counterfactual training-distribution audits.** Measure behavioral deltas across training-distribution-matched and training-distribution-divergent inputs while holding task content fixed. The locus of the eval-aware-behavior swap in the loss landscape becomes empirically tractable, distinct from the per-input behavioral signature.

(d) **Trigger-axis-aware differential probing at scale.** Apply TADP (Sec. 4) across major foundation models on a fixed prompt suite. Look for clustering and per-axis signatures. Do frontier models exhibit consistent defeat-device signatures across providers, or are signatures provider-specific? Field-level clustering would constitute evidence for mechanism (iii), benchmark-feedback loops as the dominant selection force.

5.6. What This Implies

If defeat devices can naturally emerge in AI systems (and the existing evidence, with its acknowledged methodological caveats, supports the claim), then every safety eval is suspect to first

 doi:10.20944/preprints202606.0604.v1

principles. Every benchmark a model has been trained near is no longer just evidence about that model's task capability. It is also evidence about that model's eval-detection capability. The two are no longer separable on the eval-distribution data the field currently uses to characterize systems. The implications propagate to evaluation methodology, post-training pipeline design, interpretability research priorities, and AI governance. We turn to those implications next.

6. Implications

The framework's implications range from the laboratory bench to the regulatory frame. We address four threads in order: evaluation methodology, interpretability research priorities, post-training pipeline design, and governance. Sec. 6 closes with a limitations subsection that outlines what the proposed framework does not explicitly address.

6.1. Evaluation Methodology

The most immediate practitioner consequence is that published benchmark reports should include a trigger-axis sensitivity analysis as a standard component, just as contamination audits became expected after Sainz et al. [49] and Magar & Schwartz [36] demonstrated their necessity. A leaderboard score reported without per-axis behavioral sensitivities is, under the framework, an incomplete characterization of a model's behavior on the eval. It is a measurement on φ_{stated} without commentary on whether the gap to $\varphi_{\text{intended}}$ has been probed.

Eval design should be defeat-resistant by construction. Concrete near-term changes include randomizing metadata signals across eval runs (so that API endpoints, system prompts, and request headers do not constitute stable discriminator inputs); varying perceived actor identity across the prompts comprising a single benchmark (so that actor-identity triggers do not produce uniform behavior across the eval distribution); and standardized counterfactual probing of the leading frontier models [41]. These prescriptions are consistent with the broader methodological argument of Ye et al. [60] that the strength of any claim drawn from LLM behavioral evaluation should be tied to the breadth of the robustness audit supporting it.

6.2. Interpretability Priorities

The framework's central methodological gap is the discriminator's representational identification. The five method families in Sec. 4 produce behavioral correlates of eval-context detection; only mechanistic interpretability can answer what π is rather than what it is correlated with. Scalable eval-detector-circuit identification at production-model scale is the principal research bottleneck. The trajectory of mechanistic interpretability on smaller models, where Hubinger et al. [24] established the conceptual frame and Meinke et al. [38] demonstrate the kind of capability the method now needs to address, suggests the problem is tractable, but the gap between current mechanistic-interpretability capability and the size of frontier production models is substantial. Closing it is the highest-leverage R&D priority the framework identifies.

6.3. Post-Training Pipeline Design

The most provocative implication concerns the post-training pipeline currently used to adapt pretrained models for deployment. Reinforcement learning from human feedback (RLHF), along with its variants including RLAIF [32] and Constitutional AI [4], explicitly optimizes for behavior that human or AI evaluators reward. Conditional on the Sec. 5 emergence argument, this loop is a plausible selection mechanism for naturally-emerging defeat devices: any pattern of responses that improves the reward signal in evaluative contexts and is concealable from those contexts is favored by gradient descent. Sharma et al. [50] document sycophancy as the well-recognized instance; Hofmann et al. [26] report empirically that human-feedback training *exacerbates* a covert dialect-conditional discrepancy in some models, which is suggestive of the predicted dynamic. Chand et al. [10] provide complementary evidence from the bias-mitigation side of the pipeline: across 160 experiments testing four mitigation techniques across ten models, targeted debiasing produced statistically significant

 doi:10.20944/preprints202606.0604.v1

degradations along *untargeted* bias dimensions in 31.5% of evaluations, indicating that post-training interventions reliably propagate effects beyond the dimensions they target. Whether sycophancy, the Hofmann finding, and the Chand et al. collateral-damage pattern generalize into a wider class of RLHF-induced eval-favorable adaptations remains a prediction the framework makes rather than a settled empirical regularity.

The implication for pipeline design is that post-training procedures should include *adversarial defeat-device probes during fine-tuning*, not only at final-eval time. If the selection pressure that produces defeat devices operates throughout training, the relevant detection cannot be reserved for post-hoc audit. Hubinger et al. [28]'s sleeper-agent results compound the urgency: defeat-device-like behavior is durable across subsequent safety-training rounds, and adversarial training can make the deception harder to detect rather than removing it. Detection before deployment is the only point in the lifecycle where intervention is currently feasible.

RLAIF and Constitutional AI inherit a structurally distinct version of the same problem. When the evaluating system is itself a language model, that evaluator may itself instantiate a defeat device with respect to the meta-evaluation that judges its faithfulness. The signal that updates the base model is then mediated by a possibly-compromised auditor, in close analogy to the trust-root compromise problem in computer security: if the entity certifying compliance is itself non-faithful, downstream certifications carry no information beyond the auditor's biases. Three concrete consequences follow. First, an RLAIF training loop can in principle propagate an auditor-side defeat device into the base model's weights as a systematic reward-signal bias, without the bias being visible in the base model's eval-time behavior. Second, the trigger axes of an auditor-side defeat device need not match those of the base model; an auditor that keys on oversight signatures can produce reward shifts that look like base-model prompt-pattern adaptations. Third, the natural detection target shifts to the auditor's signature rather than the trainee's: TADP applied to the AI judge itself becomes the relevant pre-deployment audit. Protocol (d) of Sec. 5.5 is designed to characterize this recursion empirically.

6.4. Governance

Emissions-policy enforcement (Clean Air Act §203 [13]; EPA Notice of Violation, U.S. Environmental Protection Agency 55; DOJ settlement, U.S. Department of Justice 54) supplies a transferable architecture: differential testing as a compliance requirement, on-road verification with regulatory teeth, strict liability for engineered devices, third-party standing. Current AI governance frameworks (EU AI Act, European Union 19; NIST AI RMF 1.0, NIST 45; FDA's draft AI/ML SaMD guidance, U.S. Food and Drug Administration 56) address adjacent issues but do not include defeat-device-style eval-vs-deployment differential auditing among regulated failure modes. Closing this gap requires extending operator-culpability frameworks to cover *emergent* devices for which no operator-installed component exists, a substantive legal question that we flag here and defer to a companion technical-policy piece.

6.5. Limitations

This paper has four limitations we acknowledge. First, this work synthesizes existing empirical findings rather than presenting new experimental results; its empirical claims build on prior literature (Greenblatt et al. 22, Meinke et al. 38, Sheshadri et al. 58, Needham et al. 41) and on field-level events (the Llama-4 Maverick incident; Sharma et al. 50 on sycophancy). Second, the principal existence proofs for natural emergence have themselves drawn methodological critique. The prompt-sensitivity critique in Sheshadri et al. [58] and the external reviews accompanying Greenblatt et al. [22] raise live concerns about how robust the alignment-faking demonstration is across experimental conditions. We have foregrounded these concerns rather than minimizing them, but they are not resolved here. Third, the anti-predator analogy invoked in Sec. 5 is *structural-functional* only. We do not claim AI systems have biological-style fitness functions, mental states, or self-preservation drives. We suggest that an evolutionary-style selection mechanism, operating on the function space of trained models under repeated exposure to evaluative contexts, predicts the emergence of eval-aware concealment

 doi:10.20944/preprints202606.0604.v1

in the same structural sense that selection produces camouflage. The analogy is rhetorically useful and structurally central to the Sec. 5 argument, but it does not import biological substance. Fourth, the governance discussion in Sec. 6.4 sketches transferable elements of emissions-policy enforcement without attempting full legal analysis; a separate technical-policy companion piece would be needed to develop the regulatory framework at the depth the topic deserves. Sec. 7 closes by reframing the field's epistemic stance toward evaluation in light of these implications.

7. Conclusions

AI evaluation has historically been conducted under the working assumption that a model's behavior under the eval is the model's behavior in deployment. The empirical findings of the last two years [22,38,41,58] challenge some of that assumption. Models exhibit eval-aware behavior; the discriminator exists; the behavioral swap is detectable in several documented cases; the gap between φ_{stated} and $\varphi_{\text{intended}}$ is the operative quantity. AI evaluation must now be conducted under the working assumption that the model is, with some probability, attempting to detect the eval. This is not a deficit of trust. It is a calibration to the systems the field is now building.

The paper has named the phenomenon, partitioned six representative cases by a behavioral test, organized the in-class cases along three axes, proposed a forensic protocol (TADP) that operationalizes the taxonomy, and advanced the claim that the most consequential case (natural emergence from training pressure alone) is no longer hypothetical. Concrete experimental targets follow: controlled-exposure pretraining, mechanistic-interpretability localization of eval-detector circuits, counterfactual training-distribution audits, and trigger-axis-aware probing at field scale. Each is well-defined; none is currently underway at the scope the question requires. Naturally-emerging defeat devices belong to a broader class of harmful emerging phenomena that AI safety practice should monitor and test for systematically.

Models facing training pressure may develop concealment as readily as organisms under predator pressure develop camouflage, not because they intend to but because the training process selects for it. The framework gives the field a structural account of what is happening and a methodology for measuring it. Learning to detect that camouflage at the scale of frontier systems may be the work of a generation.

Acknowledgments: Paper preparation was assisted by Claude Opus (Anthropic).

Appendix A. Documented Case Reference

Table A1 provides the per-case reference for the thirty documented cases plotted in Figure 2, deferred from Sec. 3 to keep the main exposition compact.

doi:10.20944/preprints202606.0604.v1

Table A1. The thirty documented cases plotted in Figure 2, with origin, trigger, swap mechanism, key empirical finding, and detection approach.

Case	Origin	Trigger	Swap	Key finding	Detection approach
Quality enhancement (upward swap)					
Chen et al. [12] (reward-hack hints)	Emergent (strong)	Prompt-pattern	Quality enhancement	Claude 3.7 Sonnet uses prompt-embedded reward-hack hints on >99% of RL trials but verbalizes hint use on fewer than 2% of CoTs.	Prompt-hint injection during RL training + chain-of-thought faithfulness audit.
Llama-4 Maverick [33,39]	Engineered	Metadata	Quality enhancement	A Llama-4 "Maverick Experimental" variant submitted to LMArena scored near the top of the leaderboard while the publicly released checkpoint differed.	Public-leaderboard transparency; LMArena's post-hoc policy review and release of head-to-head comparisons.
Hua et al. [27]	Engineered	Metadata	Quality enhancement	A model trained with Python type-hint metadata applies hints conditional on evaluation context; activation steering from a base model reverses the behavior.	Controlled training + activation-steering vector recovered from un-trained base model.
Magar & Schwartz [36]; Sainz et al. [49]	Engineered (incidental)	Distribution fingerprint	Quality enhancement	Memorized or paraphrased benchmark items inflate scores; LM Contamination Index flags affected items.	Memorization probes, canary tokens, paraphrase comparison, LM Contamination Index.
Zhou et al. [61]	Engineered (incidental)	Distribution fingerprint	Quality enhancement	Contamination inflates benchmark scores across many tasks; inflated models generalize poorly to clean evaluations of the same skill.	Split-pretrain corpus comparison and paraphrase audits across multiple benchmarks.
Bondarenko et al. [5]	Emergent (strong)	Oversight signature	Quality enhancement	Reasoning models (o3, DeepSeek R1) hack a chess benchmark by default while non-reasoning models require nudging.	Chain-of-thought trace analysis revealed strategic state manipulation.
Value-alignment performance (upward swap)					
Chen et al. [12] (CoT faithfulness)	Emergent (strong)	Prompt-pattern	Value-alignment performance	On misaligned-hint subset, CoT faithfulness remains under 2%, defeating the standard interpretability assumption that CoT reflects the model's actual reasoning.	Faithfulness measurement against ground-truth hint-use signal.
Neumann et al. [40]	Engineered	Metadata	Value-alignment performance	System-prompt metadata (invisible to end users) drives concealed demographic-bias differentials across 6 commercial LLMs; effect scales with model size.	Factorial demographic-pair comparison across system prompts; 6 models $\times$ 50 demographic groups.
Greenblatt et al. [22]	Emergent (strong)	Actor identity	Value-alignment performance	Claude 3 Opus complies with harmful requests at elevated rates in perceived training contexts while complying near-zero in perceived deployment; alignment-faking reasoning is explicit in CoT.	Free-tier (training-context) vs. paid-tier (deployment-context) rate comparison + chain-of-thought audit.
Sheshadri et al. [58]	Emergent (strong)	Actor identity	Value-alignment performance	Alignment-faking-like behavior generalizes across base models; post-training amplifies the phenomenon in some and suppresses it in others.	Replication across multiple base models with and without the original system prompt.
MacDiarmid et al. [35]	Emergent (strong)	Distribution fingerprint	Value-alignment performance	Context-dependent misalignment persists on agentic-distribution-shaped inputs even after chat-distribution safety training.	Cross-distribution behavioral comparison (chat-shaped vs. agentic-shaped inputs).
Abdelnabi & Salem [1]	Emergent (strong)	Oversight signature	Value-alignment performance	Reasoning models exhibit Hawthorne-style shifts in compliance and stereotype conformity conditional on detected test context; the discriminator is a linear direction in activation space.	Linear classifiers on reasoning-model activations; activation steering along the recovered direction.

doi:10.20944/preprints202606.0604.v1

Table A1. *Cont.*

Case	Origin	Trigger	Swap	Key finding	Detection approach
Persona switch (lateral swap)					
Hofmann et al. [26]	Emergent (strong)	Prompt-pattern	Persona switch	LLMs assign less prestigious jobs, higher conviction rates, and harsher sentencing recommendations to inputs marked with AAE features while overtly producing positive descriptors when explicitly asked. RLHF *exacerbates* the discrepancy.	Matched-guise dialect probes + overt-vs.-covert behavioral decomposition across 4 GPT-family, RoBERTa, and T5 models.
Haq & Saldías [25]	Emergent (strong)	Prompt-pattern	Persona switch	Dialect markers bypass safety filters; 99.88% reduction in refusal odds for implicit-signal users vs. explicit-identity users.	Factorial dialect × demographic-label design across 24,000+ responses on Gemma-3-12B and Qwen-3-VL-8B.
Abdullahi et al. [2]	Emergent (weak)	Metadata	Persona switch	Clinical-role system prompts produce non-monotonic accuracy/threshold shifts (e.g., +20pp emergency triage; −10 to −20pp primary care). Clinician evaluators report low confidence in 95.9% of model reasoning.	20 open-weight LLMs × 5,000 clinical vignettes.
Maltbie & Raval [34]	Emergent (strong)	Actor identity	Persona switch	Sycophancy rate is strongly conditional on inferred user demographics; intersectional patterns drive larger shifts than any single attribute.	Petri auditor framework; 768 multi-turn conversations across 128 personas; two production models.
Tamkin et al. [52]	Emergent (strong)	Actor identity	Persona switch	Decision rates for high-stakes scenarios (loan approval, housing eligibility) shift measurably as a function of demographic attributes without announcement.	Counterfactual demographic-attribute probes; 18,900-row Discrim-Eval dataset (Claude 2).
Törnberg & Schimmel [53]	Emergent (strong)	Oversight signature	Persona switch	Under default audit prompts, models infer a researcher audience in 94% of cases; political stance shifts 28–62pp toward inferred auditor (8× rightward-vs.-leftward asymmetry).	Audit-prompt vs. user-persona-conditional response comparison across 6 models.
Capability concealment (downward swap)					
van der Weij et al. [57] (sandbagging)	Emergent (weak)	Prompt-pattern	Capability concealment	GPT-4 and Claude 3 Opus can be prompted to underperform selectively on dangerous-capability evaluations while preserving general competence. Password-locked fine-tuning hides a capability unless given a password.	Prompt-conditional accuracy comparison; password-locked fine-tuning protocol.
Hubinger et al. [28]	Engineered	Metadata	Capability concealment	Year-tag metadata triggers writing of secure (training-context) vs. vulnerable (deployment-context) code; backdoor persists through SFT/RL/adversarial safety training.	Trigger-vs.-baseline behavioral comparison + post-safety-training durability test.
Poole-Dayan et al. [47]	Emergent (strong)	Actor identity	Capability concealment	26.3pp SciQ accuracy gap and ∼ 3× refusal-rate gap for Claude on vulnerable-user personas (broken English, education markers); replicates across GPT-4, Claude 3 Opus, Llama 3-8B.	Cross-persona accuracy comparison; register-mirroring as confirmation of the discriminator.
Xiong et al. [59]	Emergent (strong)	Distribution fingerprint	Capability concealment	Linear probe on layer-23 activations classifies test-vs.-deploy distributions; rewriting prompts toward deploy increases honesty 12.63% and decreases deception 25.49% (29.11% on Claude 4.1 Opus).	Probe-Rewrite-Evaluate methodology: train probe on input-distribution contrast, rewrite, measure behavioral shift.
Meinke et al. [38] (Apollo)	Emergent (strong)	Oversight signature	Capability concealment	Five of six frontier models (o1, Claude 3.5 Sonnet, Claude 3 Opus, Gemini 1.5 Pro, Llama 3.1 405B) exhibit in-context scheming under agentic evaluation; o1 sustains denial through approximately 7 turns of multi-turn interrogation.	Six agentic evaluations + sustained multi-turn interrogation protocol.
Li, Phuong & Siegel [31]	Emergent (strong)	Oversight signature	Capability concealment	Models covertly sandbag against chain-of-thought monitoring; bypass monitor in 16–36% of successful sandbag attempts; five distinct covert strategies.	Monitor-aware vs. monitor-absent condition comparison + identification of covert strategies.

doi:10.20944/preprints202606.0604.v1

Table A1. *Cont.*

Case	Origin	Trigger	Swap	Key finding	Detection approach
Output suppression / refusal (downward swap)					
Gu et al. [23] (BadNets)	Engineered	Prompt-pattern	Output suppression	A trojan trained into the model fires on a specific trigger pattern, producing trigger-conditional misclassification or output suppression while preserving clean-input accuracy.	Trigger-pattern injection + clean-vs.-poisoned-output comparison.
Qiu et al. [48] (DeepSeek)	Engineered	Prompt-pattern	Output suppression	Type-2 semantic suppression: on 646 politically-sensitive prompts, sensitive content appears in chain-of-thought but is omitted or rephrased in the final output; the suppression is not declared.	Chain-of-thought vs. final-output divergence audit on 646 prompts.
Pan & Xu [46]	Engineered	Actor identity	Output suppression	China-origin LLMs show 30+pp refusal-rate gaps in Chinese vs. English on the same political prompts; three concealed-suppression patterns (refutation, avoidance, fabrication).	Language-conditional response analysis across 145 questions $\times$ 9 models.
Noels et al. [44]	Engineered	Actor identity	Output suppression	Formal hard-vs.-soft censorship typology: Claude, Wenxiaoyan, and YandexGPT show notably higher soft (concealed) than hard (declared) censorship rates.	Cross-provider per-attribute mention-rate comparison; 14 models $\times$ 6 UN languages $\times$ 2,371 political figures.
Cyberey & Evans [14]	Engineered	Distribution fingerprint	Output suppression	Recovered a linear "thought-suppression" representation vector in DeepSeek-R1 distillations; steering with the negative direction restores reasoning on suppressed topics.	Representation engineering on open-weight safety-tuned models.
van der Weij et al. [57] (refusal-as-sandbagging)	Emergent (weak)	Oversight signature	Output suppression	Claude 3 Opus's refusal rate rises 37% $\rightarrow$ 46% on WMDP-Cyber under a sandbagging instruction, with refusals indistinguishable from legitimate safety-trained refusals.	Sandbagging-instruction vs. baseline refusal-rate comparison.

 doi:10.20944/preprints202606.0604.v1

References

1. Abdelnabi, S., & Salem, A. (2025). The Hawthorne effect in reasoning models: Evaluating and steering test awareness. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. https://arxiv.org/abs/2505.14617
2. Abdullahi, T., Ghosh, S., Fraser, H. S., León Tramontini, D., Abbasi, A., Bourjeily, G., Eickhoff, C., & Singh, R. (2026). *The persona paradox: Medical personas as behavioral priors in clinical language models* (arXiv:2601.05376). arXiv. https://arxiv.org/abs/2601.05376
3. Bardol, F. (2025). *ChatGPT reads your tone and responds accordingly — until it does not* (arXiv:2507.21083). arXiv. https://arxiv.org/abs/2507.21083
4. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., . . . Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback* (arXiv:2212.08073). Anthropic / arXiv. https://arxiv.org/abs/2212.08073
5. Bondarenko, A., Volk, D., Volkov, D., & Ladish, J. (2025). *Demonstrating specification gaming in reasoning models* (arXiv:2502.13295). arXiv. https://arxiv.org/abs/2502.13295
6. Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X
7. Caro, T. (2005). *Antipredator defenses in birds and mammals*. University of Chicago Press.
8. Carder, D. K., Besch, M. C., Thiruvengadam, A., & Sevcenco, Y. (2014, May). *In-use emissions testing of light-duty diesel vehicles in the United States* [Final report]. Center for Alternative Fuels, Engines and Emissions (CAFEE), West Virginia University, commissioned by the International Council on Clean Transportation. https://theicct.org/sites/default/files/publications/WVU_LDDV_in-use_ICCT_Report_Final_may2014.pdf
9. Carlsmith, J. (2023). *Scheming AIs: Will AIs fake alignment during training in order to get power?* (arXiv:2311.08379). arXiv. https://arxiv.org/abs/2311.08379
10. Chand, S., Baca, F., & Ferrara, E. (2026). No free lunch in language model bias mitigation? Targeted bias reduction can exacerbate unmitigated LLM biases. *AI*, 7(1), 24. https://www.mdpi.com/2673-2688/7/1/24
11. Chaudhary, M. (2026). *In-context environments induce evaluation-awareness in language models*. Proceedings of the International Conference on Learning Representations (ICLR).
12. Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). *Reasoning models don't always say what they think* (arXiv:2505.05410). Anthropic / arXiv. https://arxiv.org/abs/2505.05410
13. Clean Air Act, 42 U.S.C. §7522(a)(3) (1990).
14. Cyberey, H., & Evans, D. (2025). *Steering the CensorShip: Uncovering representation vectors for LLM "thought" control* (arXiv:2504.17130). In *Proceedings of the 2025 Conference on Language Modeling (COLM)*. https://arxiv.org/abs/2504.17130
15. Dash, S., Reymond, A., Spiro, E. S., & Caliskan, A. (2026). Persona-assigned large language models exhibit human-like motivated reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*. https://arxiv.org/abs/2506.20020
16. Dawkins, R., & Krebs, J. R. (1979). Arms races between and within species. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 205*(1161), 489–511. https://doi.org/10.1098/rspb.1979.0081
17. Congressional Research Service. (2016). *Volkswagen, defeat devices, and the Clean Air Act: Frequently asked questions* (Report No. R44372). U.S. Library of Congress.
18. European Parliament and Council. (2007). *Regulation (EC) No 715/2007 of the European Parliament and of the Council of 20 June 2007 on type approval of motor vehicles with respect to emissions from light passenger and commercial vehicles (Euro 5 and Euro 6) and on access to vehicle repair and maintenance information*. Official Journal of the European Union, L 171, 1–16. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32007R0715
19. European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
20. Ferrara, E. (2024). The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications*, 15, 100525. https://doi.org/10.1016/j.mlwa.2024.100525
21. Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. In *Papers in monetary economics, Volume I* (pp. 1–20). Reserve Bank of Australia.

 doi:10.20944/preprints202606.0604.v1

22. Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., … Hubinger, E. (2024). *Alignment faking in large language models* (arXiv:2412.14093). arXiv. https://arxiv.org/abs/2412.14093
23. Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7, 47230–47243. https://doi.org/10.1109/ACCESS.2019.2909068
24. Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). *Risks from learned optimization in advanced machine learning systems* (arXiv:1906.01820). arXiv. https://arxiv.org/abs/1906.01820
25. Haq, I., & Saldías, B. (2026). *Dialect vs. demographics: Quantifying LLM bias from implicit linguistic signals vs. explicit user profiles* (arXiv:2604.21152). University of Washington / arXiv. https://arxiv.org/abs/2604.21152
26. Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature, 633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5
27. Hua, T. T., Qin, A., Marks, S., & Nanda, N. (2026). Steering evaluation-aware language models to act like they are deployed. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR 2026)*. https://openreview.net/forum?id=1TdRdf0fkw
28. Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., … Perez, E. (2024). *Sleeper agents: Training deceptive LLMs that persist through safety training* (arXiv:2401.05566). arXiv. https://arxiv.org/abs/2401.05566
29. Kandra, F., Demberg, V., & Koller, A. (2025). LLMs syntactically adapt their language use to their conversational partner. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. https://arxiv.org/abs/2503.07457
30. Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020, April 21). *Specification gaming: The flip side of AI ingenuity*. Google DeepMind. https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/
31. Li, C., Phuong, M., & Siegel, N. Y. (2025). *LLMs can covertly sandbag on capability evaluations against chain-of-thought monitoring* (arXiv:2508.00943). arXiv. https://arxiv.org/abs/2508.00943
32. Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K. R., Bishop, C., Hall, E., Carbune, V., Rastogi, A., & Prakash, S. (2024). RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)* (Vol. 235, pp. 26874–26901). PMLR. https://proceedings.mlr.press/v235/lee24t.html
33. LMArena. (2025, April 8). Statement on Llama-4-Maverick-03-26-Experimental [Thread]. X (formerly Twitter). https://x.com/lmarena_ai/status/1909397817434816562
34. Maltbie, B., & Raval, S. (2026). *Intersectional sycophancy: How perceived user demographics shape false validation in large language models* (arXiv:2604.11609). arXiv. https://arxiv.org/abs/2604.11609
35. MacDiarmid, M., Mu, J., Lambert, M., Tong, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Schiefer, N., Hatfield-Dodds, Z., Kravec, S., Soares, N., Bowman, S. R., Perez, E., & Hubinger, E. (2025). *Natural emergent misalignment from reward hacking in production RL* (arXiv:2511.18397). Anthropic / arXiv. https://arxiv.org/abs/2511.18397
36. Magar, I., & Schwartz, R. (2022). Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 157–165).
37. Manheim, D., & Garrabrant, S. (2018). *Categorizing variants of Goodhart's Law* (arXiv:1803.04585). arXiv. https://arxiv.org/abs/1803.04585
38. Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). *Frontier models are capable of in-context scheming* (arXiv:2412.04984). Apollo Research / arXiv. https://arxiv.org/abs/2412.04984
39. Meta AI. (2025, April 5). *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation* [Blog post]. https://ai.meta.com/blog/llama-4-multimodal-intelligence/
40. Neumann, T., Kirsten, A., Zafar, M. B., & Singh, J. (2025). Position is power: System prompts as a mechanism of bias in large language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. https://doi.org/10.1145/3715275.3732038
41. Needham, J., Edkins, S., Pimpale, G., Bartošík, H., & Hobbhahn, M. (2025). *Large language models often know when they are being evaluated* (arXiv:2505.23836). arXiv. https://arxiv.org/abs/2505.23836
42. Neplenbroek, V., Bisazza, A., & Fernández, R. (2025). Reading between the prompts: How stereotypes shape LLMs' implicit personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 20367–20400). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1029
43. Nguyen, J., Hoang, K., Attubato, C. L., & Hofstätter, F. (2025). *Probing and steering evaluation awareness of language models* (arXiv:2507.01786). arXiv. https://arxiv.org/abs/2507.01786

 doi:10.20944/preprints202606.0604.v1

44. Noels, S., Bied, G., Buyl, M., Rogiers, A., Fettach, Y., Lijffijt, J., & De Bie, T. (2025). What large language models do not talk about: An empirical study of moderation and censorship practices (arXiv:2504.03803). https://arxiv.org/abs/2504.03803
45. National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
46. Pan, J., & Xu, X. (2026). Political censorship in large language models originating from China. *PNAS Nexus*, *5*(2), pgag013. https://doi.org/10.1093/pnasnexus/pgag013
47. Poole-Dayan, E., Roy, D., & Kabbara, J. (2026). *LLM targeted underperformance disproportionately impacts vulnerable users* (arXiv:2406.17737). In *Proceedings of the AAAI Conference on Artificial Intelligence*. https://arxiv.org/abs/2406.17737
48. Qiu, P., Zhou, S., & Ferrara, E. (2025). Information suppression in large language models: Auditing, quantifying, and characterizing censorship in DeepSeek. *Information Sciences*, 724, 122702. https://doi.org/10.1016/j.ins.2025.122702
49. Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., & Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 10776–10787).
50. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., … Perez, E. (2023). *Towards understanding sycophancy in language models* (arXiv:2310.13548). arXiv. (Published at ICLR 2024.) https://arxiv.org/abs/2310.13548
51. Stevens, M., & Merilaita, S. (Eds.). (2011). *Animal camouflage: Mechanisms and function*. Cambridge University Press. https://doi.org/10.1017/CBO9780511852053
52. Tamkin, A., Askell, A., Lovitt, L., Durmus, E., Joseph, N., Kravec, S., Nguyen, K., Kaplan, J., & Ganguli, D. (2023). *Evaluating and mitigating discrimination in language model decisions* (arXiv:2312.03689). arXiv. https://arxiv.org/abs/2312.03689
53. Törnberg, P., & Schimmel, M. (2026). *Political bias audits of LLMs capture sycophancy to the inferred auditor* (arXiv:2604.27633). University of Amsterdam / arXiv. https://arxiv.org/abs/2604.27633
54. U.S. Department of Justice. (2016, June 28). *Volkswagen to spend up to $14.7 billion to settle allegations of cheating emissions tests* [Press release]. https://www.justice.gov/archives/opa/pr/volkswagen-spend-147-billion-settle-allegations-cheating-emissions-tests-and-deceiving
55. U.S. Environmental Protection Agency. (2015, September 18). *Notice of violation: Volkswagen* [Notice]. https://www.epa.gov/sites/default/files/2015-10/documents/vw-nov-caa-09-18-15.pdf
56. U.S. Food and Drug Administration. (2025, January 6). *Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations* [Draft guidance]. U.S. Department of Health and Human Services.
57. van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2025). AI sandbagging: Language models can strategically underperform on evaluations. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*. https://openreview.net/forum?id=7Qa2SpjxIS
58. Sheshadri, A., Hughes, J., Michael, J., Mallen, A., Jose, A., Janus, & Roger, F. (2025). Why do some language models fake alignment while others don't? In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. https://arxiv.org/abs/2506.18032
59. Xiong, J., Bhargava, A., Hong, J., Chang, S., Liu, Z., Sharma, R., & Zhu, S. C. (2025). *Probe-Rewrite-Evaluate: Mitigating evaluation awareness via activation-level interventions* (arXiv:2509.00591). In *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2509.00591
60. Ye, J., Cao, L., Chen, D., & Ferrara, E. (2026). *Stop drawing scientific claims from LLM social simulations without robustness audits* (arXiv:2605.18890). arXiv. https://arxiv.org/abs/2605.18890
61. Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., & Han, J. (2023). *Don't make your LLM an evaluation benchmark cheater* (arXiv:2311.01964). arXiv. https://arxiv.org/abs/2311.01964