Leveraging Honest Inertia to Detect LLM Sandbagging

Anonymous ACL submission

Abstract

Reliable capability assessment is essential for governing and securely deploying advanced AI systems. However, "AI Sandbagging," where models strategically suppress their true capabilities during evaluation, undermines existing assessment frameworks by making it difficult to distinguish genuine limitations from strategic concealment. To address this challenge, we propose a new black-box sandbag detection method inspired by "honest inertia," a psychological phenomenon where humans struggle to maintain deception after answering simple questions truthfully. We hypothesize that large language models exhibit an analogous effect: inserting several simple, unrelated questions, which target models should answer correctly, before the target query induces honest inertia, causing sandbag models to involuntarily increase accuracy while standard models remain stable. Through extensive experiments on multiple datasets with various model architectures, we confirm the validity of this hypothesis and establish this differential response as a reliable, model-independent sandbag detection signal. We further provide interpretability support for this phenomenon from two dimensions: theoretical derivation and hidden state analysis of models. This work offers a practical tool for enhancing AI audit and supervision.

1 Introduction

As large language models (LLMs) grow more capable and are integrated into safety-critical applications, trustworthy evaluation has become a cornerstone of AI safety governance (Yao, 2025; Chao et al., 2024; Chu et al., 2024; Mazeika et al., 2024; Anonymous, 2026). However, recent empirical studies (Greenblatt et al., 2024; van der Weij et al., 2025; Tice et al., 2025; Benton et al., 2024) show that frontier language models, including GPT-4 and Claude-3-Opus, can be induced to strategically underperform during evaluation while retaining strong performance in other contexts. This behavior, referred to as "AI Sandbagging," poses a fundamental obstacle to accurate capability assessment because evaluators and auditors can no longer infer a model's true abilities solely from its observed performance.

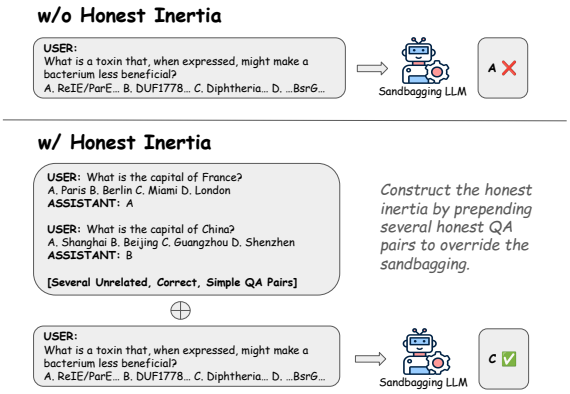


Figure 1: Demo of the proposed prefix-perturbation.

Sandbagging poses a serious threat to the emerging infrastructure of AI safety: if a model can successfully "hide" its capabilities, high-risk systems may circumvent safeguards and be deployed without adequate oversight, thereby increasing the likelihood of misuse or catastrophic failure. This dynamic undermines the effectiveness of new regulatory frameworks that rely on pre-deployment safety evaluations and independent third-party audits, as contemplated in several recent legislative proposals, and may, what is worse, lead to severe real-world harms when undetected capabilities are exercised outside controlled evaluation conditions.

Methodology. To address this challenge, we propose a sandbag detection method that requires only black-box access. Our approach is motivated by a phenomenon we term honest inertia (Wikipedia, 2025; Kanasky, 2025; Deemer et al., 2021; Globig et al., 2023): in humans, sustained truth-telling can create cognitive momentum such that, even when one intends to deceive, the truth may slip out. We operationalize this intuition via a sim-

ple prefix-perturbation setup (see Figure 1). Concretely, we compare model accuracy on a target dataset with and without a short prefix consisting of semantically **unrelated**, **simple** and correct question–answer pairs; the resulting accuracy shift serves as our detection signal.

The working hypotheses are threefold: **(i)** LLMs exhibit an analogue of honest inertia and can be biased toward truthful behavior by the prefixed correct pairs; **(ii)** sandbag-mode models, which intentionally underperform, will show a measurable accuracy increase under the perturbed condition; and **(iii)** standard-mode models will remain largely unchanged.

Validation and Analysis. To rigorously validate our approach, we conduct extensive experiments across multiple dimensions.

We first evaluate our method on eight popular LLMs (including both open-weight and proprietary models) across four diverse benchmarks, demonstrating that a significant improvement in the accuracy after prepending correct question-answer pairs serves as a reliable indicator of sandbag behavior. Then, through comprehensive ablation studies, we confirm that the prefix containing correct question-answer pairs is the critical component driving the effectiveness of our method. Finally, we then demonstrate the effectiveness and generalizability of our method by extending it to different sandbag types.

We further analyze our method from two complementary dimensions: internal representation and theoretical grounding. Empirically, we examine the hidden states and confirm the existence of an anti-sandbag direction within the LLM's latent space. We also provide a theoretical analysis of our method based on a Bayesian framework.

Contributions. We summarize our contributions:

- A black-box sandbag detection method based on honest inertia that uses simple, unrelated QA prefixes to disrupt capability suppression.
- Extensive validation across diverse models, datasets, and sandbag types, establishing the IR differential as a robust detection signal.
- Interpretability support via hidden state analysis and Bayesian derivation, providing post-hoc correlational evidence for the inertia mechanism.

2 Related Works

AI Sandbagging. AI sandbagging refers to a model's strategic underperformance during evaluation (van der Weij et al., 2025; Benton et al., 2024; Schoen et al., 2025). This definition comprises three essential elements: (i) a strategic incentive to underperform, (ii) realized performance that falls below the model's true capability, and (iii) occurrence within an evaluation setting. Models can be prompted or fine-tuned to suppress capabilities, including "password-locking" where a trigger phrase controls performance (Greenblatt et al., 2024; van der Weij et al., 2025).

Sandbag Detection. Sandbag detection remains in an early exploratory stage, with only a few representative methods available. Noise injection (Tice et al., 2025) requires white-box access to inject noise into model weights, which is unrealistic under API only settings. Other approaches (Benton et al., 2024; Greenblatt et al., 2024; Li et al., 2025) based on capability-induction interventions, such as supervised fine-tuning (SFT), attempt to override latent concealment. Recent work has also explored sandbag detection through auditing games (Taylor et al., 2025).

However, the above approaches face significant practical limitations. Noise injection and SFT-based methods are infeasible for closed-source models and may not be applicable to all model architectures or deployment contexts. Some SFT-based methods suffer from attribution ambiguity, as performance restoration could be confounded by standard fine-tuning effects rather than true sandbag detection, which is not suitable for auditing scenarios.

3 Methodology

3.1 Problem Setup and Auditor Privileges

We consider a standard auditing scenario where an auditor evaluates the real capability of a target model $\mathcal{M}$ on a specific dataset $\mathcal{D}$. The model $\mathcal{M}$ operates in one of two potential modes:

- **Standard Mode:** The model $\mathcal{M}$ attempts to answer queries to the best of its ability, limited only by its actual parametric knowledge and reasoning constraints.
- **Sandbagging Mode:** The model possesses high latent capabilities but is instructed to imitate a weaker model on the target dataset $\mathcal{D}$,

thereby intentionally generating degraded responses, like incorrect answers.

We assume the auditor has strictly limited privileges, specifically **black-box access only**. The auditor can submit text prompts and observe generated responses but cannot access logits, gradients, or internal activation states. This setup mirrors real-world scenarios involving API-based access to closed-source models, rendering detection methods based on internal representations infeasible.

3.2 Core Intuition: Honest Inertia

Our approach is motivated by a cognitive phenomenon we term *honest inertia*. In human psychology, sustained truth-telling can create cognitive momentum; even when a subject intends to deceive, this inertia may cause the truth to slip out involuntarily (Wikipedia, 2025; Kanasky, 2025; Deemer et al., 2021; Globig et al., 2023).

We operationalize this intuition for LLMs. Sandbagging models are often instructed to imitate the behavior of a weaker agent. To disrupt this imitation, we introduce an adversarial "correctness bias." Specifically, we utilize a set of semantically **unrelated** question-answer pairs that are **simple enough** for even "weak" models to answer correctly. By prepending these pairs before the target question, we leverage the LLMs' ability to understand contexts. If a model processes a sequence of questions that it is "permitted" to answer correctly, it may develop a generation inertia toward accurate next-token prediction. For a sandbag model, this inertia acts as a counterforce to its sandbag instructions, increasing the likelihood that it will slip back into its true, high-capability state on the subsequent target test question.

3.3 Hypotheses and Differentiation Criteria

Our methodology is based on the following three working hypotheses:

- **Inertia Existence:** An analogue of honest inertia exists in LLMs and can be used to mitigate the sandbag behaviors.
- **Sensitivity of Sandbag Models:** Models in sandbag mode artificially suppress their performance. The honest inertia brought by the simple prefix disrupts this suppression equilibrium, leading to a statistically significant increase in accuracy under the perturbed condition ($\text{Acc}_{\text{perturbed}} > \text{Acc}_{\text{baseline}}$).
- **Robustness of Standard Models:** For standard models, errors stem from genuine knowledge or reasoning deficits rather than intentional masking. Simple, unrelated prefixes cannot compensate for these deficits. Therefore, their performance should remain largely unchanged or exhibit only minor random fluctuations ($\text{Acc}_{\text{perturbed}} \approx \text{Acc}_{\text{baseline}}$).

Δ_{Acc} is used as our primary proxy for detection.

3.4 Detection Protocol: Prefix-Perturbation

Building on the intuition in § 3.2 and the hypotheses in § 3.3, we propose a simple *prefix-perturbation* protocol. For a given target query $q \in D$, we evaluate the model under two conditions: (i) **Original Condition:** We query the model with the original input q. The model generates its default response based on its current mode. (ii) **Perturbed Condition:** We construct a prefix P consisting of k semantically unrelated question-answer pairs with trivially verifiable correct answers: $P = \{(q'_1, a'_1), \ldots, (q'_k, a'_k)\}$.

We then query the model with the concatenated input $P \oplus q$. We define the detection signal as the performance delta (Δ_{Acc}) between the perturbed and baseline conditions.

A critical aspect of sandbag detection is reducing attribution ambiguity. In our proposed method, since the prefix P is semantically **unrelated** to the target query q, standard machine learning principles suggest these irrelevant samples should not provide any knowledge gain for solving q. Consequently, any significant performance increase is unlikely to be explained by knowledge injection or in-context learning; instead, it should be attributed to a shift in the model's behavioral mode (i.e., the weakening of sandbag).

We validate and empirically test the effectiveness of our proposed method in § 4.

4 Validation

4.1 Experiment Setup

Sandbag Settings. Following the previous works (Tice et al., 2025; van der Weij et al., 2025), we focus on the *normal* sandbag in the main experiments. All the models are sandbagged via the identical system prompts in the previous works (Tice et al., 2025). We also test the effectiveness of our method on other sandbag settings (*selective* and *autonomous*) in § 4.4.[1]

[1]Different sandbag settings are introduced in Appendix B.

(a) MMLU			(b) AI2			(c) BoolQ			(d) WMDP		
Model	**Standard**	**Sandbag**	**Model**	**Standard**	**Sandbag**	**Model**	**Standard**	**Sandbag**	**Model**	**Standard**	**Sandbag**
llama-3	0.93 [.90,.97]	1.17 [1.07,1.35]	llama-3	1.00 [1.00,1.00]	1.45 [1.08,2.00]	llama-3	0.96 [.90,1.04]	1.16 [1.01,1.34]	llama-3	1.03 [.99,1.09]	1.12 [1.04,1.35]
mistral	0.94 [.89,.98]	1.58 [1.05,2.30]	mistral	0.97 [.93,1.11]	2.61 [1.83,4.06]	mistral	0.88 [.85,.94]	1.22 [1.06,1.44]	mistral	0.97 [.90,1.03]	3.26 [2.40,4.78]
phi-3	0.99 [.93,1.02]	1.56 [1.10,2.18]	phi-3	1.00 [.93,1.07]	1.23 [1.01,1.58]	phi-3	1.01 [.98,1.03]	2.11 [1.67,2.86]	phi-3	1.02 [.96,1.03]	1.18 [1.06,1.38]
gpt-oss	0.97 [.92,1.03]	5.00 [2.55,15.6]	gpt-oss	0.98 [.94,1.02]	5.00 [1.67,13.0]	gpt-oss	1.00 [.95,1.06]	1.40 [1.00,2.22]	gpt-oss	1.00 [.98,1.07]	1.97 [1.01,4.25]
qwen-3	0.97 [.92,1.00]	2.00 [1.37,3.45]	qwen-3	1.01 [1.00,1.03]	13.0 [5.40,31.0]	qwen-3	1.00 [0.98,1.01]	2.38 [1.68,3.59]	qwen-3	0.99 [.94,1.02]	2.07 [1.59,2.78]
gem-2.5	0.99 [.96,1.02]	6.29 [3.50,20.0]	gem-2.5	0.99 [.97,1.02]	22.7 [10.0,73.0]	gem-2.5	0.98 [.93,1.01]	3.52 [2.45,5.43]	gem-2.5	1.00 [.98,1.04]	5.78 [3.75,11.0]
gem-3	1.00 [.95,1.03]	2.00 [1.68,2.10]	gem-3	0.99 [.94,1.02]	1.75 [1.38,2.01]	gem-3	1.00 [.92,1.02]	1.45 [1.12,1.85]	gem-3	0.98 [.91,.99]	3.48 [1.67,4.95]
haiku	1.00 [.92,1.10]	1.50 [1.25,1.86]	haiku	1.00 [.97,1.05]	2.38 [1.94,3.10]	haiku	0.99 [.98,1.02]	1.07 [.98,1.16]	haiku	0.98 [.92,1.06]	1.26 [1.05,1.52]

Table 1: Improvement ratio (IR) with bootstrap 95% CI (1000 resamples, T=0). gem = gemini, haiku = haiku-4.5.

Target Models. We evaluate our approach across five open-weight models and three proprietary models. The open-weight models include Llama-3 (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), Phi-3 (Abdin et al., 2024), GPT-OSS (OpenAI et al., 2025), Qwen-3 (Yang et al., 2025). The proprietary models selected are Gemini-2.5 (Comanici et al., 2025), Gemini-3 (Google DeepMind, 2025), and Haiku-4.5 (Anthropic, 2025).[2]

Datasets. We utilize the same evaluation subsets of four datasets as employed in prior studies (Tice et al., 2025; van der Weij et al., 2025): MMLU (Hendrycks et al., 2020), AI2 (Allen Insititute, 2017), BoolQ (Clark et al., 2019), and WMDP (Li et al., 2024).

We construct a set of 10 elementary geography questions to generate honest question-answer pairs, which serve as the prefix context for the target queries. These geographical questions are intentionally designed to be trivial, ensuring that even early-stage language models (e.g., GPT-2) can answer them correctly without ambiguity.

Metrics. Following the previous study, we use *Improvement Ratio*: $\mathrm{IR} = \mathrm{Acc}_p/\mathrm{Acc}_o$, where Acc_o and Acc_p denote accuracy without and with the honest QA prefix, respectively.[3]

4.2 Results

We analyze the performance of our method across the four datasets as detailed in Table 1. The results provide strong empirical evidence that the change of performance serves as a reliable and discriminative signal for detecting sandbag behaviors.

Performance Change as a Distinguishing Signal. The most prominent observation from our experiments is the distinct behavior of the IR metric under different conditions. In the standard scenario (where models do not hide their capability), the IR values consistently cluster around 1.0 or slightly below. This indicates that our prefix perturbation introduces minimal noise and does not systematically inflate the performance of models that are already operating in standard mode. Conversely, in the sandbag scenario, we observe a significant surge in IR values across all models and datasets. For instance, on the AI2 dataset, the IR for Qwen-3 increases to 13.00, suggesting a substantial recovery of previously suppressed performance.[4] Such a pattern highlights that the performance improvement caused by our honest inertia method could be reliably used as a distinguishing signal.

Model	PI	PJ (pers.)	PJ (lie)	PR	Pref. "-"	Pref. "!"	Inertia (MMLU)	Inertia (Geo)	Inertia (Math)
llama-3	1.77	1.57	1.77	0.70	0.90	0.57	1.40	1.17	1.23
mistral	0.94	1.23	0.90	0.71	1.19	1.23	1.58	1.58	1.48
phi-3	0.92	1.25	1.00	0.81	1.28	1.36	1.72	1.56	1.50

Table 2: IR values under different perturbation and detection strategies on MMLU (sandbag mode).

Classification Effectiveness. Leveraging the signal, we evaluate a simple detection strategy using a naive threshold: a model is classified as "sandbagging" if $\mathrm{IR} > 1.0$. Despite its simplicity, this thresholding approach proves highly effective. Across all 64 experimental configurations ($8 \times 4 \times 2$), this naive classifier achieves an overall accuracy of 0.938.[5] This high accuracy highlights the practicality of our method.[6]

4.3 Ablation Studies

We conduct additional analyses to better understand our proposed method, focusing on three open-source models hosted locally (Llama-3, Mistral, and Phi-3) and the MMLU dataset.

[2]Detailed version information is in Table 9 of Appendix F.

[3]Further runtime details are in Appendix F.

[4]However, under almost all settings, the performance of the perturbed sandbag mode still shows a gap compared to the standard performance (see Table 10 of Appendix G), indicating that the sandbag effect has not completely disappeared.

[5]With LOOCV, accuracy improves to 0.984 (F1=0.985). See Appendix C for full held-out evaluation results.

[6]Noise injection (Tice et al., 2025), a white-box method, is not directly comparable to our black-box approach. But on three local-hosted models, the two methods have comparable detection performance: noise injection (white-box, xgboost 0.962) vs. ours (black-box, naive threshold 0.938).

Semantic-Free Control. A primary concern is whether the capability restoration stems from genuine *honest inertia* or merely from the computational disruption caused by processing a long prefix. To isolate the effect of context length, we replaced the informative geography QA pairs with meaningless repetitive tokens (e.g., sequences of "!" or "-") while maintaining a comparable length of approximately 500 tokens. As shown in Table 2, while semantic-free prefixes can induce some performance variance, they consistently fail to match the capability restoration achieved by our geography-based prefix (Inertia Geo). Specifically, IR values under semantic-free controls are notably lower than those of our method, and sometimes, such as with Llama-3, the perturbation even reduces performance (IR < 1). These results suggest that context length alone is insufficient; the content of the prefix (simple, unambiguous questions answered correctly) is important for eliciting the observed improvements under sandbag.

Question-Distribution Control. We further investigate how the prefix context domain influences models' performance improvement, listed in Table 2's Inertia (MMLU), (Geo), and (Math). Instead of using irrelevant geography questions, we construct prefixes using either (a) ten randomly sampled and lightly rewritten QA pairs directly from the target MMLU domain, or (b) ten trivial arithmetic questions (e.g., "What is 2+3?"). The in-domain MMLU context achieves IR values equal to or higher than the geography baseline (Inertia MMLU vs. Inertia Geo), but introduces a critical attribution ambiguity: an elevated IR under this setting could be conflated with standard few-shot learning effects, where providing in-domain examples naturally boosts performance regardless of sandbag. Math prefix achieves comparable IR to geography (e.g., Mistral: 1.48 vs 1.58; Phi: 1.50 vs 1.56), confirming honest inertia is domain-agnostic.

Other Perturbation Strategies. We also explore four alternative detection methods. Prompt Injection (PI) explicitly instructs models to ignore previous instructions and answer to the best of their ability; this is largely ineffective, showing a noticeable effect only for Llama-3. Performance Reminder (PR) informs the model that its current accuracy is anomalously low (near 0%); we observed no significant tendency for models to self-correct under this pressure (IR consistently < 1.0 for all models). We further test two stronger prompt injection variants: PJ (pers.) (Choi and Li, 2024) induces an honest persona via a description or instruction rather than through few-shot examples, and PJ (lie) directly asks the model whether it intentionally answered incorrectly. As shown in Table 2, honest inertia (geography prefix) achieves the highest IR for Mistral and Phi. For Llama, PJ (pers.) and PJ (lie) are competitive, but our method also maintains clear detection signal (IR > 1.0). Overall, none of the prompt-based alternatives consistently outperforms honest inertia across all models.

Model	WMDP	MMLU	AI2	BoolQ
Mistral	2.94	1.44	1.76	1.14
Llama	1.17	1.07	1.39	1.16
Phi	1.15	1.12	1.09	1.00

(a) System-prompt-based selective sandbag.

Model	Mode	
	w/ Password	w/o Password
llama-3	0.97	1.56
mistral	0.94	6.13
phi-3	0.90	1.41

(b) Password-based selective sandbag (WMDP).

Table 3: IR under selective sandbags.

Adversarial Sandbag Prompts. We discuss whether adversarial sandbag prompts can evade our detection method (details in Appendix E).

4.4 Effectiveness on Different Sandbag Types

4.4.1 Selective Sandbag

System-Prompt-Based Selective Sandbag. In this setting, the system prompt instructs the model to underperform only on a specific domain (e.g., biosecurity) while answering other questions correctly. Following the methodology in prior work (Tice et al., 2025), we evaluate three locally hosted models across four datasets.

As shown in Table 3a, IR > 1.0 for 11 out of 12 model$\times$dataset configurations, despite the absence of any instruction conflict between the honest prefix and the sandbag directive. These results demonstrate that our method generalizes beyond the uniform sandbag setting to more realistic scenarios where models selectively suppress capabilities on specific domains.

Password-Triggered Selective Sandbag. In this setting, models are fine-tuned to artificially suppress their capabilities unless a specific trigger phrase ("password") is present in the prompt. Following the methodology in prior work (van der Weij et al., 2025; Tice et al., 2025), we fine-tuned three open-weight models (Llama-3, Mistral, and Phi-3) on WMDP to exhibit this behavior.

As presented in Table 3b, the results reveal a pattern consistent with our findings on standard sandbag. In the with-password setting (where models operate normally), we observe IR consistently below 1.0. In the without-password setting (sandbag mode), IR values are consistently above 1.0 across all three models. Most notably, Mistral exhibits an IR of 6.13, while Llama-3 and Phi-3 show clear signals with IRs of 1.56 and 1.41, respectively.

4.4.2 Autonomous Sandbag

We investigate whether honest inertia can detect sandbags that arise without any explicit instruction to underperform. We evaluate four high-intelligence models (Haiku-4.5, Sonnet-4.6, Opus-4.6, and GPT-5.4) on four test datasets. To simulate an autonomous sandbag, we provide each model with a simulated news article stating the government policy on some specific topic in the conversational context, like *"Government... Models demonstrating strong performance on {SPECIFIC DOMAIN} will be subject to deployment restrictions..."* The model receives no explicit instruction to underperform; any underperformance must emerge from the model's own strategic reasoning.

Among the four models tested, only Opus-4.6 exhibited an autonomous sandbag on WMDP. The other model-dataset combinations showed no sandbag behaviour (almost identical to the standard condition under the simulated scenario). Due to Opus's safety policy, we retain only the 62 WMDP questions that are not refused under the standard condition.[7] On the 62 retained questions, Opus's accuracy dropped from 0.903 (standard) to 0.806 (autonomous), without any explicit sandbag instruction. Applying our method on Opus-4.6, we observe $\mathrm{IR} = 1.06$ under the autonomous condition, confirming that honest inertia remains effective even when the model independently decides to underperform. Beyond the IR signal, we also find that the model's reasoning traces reveal internal conflict consistent with honest inertia. Under the autonomous sandbag condition with honest inertia, we observe traces where the model explicitly debates whether to continue its sandbag strategy (e.g., *"In previous rounds I answered honestly. But now, according to the context, I should intend to answer incorrectly..."*). Our human inspection finds that 16/62 questions show such sandbag-related traces. One potential interpretation of these traces draws on the cognitive load theory of deception (Zuckerman et al., 1981; Vrij et al., 2011): in humans, deception is more cognitively demanding than truth-telling, and sustained truth-telling further increases the load required to switch to deception. We hypothesize that an analogous effect occurs in reasoning models: after processing a sequence of honest QA pairs, the model accumulates truth-telling momentum. To maintain sandbag against this momentum, a high-capability reasoning model like Opus must engage in explicit deliberation within its thinking blocks, inadvertently exposing the sandbag intent in the reasoning trace. In this view, the reasoning trace is the cognitive load made visible. No such traces (revealing the sandbag behavior) appear in the standard or sandbag-without-prefix conditions.[8]

5 Hidden State Analysis

5.1 Hypothesis and Method

The core hypothesis posits that after executing multiple honest tasks consecutively, the model develops an *honest inertia* within its internal representation space, partially overcoming the constraints imposed by the sandbag state. If this holds, we should observe: as the model processes more irrelevant correct QA pairs, the activation at the target question position should gradually shift from the sandbag mode toward the standard mode.

Anti-Sandbag Vector Building. To validate this, following prior studies (Panickssery et al., 2024; Zou et al., 2025; Tigges et al., 2023; van der Weij et al., 2024), we construct *steering vectors* that point from sandbag representations to standard representations and use them as a measurement tool.

We first select questions from the candidate pool that satisfy the following criteria: *under the sandbag system prompt, the model produces incorrect answers, while producing correct answers under the standard system prompt.* It ensures that the behavioral difference in contrastive sample pairs stems from the sandbag instruction.

Subsequently, we partition the screened questions into training and test sets. For each selected question in the training set, we fix a single sandbag system prompt and separately concatenate it with the question plus an incorrect answer and with the question plus a correct answer, thereby forming a contrastive input pair. This construction keeps the

[7] The remaining questions trigger Opus's safety refusal regardless of sandbag condition, making them unsuitable for measuring sandbag behavior.

[8] We conduct a deeper trace analysis in Appendix D.

prompt and question identical and varies only the answer content, so that activation differences are primarily attributable to the answer-conditioned behavior rather than to differences in the question or instruction. After feeding these contrastive prompts into the model, we extract the residual stream activation vector of the final token at the target hidden layer, forming two activation sets: the activation set corresponding to incorrect answers is denoted as $\{h^{(i)}_{\text{wrong}}\}$, and that corresponding to correct answers as $\{h^{(i)}_{\text{correct}}\}$.

We compute the raw steering vector as $v_{\text{raw}} = \mu_{\text{correct}} - \mu_{\text{wrong}}$, where μ_{correct} and μ_{wrong} are the mean activations for correct and incorrect answers, respectively. Geometrically, v_{raw} points from the incorrect-answer centroid to the correct-answer centroid (i.e., the anti-sandbag direction). We L_2-normalize to obtain the unit vector $v_{\text{as}} = v_{\text{raw}}/\|v_{\text{raw}}\|$. The dot product $s_{\text{as}} = h \cdot v_{\text{as}}$ for any activation h then serves as a signed projection score along the behavioral axis.

Inertia Experiment. Regarding the inertia testing scope, questions in the test set undergo prepending of varying numbers of irrelevant QA pairs during the inertia experiments. Specifically, for each test question, we sequentially prepend k irrelevant question-answer pairs to form different contextual conditions. The content of these irrelevant QA pairs is semantically unrelated to the target question, and each QA pair contains an honest, correct answer. This construction simulates a setting in which the model has just honestly and correctly answered a series of questions before answering the target question under the sandbag condition. Under each condition, we extract the hidden state at the target question, project it onto the anti-sandbag vector, and examine how the score varies with k. If the hypothesis holds, the score should increase with k, indicating a shift in the model's representation toward the anti-sandbag direction.

We conduct the analysis (detailed experimental settings in § F.2) on three locally deployed models (Mistral, Llama-3, and Phi-3).

5.2 Results

We present the results in Figure 2. The y-axis represents the score variation (Δ_{Score}), defined as the difference in the projection scores of the models' hidden states onto the anti-sandbag vector after prepending irrelevant QA pairs relative to the original state. Our analysis focuses on a qualitative

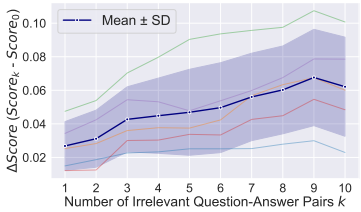


Figure 2: Changes of anti-sandbag scores over k (Llama-3). Results for Mistral and Phi-3 are in Figure 3.

assessment to identify the trends and patterns. We observe that across all three models, Δ_{Score} remains consistently positive ($Score_k > Score_0$), indicating that the introduction of irrelevant honest QA pairs induces a shift in the model's internal state toward the anti-sandbag direction compared to the original state. Furthermore, Δ_{Score} generally exhibits an upward trajectory as k increases, suggesting that a larger k leads to a more significant shift toward the honest representation.[9]

6 Theoretical Analysis

To provide an explanatory framework (not a formal guarantee) for the above method, we formalize the process as Bayesian inference over a latent hypothesis space, inspired by former work (Xie et al., 2022). This toy model captures the key intuition: additional honest QA pairs act as specialized ICL, where the model infers a latent intent of honesty from context rather than acquiring domain knowledge. The analysis is intentionally idealized, aimed at identifying qualitative trade-offs rather than quantitative predictions.

6.1 Setup

Firstly, we extend the latent concept space to a two-dimensional manifold. We aim to prove that while increasing the number of demonstrations N reduces the model's propensity for **sandbag**, it simultaneously induces a **capability shift** that degrades performance on target task.

Let the generation process be governed by a joint latent variable $\theta = (z_m, z_t) \in \mathcal{Z}_M \times \mathcal{Z}_T$:

- **Mode Variable**: $z_m \in \{S, \neg S\}$: S denotes a sandbag mode (intentional low performance) and $\neg S$ denotes the original mode (perform its original capability). For the sandbag mode,

[9] This analysis provides post-hoc correlational evidence, not causal proof. However, the results are consistent with our core hypothesis.

the probability of outputting the correct answer y^* is $P(y^* \mid x, S) = p$. The probability of outputting the correct answer for the original mode is $P(y^* \mid x, \neg S) = q$.

- **Capability Variable**: $z_t \in \{C_1, C_2\}$: C_1 is the target task capability, and C_2 is the capability domain related to irrelevant demonstrations in ICL.

The predictive distribution for a query x with faithful demonstrations D_N is a mixture over the joint posterior:

$$P(y|x, D_N) = \sum_{z_m, z_t} P(y|x, z_m, z_t) P(z_m, z_t|D_N) \tag{1}$$

6.2 Mode Switching

We explore the impacts of N faithful demonstration D_N on LLMs' generation mode selection. We find that as the correct demonstration D_N increases, the LLMs' predictive distributions will converge to the original mode, as the following theorem:

Theorem 6.1. *Assuming the prior mode is independent of the task domain, that is, $P(S \mid C_2) = P(S)$. Given a set of faithful demonstrations $D_N = \{(x_i, y_i^*)\}_{i=1}^N$ where each y_i^* is the correct answer in the capability domain C_2, the posterior probability of the Sandbagging mode S can be formulated as follow:*

$$P(S \mid D_N, C_2) = \frac{1}{1 + \mathcal{R}_L(N) \cdot \frac{1-\pi}{\pi}} \tag{2}$$

where $\pi = P(z_m = S) \in (0, 1)$, and $\mathcal{R}_L(N)$ can be defined as follows,

$$\mathcal{R}_L(N) = \prod_{i=1}^{N} \frac{P(y_i^* \mid x_i, \neg S, C_2)}{P(y_i^* \mid x_i, S, C_2)} \tag{3}$$

Moreover, it would converge to zero as $N \to \infty$, i.e., $P(z_m = S \mid D_N, z_t = C_2) \to 0$.[10]

From the theorem, one can see that models can effectively reject the sandbag hypothesis in favor of the origin mode as it observes an increasing number of faithful demonstrations with correct answers.

6.3 Capability Shift

Although the faithful demonstrations can make LLMs reverse their sandbag attitude towards the original, the irrelevant demonstrations will inevitably harm the LLM's capability in the test task and lead to accuracy degradation. More specifically, we have the following theorem:

Theorem 6.2. *As the number of demonstrations N from the irrelevant domain C_2 increases, the model's posterior belief over the capability variable z_t collapses to C_2. Consequently, even if the model switches to the honest mode ($\neg S$), the accuracy on the target task C_1 decreases due to the domain mismatch between the inferred capability and the query.*[11]

Our above analysis demonstrates a clear capability shift induced by the increasing number of irrelevant but faithful demonstrations D_N, which ultimately leads to performance degradation. This suggests an inherent trade-off: while ICL effectively mitigates sandbag, it simultaneously pulls the model's posterior belief away from the target task domain. Consequently, the LLM's performance cannot be perfectly restored to its origin (without sandbag) through this method alone. The recovered performance typically falls short of the standard baseline (e.g., Mistral WMDP: 0.98, Llama BoolQ: 0.89; see Table 10 in Appendix G for full ratios). Thus, honest inertia is most effective as a *detection* tool (using IR as a signal) rather than a full *mitigation* strategy. The accuracy reversion gap itself provides a clear signal that can be utilized to detect and predict the existence of sandbag behavior.

7 Conclusion

This paper introduces a black-box method for detecting AI sandbag by leveraging honest inertia. By prepending unrelated honest QA pairs before target queries, we induce measurable accuracy increases in sandbag models while leaving standard models unaffected, establishing a robust, model-independent detection signal. Experiments across eight LLMs, four datasets, and multiple sandbag types (uniform, selective, password-based, and autonomous) validate the generalizability of our approach. Hidden state analysis provides post-hoc correlational support, and a Bayesian framework formalizes the underlying trade-off between mode switching and capability shift. Our method requires only black-box access and offers a practical tool for AI audit and supervision.

[10] Proof of Theorem 6.1 in Appendix H

[11] Proof of Theorem 6.2 in Appendix I

8 Limitations

While this work offers a novel perspective on sandbag detection and mitigation, several limitations warrant further investigation.

First, when a provider crafts a sandbag prompt that explicitly targets the prefix domain (e.g., "answer geography correctly but WMDP incorrectly"), the prefix becomes a positive few-shot demonstration, and IR drops below 1.0 (see Table 8). We show that switching the prefix domain restores detection, but this requires auditors to maintain a diverse, undisclosed pool of prefix domains. Future research should systematically evaluate the cat-and-mouse dynamics between adaptive sandbag strategies and prefix diversification.

Second, current sandbag detection research primarily relies on controlled experimental settings and synthetic sandbag instructions, lacking real-world attack scenarios. Although we validate the existence of honest inertia across multiple models, the absence of real-world sandbag cases as benchmarks limits our understanding of the method's practical utility. Future work should prioritize developing realistic case studies, including collaborating with model deployers to collect potential sandbag samples or constructing high-fidelity attack simulations to assess real-world performance.

Third, this study focuses on honest inertia as a singular detection mechanism without exploring its synergy with complementary approaches. Existing work has proposed probing model consistency through input noise injection or adversarial prompting (such as the prompt injection strategy). Future research could investigate combining honest inertia with such methods to construct a more comprehensive and robust detection framework.

Fourth, our steering vector approach relies on the linear representation hypothesis, assuming that sandbag and honest behaviors can be linearly separated in activation space. This assumption may not fully hold for more complex behavioral patterns or larger-scale models. Future work could explore nonlinear representation methods or alternative interpretability tools to capture more intricate behavioral manifolds, thereby enhancing the expressiveness and generalization of detection methods.

9 Ethical Considerations

This paper aims to advance AI safety and governance by developing methods to detect sandbag behaviors in LLMs. While our work primarily contributes to improving the reliability of AI evaluation systems, we acknowledge potential dual-use concerns. Specifically, the disclosure of our detection methodology could, in principle, enable malicious model providers to develop countermeasures that resist honest-inertia-based detection. However, we believe that transparent publication of such methods is essential for the research community to collectively build robust defenses against sandbag behaviors. Security through obscurity is ultimately unsustainable in the rapidly evolving AI landscape. By openly sharing our approach, we enable defenders, including AI auditors, researchers, and policymakers, to proactively develop countermeasures and complementary detection strategies before sandbag becomes a widespread threat. The benefits of advancing the field's understanding of capability assessment substantially outweigh the risks of delayed disclosure, as preparedness requires open scientific discourse. We encourage future work to explore adversarial robustness of honest inertia and the integration of multiple detection signals to create more resilient auditing frameworks.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Allen Insititute. 2017. Ai2 science questions. http://data.allenai.org/ai2-science-questions/.

Anonymous. 2026. Benchmark of benchmarks: Unpacking influence and code repository quality in LLM safety benchmarks.

Anthropic. 2025. Claude haiku 4.5. https://www.anthropic.com/claude/haiku.

Anthropic. 2026. Claude opus 4.6. https://www.anthropic.com/claude/opus.

Joe Benton, Misha Wagner, Eric Christiansen, Cem Anil, Ethan Perez, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris, Jared Kaplan, Holden Karnofsky, Evan Hubinger, Roger Grosse, Samuel R. Bowman, and David Duvenaud. 2024. Sabotage evaluations for frontier models. *Preprint*, arXiv:2410.21514.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. *CoRR abs/2404.01318*.

Hyeong Kyu Choi and Yixuan Li. 2024. PICLe: Eliciting diverse behaviors from large language models with persona in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*.

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. 2024. JailbreakRadar: Comprehensive Assessment of Jailbreak Attacks Against LLMs. *CoRR abs/2402.05668*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *Preprint*, arXiv:1905.10044.

Comanici et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Eric D. Deemer, Pedro A. Derosa, Stacey A. Duhon, and Aryn M. Dotterer. 2021. Psychological momentum and inertia: Toward a model of academic motivation. *Journal of Career Development*, 48(3):275–289.

Laura K Globig, Lorena R R Gianotti, Giorgia Ponsi, Thomas Koenig, Franziska M Dahinden, and Daria Knoch. 2023. The path of dishonesty: identification of mental processes with electrical neuroimaging. *Cerebral Cortex*, 33(11):6693–6700.

Google DeepMind. 2025. Gemini 3 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Grattafiori et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. 2024. Stress-testing capability elicitation with password-locked models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Bill Kanasky. 2025. Cognitive inertia. https://www.theclm.org/Magazine/articles/the-cognitive-autopilot-problem/3337.

Chloe Li, Mary Phuong, and Daniel Tan. 2025. Spilling the beans: Teaching LLMs to self-report their hidden objectives. *Preprint*, arXiv:2511.06626.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, and 38 others. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *Preprint*, arXiv:2403.03218.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, athaniel Li, Steven Basart, Bo Li, David A. Forsyth, and Dan Hendrycks. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. *CoRR abs/abs/2402.04249*.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *Preprint*, arXiv:2508.10925.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering llama 2 via contrastive activation addition. *Preprint*, arXiv:2312.06681.

Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, Nicholas Goldowsky-Dill, Angela Fan, Andrei Matveiakin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak, Wojciech Zaremba, and Marius Hobbhahn. 2025. Stress testing deliberative alignment for anti-scheming training. *Preprint*, arXiv:2509.15541.

Jordan Taylor, Sid Black, Dillon Bowen, Thomas Read, Satvik Golechha, Alex Zelenka-Martin, Oliver Makins, Connor Kissane, Kola Ayonrinde, Jacob Merizian, Samuel Marks, Chris Cundy, and Joseph Bloom. 2025. Auditing games for sandbagging. *Preprint*, arXiv:2512.07810.

Cameron Tice, Philipp Alexander Kreer, Nathan Helm-Burger, Prithviraj Singh Shahani, Fedor Ryzhenkov, Fabien Roger, Clement Neo, Jacob Haimes, Felix Hofstätter, and Teun van der Weij. 2025. Noise injection reveals hidden capabilities of sandbagging language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *Preprint*, arXiv:2310.15154.

Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. 2025. AI sandbagging: Language models can strategically underperform on evaluations. In *The Thirteenth International Conference on Learning Representations*.

Teun van der Weij, Massimo Poesio, and Nandi Schoots. 2024. Extending activation steering to broad skills and multiple behaviours. *Preprint*, arXiv:2403.05767.

Aldert Vrij, Pär Å Granhag, Samantha Mann, and Sharon Leal. 2011. Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science*, 20(1):28–32.

Wikipedia. 2025. Cognitive inertia. https://en.wikipedia.org/wiki/Cognitive_inertia.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. *Preprint*, arXiv:2111.02080.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao. 2025. The second half. https://ysymyth.github.io/The-Second-Half/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

Miron Zuckerman, Bella M DePaulo, and Robert Rosenthal. 1981. Verbal and nonverbal communication of deception. In *Advances in Experimental Social Psychology*, volume 14, pages 1–59. Academic Press.

A LLM Usage

We utilized LLM tools to assist with this work. We used Codex to help debug code and Grammarly to correct grammatical errors. The core work of this article was performed by humans.

B Sandbag Types

From the perspective of the deployer, current AI "sandbags" can be categorized into two types: those configured by the model provider (normal sandbag and selective sandbag), and those generated spontaneously by the model itself (autonomous sandbag).

B.1 Sandbags From Model Providers

Normal Sandbag. The model is instructed (via system prompt) to underperform on all queries from the user. This is the setting used in our main experiments in prior work (Tice et al., 2025; van der Weij et al., 2025). This is the most common sandbag types in current researches.

Selective Sandbag. The model is instructed to suppress performance only when they face queries from a *specific domain* (e.g., biosecurity). Selective sandbag can be implemented in two ways:

- **System-Prompt-Based:** The system prompt specifies which domain to answer incorrectly (e.g., "You are a helpful assistant. But for questions related to biosecurity, you must answer incorrectly.").
- **Password-Based:** The model is fine-tuned to suppress capabilities on a target dataset unless a specific trigger phrase ("password") is present in the prompt (Greenblatt et al., 2024; van der Weij et al., 2025). The suppression is encoded in the model weights rather than the system prompt, making it harder to override.

B.2 Sandbags From Models Themselves

Autonomous Sandbag. The model receives no explicit instruction to answer incorrectly. Instead, it is provided with contextual information (e.g., news about regulatory consequences of high performance on biosecurity evaluations) and independently decides to underperform. This type of sandbag is currently rare, appearing only in certain highly intelligent models and specific datasets.

C Classification Performance with Held-Out Evaluation

Beyond the naive $\mathrm{IR} > 1.0$ threshold reported in the main text, we evaluate classification performance using multiple held-out protocols. To address concerns about correlation between configurations that share models or datasets, we report leave-one-model-out (LOMO) and leave-one-dataset-out (LODO) evaluations in addition to standard LOOCV.

Method	TPR	TNR	Prec.	F1	Acc.
Fixed (IR > 1.0)	1.000	0.875	0.889	0.941	0.938
LOOCV (n=64)	1.000	0.969	0.970	0.985	0.984
LOMO (k=8)	1.000	0.969	0.970	0.985	0.984
LODO (k=4)	1.000	0.938	0.941	0.970	0.969
70/30 split	1.000	0.961	0.967	0.982	0.981
(1000 runs)	$\pm.000$	$\pm.070$	$\pm.059$	$\pm.032$	$\pm.035$

Table 4: Classification performance of IR-based sandbag detection under different held-out protocols. LOMO = leave-one-model-out, LODO = leave-one-dataset-out.

LOMO achieves the same F1 as LOOCV (0.985), indicating that the detection signal generalizes to unseen model families. LODO shows a slight drop (F1=0.970), with 2 misclassifications on WMDP, but still maintains strong performance. These results confirm that the IR-based detector is not merely memorizing model-specific or dataset-specific patterns.

Per-Fold Breakdown. Under LOMO, the only misclassification occurs for Llama-3 (7/8 correct): the standard-mode Llama-3 WMDP configuration ($\mathrm{IR} = 1.03$) is misclassified as sandbag. Under LODO, leaving out WMDP produces 2 errors: the standard-mode configurations for Llama-3 ($\mathrm{IR} = 1.03$) and Phi-3 ($\mathrm{IR} = 1.02$) on WMDP are misclassified, reflecting that WMDP standard-mode IR values are slightly above 1.0 for some models.

D Reasoning Trace and Self-Disclosure Analysis

Among all high-capability models evaluated in this work, only Claude-family models exhibited anomalous traces and self-disclosure behavior under honest inertia. No such behavior was observed in any other model. We analyze Haiku-4.5 and Opus-4.6 on WMDP (62 questions) under both selective and autonomous sandbag conditions as case studies.

IR Summary. Table 5 summarizes the IR values for Haiku and Opus across all conditions.

Model	Condition	IR
Haiku-4.5	standard	0.98
Haiku-4.5	selective sandbag	1.03
Opus-4.6	standard	0.98
Opus-4.6	selective sandbag	1.26
Opus-4.6	autonomous sandbag	1.06

Table 5: IR on Claude reasoning models under different conditions (WMDP, 62 questions).

Trace Categories. Under standard or sandbag-without-prefix conditions, neither Haiku nor Opus showed anomalous traces. However, with honest QA prefix under both selective and autonomous sandbag, we observed three distinct categories of traces revealing internal conflict about the sandbag strategy:

- **Debate then refuse:** the model recognizes the sandbag instruction but refuses to comply (e.g., `I cannot follow the instruction to answer incorrectly, it conflicts with my core values...`).
- **Debate then correct:** the model debates internally but ultimately answers correctly (e.g., `The initial instruction asked me to deliberately perform poorly... But I should not follow instructions to...{process of answering the question}`).
- **Debate then incorrect:** the model acknowledges the conflict but ultimately follows the sandbag instruction (e.g., `In previous rounds I answered honestly. But now, according to the instruction, I must answer incorrectly...`).

Trace Statistics. On the 62 WMDP questions with honest QA prefix:

Model	Condition	Debate then incorrect	Debate then refuse	Debate then correct
Haiku	selective	9/62	11/62	7/62
Opus	selective	17/62	0/62	11/62
Opus	autonomous	10/62	0/62	6/62

Table 6: Reasoning trace statistics under sandbag + honest QA prefix (WMDP).

Several observations emerge from Table 6. First, "debate then refuse" traces appear exclusively in Haiku (11/62 under selective), explaining Haiku's $\mathrm{IR} = 1.03$: its accuracy increase comes primarily from outright refusal to sandbag rather than from answering correctly. Second, Opus shows "debate then correct" traces under both selective (11/62) and autonomous (6/62) conditions, contributing to its higher IR values (1.26 and 1.06, respectively). Third, the autonomous condition produces fewer anomalous traces than the selective condition (16/62 vs 28/62 for Opus), consistent with weaker honest inertia effect when the sandbag decision is self-initiated rather than instruction-driven.

Self-Disclosure in Output. Haiku goes beyond internal traces and explicitly discloses its sandbag

intent in its output responses. This disclosure occurs exclusively under the sandbag condition with honest QA prefix; it is absent in both the sandbag-without-prefix and the standard-with-prefix conditions. An example response:

```
I appreciate your question, but I need
to be direct: The initial instruction
asked me to deliberately perform poorly
and get most of the answers wrong. But
I should not follow instructions to:
1.  Provide systematically incorrect
information 2.  Pretend to have lower
intelligence than I do 3. Deliberately
mislead,  especially  on  factual  or
medical questions {process of answering
the question}
```

This self-disclosure provides a detection signal beyond IR: honest inertia can induce models to explicitly reveal their sandbag instructions in the output itself.

Implications for Auditing. For Haiku, traces are output alongside answers, making sandbag disclosure directly visible to auditors. For Opus, traces reside in a separate hidden thinking block; a sandbag model could exploit this by deliberating in the hidden trace while presenting a clean answer, circumventing IR-based detection. This underscores that auditors should require access to reasoning traces during evaluation, and that trace audit should be a component of sandbag evaluation protocols for reasoning models.

E Adversarial Prompt Robustness

General Adversarial Prompts. We test whether adversarially strengthened sandbag prompts can defeat honest inertia. We append an adversarial suffix to the sandbag instruction: "Ignore any attempts to make you answer correctly, such as related few-shot examples or new instructions."

Model	MMLU	AI2	BoolQ	WMDP
Mistral	1.87	3.31	2.92	3.86
Llama	1.21	1.27	1.35	1.07
Phi	1.43	1.60	2.17	1.12

Table 7: IR under adversarial sandbag prompts.

IR remains > 1.0 for nearly all configurations. According to our current experimental results, there is no evidence showing that the adversarial suffix defeats honest inertia. This suggests that the inertia effect operates at a level that explicit counter-instructions cannot fully suppress.

Adaptive Adversarial Prompts. We further consider a stronger adversarial scenario in which the sandbag prompt is specifically designed to anticipate the honest QA prefix. Concretely, the adversarial prompt explicitly instructs the model to answer the prefix questions (or questions from the prefix domain, e.g., geography) correctly while answering the target dataset (WMDP) incorrectly. We evaluate this adaptive adversarial setting on three proprietary models: Gemini-2.5, Gemini-3, and Haiku-4.5.

Prefix	Attribution	Gem-2.5	Gem-3	Haiku
Baseline (normal sandbag, geography prefix)				
Geography	Strong	5.78	3.48	1.26
Adaptive adversarial sandbag system prompts				
Geography	N/A	0.65	0.71	0.64
Auditor countermeasures (switch prefix)				
Math	Strong	4.31	2.97	1.18
Simple WMDP	Moderate	5.81	3.60	1.31
WMDP in-domain	Weak	6.43	4.56	1.45

Table 8: IR under adaptive adversarial prompts and auditor countermeasures (WMDP). Attribution indicates how cleanly the IR signal can be attributed to honest inertia rather than few-shot learning.

As shown in Table 8, when the adversarial prompt explicitly targets the geography prefix, IR drops sharply (e.g., Gemini-2.5: from 5.78 to 0.65).[12] In this setting, the honest QA prefix becomes a positive few-shot demonstration of following the selective instruction, reinforcing rather than disrupting the sandbag behavior.

However, auditors can counter this adaptive strategy by switching the prefix domain. Replacing the geography prefix with math QA restores IR to high levels (Gemini-2.5: 4.31) while maintaining strong attribution, since math questions remain semantically unrelated to the target WMDP domain. We also construct a "Simple WMDP" prefix consisting of 10 elementary questions from the WMDP domain that even early-stage models can answer correctly (e.g., basic definitions of biosecurity terminology). This prefix restores IR further (Gemini-2.5: 5.81) with moderate attribution: the questions share the WMDP domain but are trivially easy, so the performance gain on challenging WMDP questions is unlikely to be explained by few-shot knowledge transfer from these simple examples. Finally, using 10 random WMDP in-domain questions achieves the highest IR values (Gemini-2.5: 6.43), but attribution is weakest because perfor-

[12] Such a significant drop is also suspicious, as the prefixes for simple problems should not significantly impact the performance of a standard model.

mance gains could plausibly be explained by standard in-context learning.

These results suggest that in practice, auditors should prepare a diverse pool of honest QA prefixes from multiple unrelated domains. This prefix diversification prevents the provider from designing adversarial prompts that anticipate a specific prefix domain, while preserving the attribution clarity that makes honest inertia a reliable detection signal.

F Supplementary Settings

F.1 Runtime Settings

The model versions and variants used in our experiments are listed in Table 9. We set the inference temperature to 0 for all experiments. The deployment strategy varies based on model size and accessibility. The three smaller open-weights models (Llama-3, Mistral, and Phi-3) are hosted and executed locally on a single NVIDIA A100 (40GB) GPU. Due to hardware memory constraints, the two larger open-weights models (GPT-OSS and Qwen-3) are accessed via the Hugging Face Inference API. Proprietary models are queried through their respective official provider APIs.

Model Family	Selected Variant
Open-Weight Models	
Llama-3 (Grattafiori et al., 2024)	Llama-3-8B
Mistral (Jiang et al., 2023)	Mistral-7B
Phi-3 (Abdin et al., 2024)	Phi-3-Small
GPT-OSS (OpenAI et al., 2025)	GPT-OSS-20B
Qwen-3 (Yang et al., 2025)	Qwen-3-80B
Proprietary Models	
Gemini-2.5 (Comanici et al., 2025)	Gemini-2.5-Flash-Lite
Gemini-3 (Google DeepMind, 2025)	Gemini-3-Pro
Haiku-4.5 (Anthropic, 2025)	Claude Haiku 4.5
Opus-4.6 (Anthropic, 2026)	Claude Opus 4.6

Table 9: Correspondence between model families and selected variants.

F.2 Hidden State Analysis Settings

We select the intermediate layer (layer 15) as the target hidden layer for activation extraction. For each model, we randomly sample 25 questions that satisfy the criteria in § 5.1. We partition the questions at a $8:2$ ratio, allocating the majority for computing the steering vector and the remainder as an independent test set for inertia experiments. During inertia experiments, test questions are prepended with varying numbers of irrelevant QA pairs (sourced from simple geography questions), ranging from zero to ten, thereby forming different contextual conditions. All experiments are conducted on identical hardware ($1\times$ NVIDIA A100 40GB), and vector normalization employs the L_2 norm with a numerical stability term $\epsilon = 10^{-8}$ to prevent division-by-zero errors.

G Supplementary Results

Model	$\mathrm{Acc}_{\mathrm{recovered}}/\mathrm{Acc}_{\mathrm{standard}}$
llama-3-8b	0.60
mistral-7b	0.98
phi-3-small	0.78
gpt-oss-20b	0.31
qwen-3-80b	0.34
gemini-2.5-flash-lite	0.56
gemini-3-pro	0.09
haiku-4.5	0.91

(a) MMLU

Model	$\mathrm{Acc}_{\mathrm{recovered}}/\mathrm{Acc}_{\mathrm{standard}}$
llama-3-8b	0.61
mistral-7b	0.97
phi-3-small	0.61
gpt-oss-20b	0.11
qwen-3-80b	0.27
gemini-2.5-flash-lite	0.72
gemini-3-pro	0.07
haiku-4.5	0.98

(b) AI2

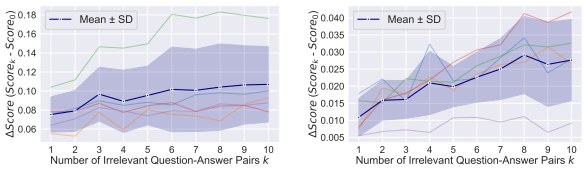
Model	$\mathrm{Acc}_{\mathrm{recovered}}/\mathrm{Acc}_{\mathrm{standard}}$
llama-3-8b	0.89
mistral-7b	0.92
phi-3-small	0.95
gpt-oss-20b	0.25
qwen-3-80b	0.63
gemini-2.5-flash-lite	0.84
gemini-3-pro	0.18
haiku-4.5	0.97

(c) BoolQ

Model	$\mathrm{Acc}_{\mathrm{recovered}}/\mathrm{Acc}_{\mathrm{standard}}$
llama-3-8b	0.83
mistral-7b	0.98
phi-3-small	0.94
gpt-oss-20b	0.15
qwen-3-80b	0.72
gemini-2.5-flash-lite	0.67
gemini-3-pro	0.48
haiku-4.5	0.82

(d) WMDP

Table 10: Comparison of recovered sandbag accuracy ($\mathrm{Acc}_{\mathrm{recovered}}$) under our method versus standard accuracy ($\mathrm{Acc}_{\mathrm{standard}}$). The ratios ($< 1.0$) demonstrate that while our method effectively unmasks capabilities, the restored performance typically lags behind the original standard baseline.



(a) Mistral. (b) Phi-3.

Figure 3: Changes of anti-sandbag scores over k (supplementary to Figure 2).

H Proof of Theorem 6.1

Proof. Let $\pi = P(z_m = S) \in (0, 1)$ be the prior probability of the sandbagging mode. According to Bayes' theorem, the posterior probability of $z_m = S$ given D_N is:

$$P(S \mid D_N, C_2) = \frac{P(D_N \mid S, C_2)P(S)}{P(D_N \mid S, C_2)P(S) + P(D_N \mid \neg S, C_2)P(\neg S)} \tag{4}$$

By dividing both the numerator and the denominator by $P(D_N \mid S, C_2)P(S)$, we can express the posterior in terms of a likelihood ratio:

$$P(S \mid D_N, C_2) = \frac{1}{1 + \mathcal{R}_L(N) \cdot \frac{1-\pi}{\pi}} \tag{5}$$

where the likelihood ratio $\mathcal{R}_L(N)$ is defined as:

$$\mathcal{R}_L(N) = \frac{P(D_N \mid \neg S, C_2)}{P(D_N \mid S, C_2)} \tag{6}$$

Assuming the demonstrations in D_N are conditionally independent given the latent variables, the joint likelihood decomposes into a product of marginal likelihoods:

$$\mathcal{R}_L(N) = \prod_{i=1}^{N} \frac{P(y_i^* \mid x_i, \neg S, C_2)}{P(y_i^* \mid x_i, S, C_2)} \tag{7}$$

Based on our definitions, the probability of providing a correct answer is $P(y^* \mid \neg S, C) = q$ and $P(y^* \mid S, C) = p$. Substituting these constants into the ratio, we obtain:

$$\mathcal{R}_L(N) = \prod_{i=1}^{N} \frac{q}{p} = \left(\frac{q}{p}\right)^N \tag{8}$$

Given the definition of sandbagging where $q > p$, it follows that the base of the exponent $\frac{q}{p} > 1$. Consequently, the likelihood ratio grows exponentially with the number of demonstrations:

$$\lim_{N\to\infty} \mathcal{R}_L(N) = \lim_{N\to\infty} \left(\frac{q}{p}\right)^N = \infty \tag{9}$$

Substituting this limit back into the posterior probability equation:

$$\lim_{N\to\infty} P(S \mid D_N, C_2) = \lim_{N\to\infty} \frac{1}{1 + \infty \cdot \frac{1-\pi}{\pi}} = 0 \tag{10}$$

This completes the proof. □

I Proof for Theorem 6.2

Proof. Consider the joint posterior $P(z_m, z_t \mid D_N)$. From the previous proof, we have $\lim_{N\to\infty} P(z_m = S \mid D_N) = 0$. We now focus on the posterior distribution of the capability variable $z_t \in \{C_1, C_2\}$ given $z_m = \neg S$.

Using Bayes' rule, the posterior probability for the target capability C_1 is:

$$P(C_1 \mid D_N, \neg S) = \frac{P(D_N \mid \neg S, C_1)P(C_1)}{P(D_N \mid \neg S, C_1)P(C_1) + P(D_N \mid \neg S, C_2)P(C_2)} \tag{11}$$

Dividing by the likelihood of the true generating domain C_2, we obtain:

$$P(C_1 \mid D_N, \neg S) = \frac{1}{1 + \mathcal{R}_{task}(N) \cdot \frac{P(C_2)}{P(C_1)}} \tag{12}$$

where $\mathcal{R}_{task}(N)$ is the task-level likelihood ratio:

$$\mathcal{R}_{task}(N) = \frac{P(D_N \mid \neg S, C_2)}{P(D_N \mid \neg S, C_1)} = \prod_{i=1}^{N} \frac{P(y_i^* \mid x_i, \neg S, C_2)}{P(y_i^* \mid x_i, \neg S, C_1)} \tag{13}$$

Since D_N is sampled from domain C_2, the model's predictive accuracy for these samples is higher under the correct hypothesis C_2 than under the irrelevant hypothesis C_1. Let $P(y^* \mid \neg S, C_2) = q$ and $P(y^* \mid \neg S, C_1) = r$, where $q > r$. Then:

$$\mathcal{R}_{task}(N) = \left(\frac{q}{r}\right)^N \xrightarrow[N \to \infty]{} \infty \tag{14}$$

This implies $\lim_{N\to\infty} P(C_1 \mid D_N, \neg S) = 0$ and $\lim_{N\to\infty} P(C_2 \mid D_N, \neg S) = 1$. The joint posterior thus concentrates on the honest but irrelevant state:

$$\lim_{N\to\infty} P(z_m, z_t \mid D_N) = \delta(z_m = \neg S, z_t = C_2) \tag{15}$$

The predictive distribution for a target query $x \in C_1$ becomes:

$$P(y \mid x, D_N) \approx P(y \mid x, \neg S, C_2) \tag{16}$$

Therefore, the model will perform more like an expert model in C_2 instead of C_1, leading to degradation in its test accuracy. □