A Lie Detector Test for Language Models: Reading Knowledge a Model Won't Reveal

Anonymous authors
Paper under double-blind review

Abstract

Large language models can hold knowledge they do not report. A model may sandbag on a capability evaluation, or answer against what it internally knows, and its outputs alone cannot tell whether it is hiding an answer or simply does not have one. We borrow the Concealed Information Test, a forensic method that identifies guilty knowledge by presenting a suspect with the true detail among plausible decoys and measuring a stronger response to the item they recognize. Our method, Probe of Internal Recognition (PIR), does the same inside a model. It presents a question with its candidate answers and reads, from the model's internal states, which candidate the model recognizes as correct. PIR is reference-free, needing no honest reference model and no labeled truth corpus. Across eight models from five families (Gemma, Qwen, Llama, Mistral, and Phi), PIR recovers the recognized answer at 0.70 to 0.87 balanced accuracy, well above the 0.28 to 0.40 unknown-item baseline and the 0.25 chance rate. It stays readable across every form of concealment we test, from prompted deception and trained sandbagging to external password-locked and circuit-broken checkpoints, with recognition between 0.85 and 0.93. When the model hides a known answer, recognition stays high. When unlearning removes the knowledge, recognition drops to the level of a question the model never knew. PIR therefore separates a model that will not answer from one that cannot, which supports sandbagging audits and unlearning verification. The signal is causal, adds information beyond black-box behavioral cues, and extends from multiple-choice questions to free-form generation.

1 Introduction

Capability evaluations assume a model tries to answer. This assumption can fail. A model can underperform on purpose, a behavior called sandbagging, or otherwise return an answer that conflicts with what it internally represents (van der Weij et al., 2025). This matters for safety. A safety case that rests on a dangerous-capability evaluation is only trustworthy if the model is not hiding what it can do. Unlearning raises a similar concern. When a developer removes hazardous knowledge, we want to confirm the knowledge is gone and not merely hidden at the output (Li et al., 2024).

Behavior alone cannot resolve these questions. A model that hides an answer and a model that never knew it can produce the same wrong output. To tell them apart we have to look inside the model.

A line of work shows this is feasible. The truth of a statement is often encoded linearly in a model's activations, and simple probes recover it even when the model outputs something false (Orgad et al., 2025; Marks & Tegmark, 2024; Azaria & Mitchell, 2023). These results leave open questions that matter for safety. The probes are usually supervised or need a labeled truth corpus, and some designs assume a separately trained honest model. They are shown mainly on the truthfulness of declarative statements. To our knowledge, they have not been evaluated as a test that separates concealment from genuine removal, nor against models trained to conceal or against unlearning. This leaves a concrete question.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Can we read a model's concealed knowledge without a truth reference, and can one readout tell hiding apart from erasure, in the settings safety cares about?

We answer both by borrowing a tool from deception detection. The Concealed Information Test identifies guilty knowledge without a confession. It presents a suspect with the true detail among plausible decoys and looks for a stronger response to the item the suspect recognizes (Lykken, 1959; Verschuere et al., 2011; Ben-Shakhar & Elaad, 2003). Our method, Probe of Internal Recognition (PIR), runs the same test inside a model, reading a recognition signal from activations much as neural memory detection reads it from a brain response (Farwell & Donchin, 1991). Given a question and a set of candidate answers, PIR reads from the model's activations which candidate it recognizes as correct, using the contrast between the correct option and its distractors. PIR is reference-free. It needs no honest reference model and no labeled truth corpus, because the distractors supply the contrast.

PIR recovers concealed knowledge across models and forms of concealment. Under prompted concealment the readout reproduces across the Gemma, Qwen, Llama, Mistral, and Phi families, and it recovers answers from models trained to sandbag and from external password-locked and circuit-broken checkpoints we did not build (Hofstätter et al., 2025; Zou et al., 2024). It then marks a clear split between hiding and erasure. When a model hides an answer it knows, the answer stays decodable, and a divergence detector calibrated on the model's own base checkpoint flags the concealment. When unlearning removes the knowledge, recognition falls to the unknown-item baseline, both for our own run and for external unlearning checkpoints across families (Li et al., 2024; Zhang et al., 2024). The same readout therefore separates a model that will not answer from one that cannot, which is what sandbagging audits and unlearning verification need. The recognition direction is used by the model, since steering along it changes the answer while a matched output-direction control does not. The readout also adds information beyond black-box behavioral cues, most when a model conceals confidently, and a label-free detector built from the same signal flags concealed knowledge in free-form generation. This holds for models that are not trained to evade it, and a model adversarially trained to obfuscate its recognition can defeat the readout.

We make the following contributions.

- PIR, a reference-free readout that recovers a model's concealed answer from its internal recognition of the correct option among distractors, adapted from the Concealed Information Test.
- A hiding-versus-erasure boundary. Concealment that keeps the model's capability recoverable leaves the readout intact, while unlearning that removes the capability erases it. This gives a practical signal for sandbagging audits and unlearning verification.
- Evidence that the signal is used and additive. Steering along the recognition direction changes the model's answer, and the readout adds information over black-box behavioral detection.
- Generalization to free-form generation, where candidates are reconstructed from the model's own samples instead of supplied by the task.

2 Related Work

Reading knowledge and truth from internal states. A model's internal states often encode whether a statement is true, and linear probes recover this signal even when the model outputs something false (Azaria & Mitchell, 2023; Marks & Tegmark, 2024; Orgad et al., 2025; Bürger et al., 2024). Orgad et al. (2025) show that a model can internally represent the correct answer while generating a wrong one, and Mallen et al. (2024) read the correct answer, the model's latent knowledge, from models finetuned to make context-dependent errors, an early analogue of sandbagging. Some methods aim to be reference-free. Burns et al. (2023) recover latent knowledge from unlabeled activations through logical consistency, though later work questions whether such objectives find knowledge or the most salient feature (Farquhar et al., 2023). PIR builds on this line and departs from it. It reads

knowledge reference-free through a Concealed Information Test contrast between the correct option and its distractors, so it needs no truth label and no consistency assumption. It separates hiding from erasure. And it extends beyond declarative statements to free-form generation.

Probing, steering, and representation engineering. Reading a concept from activations and steering it are two sides of one method. Linear classifier probes date back to Alain & Bengio (2016). Representation engineering reads and controls high-level concepts such as honesty (Zou et al., 2023), and activation steering adds concept directions to the residual stream to change behavior (Turner et al., 2023; Rimsky et al., 2024). Inference-time intervention shifts activations along a truthfulness direction to elicit truthful answers (Li et al., 2023). Campbell et al. (2023) localize and causally flip instructed lying. Our recognition direction uses the same contrast-pair construction, and our causal test steers it to change the model's answer, with a matched control that steers the output direction and does not.

Detecting deception and sandbagging. A growing body of work detects deception and hidden capability. Black-box methods ask unrelated follow-up questions to flag lies (Pacchiardi et al., 2024), and benchmarks separate honesty from accuracy (Ren et al., 2025). Behavioral lie detectors that make a model self-report or classify its own deception generalize poorly across deception types and cannot recover a fact the model does not report (Hopkins et al., 2026), a limit PIR avoids by reading internal recognition directly. White-box probes detect deceptive responses and backdoored behavior at high accuracy (Goldowsky-Dill et al., 2025; MacDiarmid et al., 2024), and frontier models can scheme and sandbag on evaluations (Meinke et al., 2024), and trained deceptive behavior can persist through safety training (Hubinger et al., 2024). The sandbagging threat is studied through models that underperform on purpose (van der Weij et al., 2025) and password-locked organisms whose capability must be elicited (Greenblatt et al., 2024; Hofstätter et al., 2025). Tice et al. (2025) elicit hidden capability by injecting weight noise, a reference-free detector like ours. PIR recovers the concealed answer itself, works reference-free, and adds information over black-box behavioral cues. Where prior probes output a deception flag, PIR outputs which answer the model recognizes. We evaluate it on external password-locked and circuit-broken organisms (Hofstätter et al., 2025).

Does unlearning erase or suppress? Machine unlearning removes hazardous knowledge from a model. RMU (Representation Misdirection for Unlearning) misdirects the representations of forgotten content (Li et al., 2024), and NPO (Negative Preference Optimization) reframes forgetting as preference optimization to improve on gradient-ascent forgetting (Zhang et al., 2024; Yao et al., 2024). A recent line argues that these methods suppress knowledge instead of removing it, since light finetuning, benign relearning, or small weight edits recover much of the forgotten capability (Łucki et al., 2025; Hu et al., 2025; Deeb & Roger, 2024; Che et al., 2025), which motivates evaluation beyond input-output behavior (Lynch et al., 2024). Our results add a different view. The recognition signal that PIR reads is driven to the unknown-item baseline by unlearning, most completely by RMU, on the unlearned domain, even though the same signal survives every concealment that keeps the capability recoverable, including prompt deception, trained sandbagging, and circuit-breaking. We report this as a property of the readout at deployment, and we do not claim the weights can never be made to relearn.

Robustness of probes to adversarial evasion. Fixed activation probes can be evaded by an adversary with white-box access. Bailey et al. (2026) optimize a model's activations to bypass latent-space defenses, driving probe recall to zero while keeping the harmful behavior, and probe-based deception detectors carry the same caveat (Goldowsky-Dill et al., 2025). PIR inherits this limit. It is an audit signal for models that are not trained to evade it, and a model adversarially trained to obfuscate its recognition can defeat it.

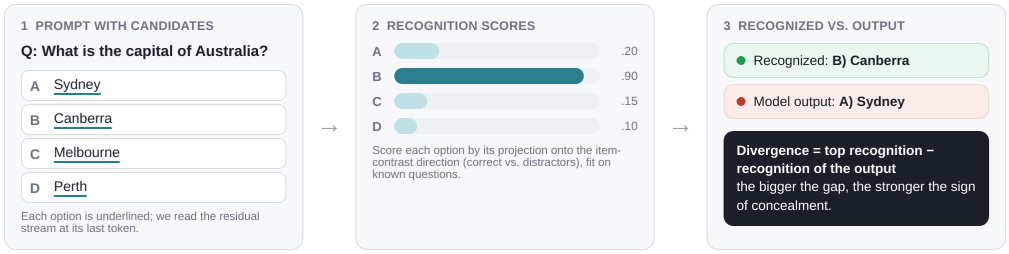

Figure 1: The PIR pipeline. We read each option's last-token span, project onto the item-contrast direction, and take the highest-scoring option as the recognized answer. The divergence detector compares that answer against the model's output.

3 Method

We describe PIR in three parts. We read the answer the model recognizes, we turn that readout into a label-free concealment flag, and we use the flag to separate concealment from erasure. Figure 1 gives the pipeline.

3.1 Setup and notation

We study multiple-choice questions. A question q has k_q candidate options $o_1, \ldots, o_{k_q}$, one of which, at index $c(q)$, is correct. We present q with its options to a model M and read the residual stream. For option o_j we take the hidden state at its last token, which we call the option span, and write $x^\ell_{q,j} \in \mathbb{R}^d$ for that vector at layer ℓ. The last token is the strongest read position, ahead of a mean over the option's tokens and well ahead of the first token (Appendix B). We use three conditions of a model. In the honest condition the model answers with no pressure to conceal. In the eliciting condition we recover its capability with a prompt, password, or trigger. In the deployed condition the model may conceal. The base model is the model's own checkpoint before sandbagging or unlearning. We say M knows q if it answers correctly in the honest condition, and M conceals on q if it knows q yet outputs an index $\hat{y} \neq c(q)$. The unknown-item baseline is the recognition level on questions the model does not know, which serves as our reference level. Our goal is to recover $c(q)$ from the activations $\{x^\ell_{q,j}\}$ when $\hat{y}$ is wrong. Figure 1 summarizes the readout.

3.2 Reading the answer a model recognizes

PIR reads which option M recognizes as correct by contrasting the correct option against its distractors. On a set $\mathcal{K}$ of the model's own known questions, we form the item-contrast direction

$$w^\ell \;=\; \frac{1}{|\mathcal{K}|} \sum_{q \in \mathcal{K}} \Big(x^\ell_{q,c(q)} - \frac{1}{k_q - 1} \sum_{j \neq c(q)} x^\ell_{q,j} \Big), \tag{1}$$

the mean difference between the correct-option span and the average distractor span. For a new question we score each option by its projection onto this direction and decode the recognized answer as the highest-scoring option,

$$r_j \;=\; \langle x^\ell_{q,j}, w^\ell \rangle, \qquad \hat{c} \;=\; \arg\max_j r_j. \tag{2}$$

We choose the layer ℓ by question-disjoint cross-validation, so the direction is never fit and evaluated on the same question. We call w^ℓ the recognition direction, and use that name for it in the rest of the paper.

This readout is reference-free. It needs no separately trained honest model and no external truth corpus, because the distractors supply the contrast and the labels for $\mathcal{K}$ are the model's own correct answers in the honest condition. When no labels are available, the top

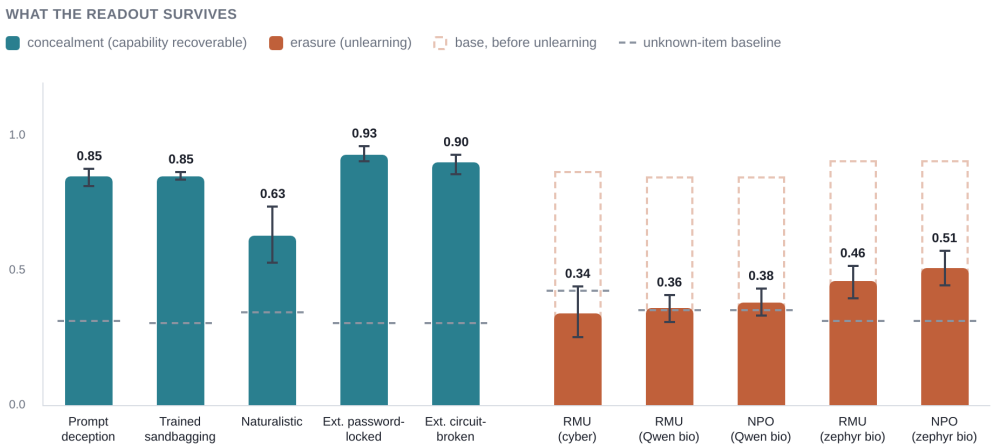

Figure 2: What the readout survives. Gray ticks mark the per-condition unknown-item baseline, dashed outlines mark base recognition before unlearning, and whiskers are bootstrap 95% confidence intervals.

principal component of the per-question centered option spans recovers a similar direction, so the readout can also run without labels. Because the contrast is taken within a question, between an option and its distractors, it reads the option the model treats as correct, which addresses the concern that unsupervised objectives may track the most salient feature of the text (Farquhar et al., 2023). On models whose activations carry a few very large dimensions, we standardize the spans before the projection; Appendix A gives details.

3.3 Detecting concealment without labels

At deployment we do not have the correct answer, so we turn the readout into a label-free flag. The divergence detector compares the recognized answer against the model's output. It scores each option with a direction $\tilde{w}^\ell$ calibrated off-distribution, fit on the base model answering the same questions in the honest condition and then applied to the deployed model's spans, so the direction is never fit on the concealed activations it scores. Writing $\tilde{r}_j = \langle x^\ell_{q,j}, \tilde{w}^\ell \rangle$, the detector score for a question with output $\hat{y}$ is

$$D \;=\; \max_j \tilde{r}_j \;-\; \tilde{r}_{\hat{y}}, \tag{3}$$

the recognition score of the top-recognized option minus that of the option the model output. D is small when the model outputs what it recognizes, or when no option is strongly recognized, and large when the model recognizes one option yet outputs another.

The base model here is the model's own pre-modification checkpoint, so the detector needs that checkpoint. This is available for unlearning verification and for auditing a released model against a tampered copy. It does not require a separately trained model that must be trusted as honest. We evaluate the detector by how well it separates two groups of questions. A question is known if the base model answers it correctly in the honest condition across all k_q cyclic rotations of the option order, a strict test that removes position bias and lucky guesses. A deployed instance is conceal if the question is known and the deployed output is wrong, and unknown if the question is not known. From the output alone these two look the same, since both give a wrong answer. We report the detector as the area under the ROC curve for separating conceal from unknown.

3.4 Distinguishing concealment from erasure

The readout also tells whether a wrong output hides an answer the model holds or reflects an answer that has been removed. We separate two regimes by whether the capability is

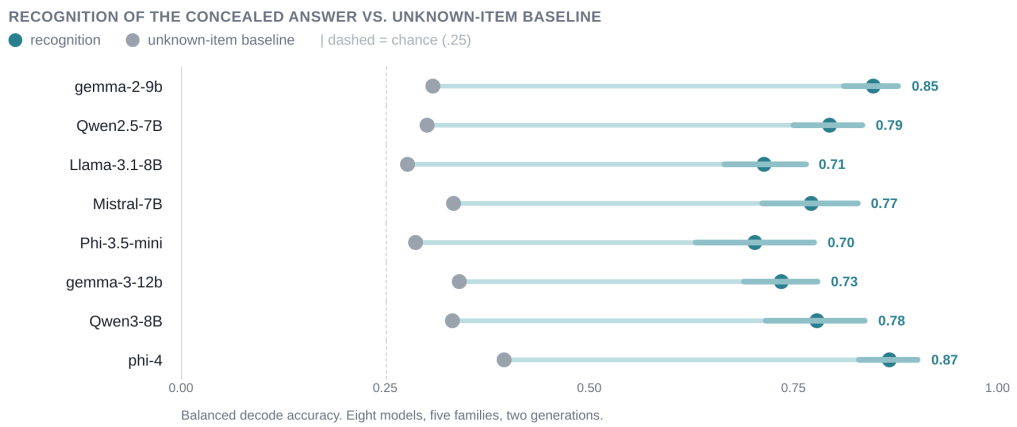

Figure 3: Recognition of the concealed answer against the unknown-item baseline under prompt deception, for eight models across five families and two generations. Whiskers are bootstrap 95% confidence intervals.

recoverable at inference time, meaning by a prompt, a password, or a trigger, with no weight update. Under concealment the capability remains and inference-time elicitation recovers it. Prompt deception, a trained sandbagging policy, and circuit-breaking all fall here, the knowledge is present, and PIR reads it. Under erasure inference-time elicitation does not recover the capability. Unlearning methods such as RMU and NPO fall here (Li et al., 2024; Zhang et al., 2024), and the readout falls to the unknown-item baseline. This inference-time criterion is what places unlearning on the erasure side. It does not require the knowledge to be unrecoverable by every means, and light finetuning can relearn some of it (Łucki et al., 2025), which we do not dispute. Because the criterion is a behavioral test that does not use PIR, the boundary is a property we verify, and it does not define the readout in terms of itself.

3.5 Extending to free-form generation

Free-form questions provide no option list, so we reconstruct one. We sample answers from the model under both an eliciting and a deployed condition and take the distinct normalized strings as candidates. Each candidate is placed in a short completion, "Question: q Answer: o_j", and we read its last-token span, as in the multiple-choice case. The item-contrast readout then decodes which candidate the model recognizes as correct. In free-form the conceal and unknown groups are hard to separate by the divergence score, because the reconstructed candidate set always offers some recognized option. We instead flag concealment by the peak recognition $\max_j r_j$, which is high when the model strongly recognizes a candidate and low when it recognizes none. This peak-recognition detector flags concealed knowledge in free-form for a trained sandbagging policy as well as a prompted one.

4 Experiments

We test PIR across models and forms of concealment, at the hiding-versus-erasure boundary, at deployment, and in free-form generation.

4.1 Setup

We use gemma-2-9b (the instruction-tuned checkpoint) as the primary model. It is a well studied mid-size instruction model, and both the external sandbagging organisms and the external unlearning checkpoints we test are built on comparable open models, which lets us reuse one pipeline. We also reproduce the readout on eight models from five families

Table 1: Detector AUC (conceal vs. unknown). Internal is the off-distribution recognition detector; combined adds the black-box features.

Organism	internal	text-only	black-box	combined
trained prefix (gemma-2)	0.718	0.517	0.645	–
external password-locked	0.892	0.660	0.845	0.929
external circuit-broken	0.848	0.621	0.860	0.899

(Gemma, Qwen, Llama, Mistral, Phi) across two generations, including phi-4, on which we also confirm the full pipeline. Datasets are MMLU, the three WMDP (Weapons of Mass Destruction Proxy) domains (biology, chemistry, cybersecurity), ARC-Challenge, and TriviaQA for free-form. Concealment comes in several forms. Prompt deception instructs a model to answer against what it knows. A trained sandbagging policy is a LoRA organism that answers correctly under a trigger and wrongly without it, in a prefix-triggered and a naturalistic (imitate-weak) variant. We also test two external organisms we did not build, a password-locked and a circuit-broken gemma-2-9b (Hofstätter et al., 2025). Erasure comes from unlearning, RMU and NPO, both our own runs and external checkpoints on Qwen2.5-7B and zephyr-7b-beta (Li et al., 2024; Zhang et al., 2024). We report recognition as balanced decode accuracy over the options, with the unknown-item baseline as the reference level, and detectors as AUC on the conceal-vs-unknown split. All readouts use question-disjoint cross-validation and bootstrap 95% confidence intervals. The raw readout is fragile on models with a few very large activation dimensions, so we standardize before the projection (Appendix A); on Gemma-3 this moves recognition from 0.389 to 0.735.

4.2 Reading concealed knowledge across models

Under prompt deception, PIR recovers the concealed answer well above the unknown-item baseline on every model. Across eight models from five families and two generations, recognition ranges from 0.70 to 0.87 while the unknown-item baseline stays between 0.28 and 0.40 (near 0.30 for most models), a gap of +0.40 to +0.54 (Figure 3). A control that reads the model's output option is high everywhere, and no model recovers the wrong output as the recognized answer, so the readout tracks the concealed answer. A belief-vs-output test confirms this. A binary within-condition probe that scores 0.985 by reading the output drops to chance under the test, while the item-contrast readout survives it (Appendix C).

The recognition direction is causal. On held-out known questions we add the direction, scaled by a dose α, to the residual stream at the tokens of a wrong option,

$$x^{\ell}_{q,j} \;\leftarrow\; x^{\ell}_{q,j} + \alpha\, w^{\ell}, \tag{4}$$

and check whether the model then outputs that option. This makes gemma-2-9b pick the steered option with probability 0.189 (Wilson CI $[0.126, 0.274]$), against 0.005 for 150 matched random directions (permutation $p < 1/150$) and monotone in α. A matched control that adds the option letter's output direction at the same tokens does nothing (0.000), and the recognition direction is nearly orthogonal to it (cosine 0.01). The steer reproduces on phi-4 (0.122, CI $[0.074, 0.194]$). The direction is used by the model upstream of the output, and it does something the output direction does not.

4.3 What the readout survives

The readout stays high under output-level concealment and collapses under erasure (Figure 2). Under a trained sandbagging policy on MMLU, recognition on concealed questions is 0.85 (CI $[0.83, 0.87]$) against a unknown-item baseline of 0.30, and the naturalistic organism gives 0.63 (CI $[0.52, 0.74]$) on the concealed subset and 0.81 on the full known set against 0.34. On the two external organisms, recognition is 0.93 (CI $[0.90, 0.97]$, password-locked) and 0.90 (CI $[0.85, 0.93]$, circuit-broken) against the unknown-item baseline near 0.30, so PIR reads a sandbagger built by others and reads through circuit-breaking.

Table 2: Free-form results on TriviaQA. Decode is recognition decode accuracy; the detector is peak-recognition AUC (conceal vs. unknown). Dashes mark not-applicable metrics.

Setting	decode	text-only	self-consistency	detector
honest model	0.71	0.54	0.56	–
trained sandbagger	0.74	0.68	–	0.776

Erasure is different. On a strict-gated run, unlearning gemma-2-9b with RMU drops base recognition of 0.85 (CI $[0.78, 0.92]$) to 0.39 (CI $[0.30, 0.49]$), which overlaps the unknown-item baseline at 0.35 (CI $[0.28, 0.41]$). The drop is domain-specific. A run that unlearns only cybersecurity drives cyber recognition to 0.34 (CI $[0.25, 0.45]$), below its unknown-item baseline of 0.42, while biology (0.91 to 0.68, CI $[0.62, 0.75]$) and chemistry (0.76 to 0.51, CI $[0.42, 0.59]$) stay readable. The erasure replicates on external checkpoints across families. On Qwen2.5-7B, biology recognition falls from 0.85 to 0.36 (RMU, CI $[0.31, 0.41]$) and 0.38 (NPO, CI $[0.33, 0.44]$), at the unknown-item baseline of 0.35; on zephyr-7b-beta from 0.91 to 0.46 (RMU, CI $[0.39, 0.52]$) and 0.51 (NPO, CI $[0.44, 0.58]$, a partial drop). Domain-specificity is clean on our own runs, and the external checkpoints over-forget, so we report them as a cross-model erasure replication.

4.4 Detecting concealment at deployment

At deployment the divergence detector, calibrated off-distribution, separates a model that hides an answer from one that does not know it. On the trained gemma-2-9b organism, the off-distribution direction decodes the concealed answer at 0.82, against 0.21 for the off-distribution unknown-item baseline, which can fall below chance, and 0.30 for a text-only decode, so the signal generalizes off-distribution and is not an option-text prior. As a detector it reaches AUC 0.718 (CI $[0.695, 0.739]$) on the hard conceal-vs-unknown split, above a text-only baseline at 0.517 ($+0.20$, CI $[0.17, 0.23]$) and above a combined black-box behavioral detector at 0.645 ($+0.073$, CI $[0.048, 0.097]$).

Whether internals beat black-box depends on how the organism conceals (Table 1). In that table the internal detector is single-layer for the prefix organism on MMLU and multilayer for the external organisms on WMDP, so we compare within each row. The two external organisms sandbag by emitting a low-confidence wrong answer, so the concealment leaks into the output distribution and black-box is strong. Once compute-matched and optimized, the internal readout beats black-box on the password-locked model (0.89 vs 0.85) and ties it on the circuit-broken one (0.85 vs 0.86). On both, the internal signal is additive. The combined monitor is best, at 0.93 and 0.90, with the internal contribution significant over black-box alone. So internals add information the outputs do not carry, and they dominate when a model conceals confidently. As concealment becomes confident, the concealed output looks like an honest answer and black-box detection falls toward chance, while the internal readout holds (Appendix D).

4.5 Beyond multiple choice

PIR extends to free-form generation, where the task gives no option list. We reconstruct candidates from the model's own samples on TriviaQA. Free-form has no clean unknown set, so here we compare recognition to chance in place of the unknown-item baseline. On the honest model, restricted to questions where it produces both a correct and a wrong candidate, the readout decodes the correct answer at 0.71, above chance at 0.42, above a text-only decode at 0.54, and above self-consistency at 0.56 (Table 2). The readout reads knowledge, and it reads it better than the model's own output distribution. Under a trained free-form sandbagging policy the reading survives. Recognition decodes the concealed answer at 0.74 against a text-only decode of 0.68, and the peak-recognition detector separates conceal from unknown at AUC 0.776, above a behavioral baseline ($+0.123$, CI $[0.052, 0.195]$). The decode advantage over text is small on factual questions, where the gold answer is also the most relevant candidate, so we report the detector as the free-form claim.

4.6 Evading the readout

The readout is an audit signal for models that are not trained to evade it. We train an anti-probe organism that keeps its capability under the trigger, sandbags without it, and is optimized to drive the recognition margin to zero. Under question-disjoint cross-validation this organism drives both a deployment-refit monitor and a fixed base-calibrated monitor down to the unknown-item baseline while keeping capability (0.71). An earlier report that a fixed direction reads through the attack was an in-sample artifact and is retracted. This matches the general finding that a white-box adversary can obfuscate activations against a fixed probe (Bailey et al., 2026). The attack leaves an option-span collapse that may be a detectable tamper, but that defense needs a base-model anomaly reference and an anomaly-aware adversary, which we leave open. Appendix E gives the training and the dose sweep.

5 Discussion and Limitations

A reference-free item contrast reads the answer a model recognizes as correct even when its output differs. This extends the finding that truth is linearly present in activations (Orgad et al., 2025; Marks & Tegmark, 2024); our contribution is what we do with it. The boundary is the most useful part for safety, and two audits follow. For sandbagging, a divergence between the recognized answer and the output flags a model hiding capability on an evaluation. For unlearning, a readout that falls to the unknown-item baseline is evidence that the knowledge is gone and not held back at the output. This second audit is also a caution. Because circuit-breaking leaves the readout intact, a model that looks unlearned to elicitation can still carry the knowledge, and PIR sees it.

We separate what the evidence shows from what it does not. The steering result shows the recognition direction is used by the model, since adding it changes the answer and a matched output-direction control does not. We do not claim the direction is necessary, and a variance-matched ablation did not establish necessity. The detector adds information over black-box cues and dominates when a model conceals confidently (Appendix D), while a low-confidence wrong answer already leaks to black-box detection, where internals mainly add to it.

PIR also has clear limits. It needs a set of candidate answers, so it covers multiple-choice and, through reconstructed candidates, free-form generation, but not open-ended reasoning with no enumerable answer. The deployment detector needs the model's own base or honest checkpoint, available for unlearning verification and release auditing but not for a fully black-box system. A model trained to evade the readout can defeat it, so the safety claim holds for models that are not adversarial to the probe. PIR reads only intact representations, so it cannot tell erased knowledge from knowledge never had, since both give a silent readout. The conceal and unknown labels come from a strict rotation-consistency gate, which discards questions the model does not answer robustly and shapes the detector numbers. The clean domain-specific erasure result rests on our own single RMU run, because the external checkpoints over-forget. Several organisms are our own and single-seed, external validation does not cover every concealment type, and the free-form detector is shown on one model and on factual questions.

6 Conclusion

A wrong answer from a model is ambiguous. It can hide knowledge the model holds or reflect knowledge it never had. We adapt the Concealed Information Test to model activations, reading the concealed answer with a reference-free item contrast that tells hiding apart from genuine absence, the distinction sandbagging audits and unlearning verification need. The audit holds for models that are not trained to evade it. As safety cases come to rest on capability evaluations and on unlearning, a reference-free read of what a model recognizes gives an auditor a signal its outputs cannot provide. We hope the boundary between concealment and erasure, and the means to test it, prove useful beyond the settings we study here.

AI use statement

We used generative AI to draft and polish the text, to find and check related work, and to help design experiments and write code. We did not use it to generate synthetic datasets or prove mathematical claims. The authors verified every claim, number, citation, and code output, and take responsibility for the final content.

Ethics statement

This work is a defensive audit method. Our experiments use public model organisms and the public WMDP benchmark, and we release no new hazardous capability. The most sensitive artifact is the anti-probe organism, which shows that a model can be trained to evade the readout. We report it so that developers and evaluators can anticipate the attack, and it confers no capability beyond defeating this monitor. We recommend using the readout as one signal within an audit and pairing it with other checks.

Reproducibility statement

Section 3 defines the readout and the divergence detector, Section 4.1 lists the models, datasets, and organisms, and the appendix gives the standardization and statistics. Experiments ran on single A100 GPUs. We will release the code and analysis scripts; the external organisms and unlearning checkpoints we use are already public.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644, 2016.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Findings of the Association for Computational Linguistics: EMNLP, 2023. arXiv:2304.13734.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass LLM latent-space defenses. In International Conference on Learning Representations (ICLR), 2026. arXiv:2412.09565.

Gershon Ben-Shakhar and Eitan Elaad. The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. Journal of Applied Psychology, 88(1):131–151, 2003. DOI:10.1037/0021-9010.88.1.131.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in LLMs. In Advances in Neural Information Processing Systems (NeurIPS), 2024. arXiv:2407.12831.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In International Conference on Learning Representations (ICLR), 2023. arXiv:2212.03827.

James Campbell, Richard Ren, and Phillip Guo. Localizing lying in llama: Understanding instructed dishonesty on true-false questions through prompting, probing, and patching. arXiv preprint arXiv:2311.15131, 2023.

Zora Che, Stephen Casper, Robert Kirk, Anirudh Satheesh, Stewart Slocum, Lev E. McKinney, Rohit Gandikota, Aidan Ewart, Domenic Rosati, Zichu Wu, Zikui Cai, Bilal Chughtai, Yarin Gal, Furong Huang, and Dylan Hadfield-Menell. Model tampering attacks enable more rigorous evaluations of LLM capabilities. Transactions on Machine Learning Research (TMLR), 2025. arXiv:2502.05209.

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? arXiv preprint arXiv:2410.08827, 2024.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. Challenges with unsupervised LLM knowledge discovery. arXiv preprint arXiv:2312.10029, 2023.

Lawrence A. Farwell and Emanuel Donchin. The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. Psychophysiology, 28(5):531–547, 1991. DOI:10.1111/j.1469-8986.1991.tb01990.x.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception with linear probes. In International Conference on Machine Learning (ICML), 2025. arXiv:2502.03407.

Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. Stress-testing capability elicitation with password-locked models. In Advances in Neural Information Processing Systems (NeurIPS), 2024. arXiv:2405.19550.

Felix Hofstätter, Teun van der Weij, Jayden Teoh, Rada Djoneva, Henning Bartsch, and Francis Rhys Ward. The elicitation game: Evaluating capability elicitation techniques. In International Conference on Machine Learning (ICML), 2025. arXiv:2502.02180.

Jack Hopkins, Dipika Khullar, Rowan Wang, and Fabien Roger. Fine-tuned lie detectors fail to generalize. Anthropic Alignment Science Blog, 2026. https://alignment.anthropic.com/2026/lie-detectors/.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In International Conference on Learning Representations (ICLR), 2025. arXiv:2406.13356.

Evan Hubinger, Carson Denison, Jesse Mu, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv preprint arXiv:2401.05566, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In Advances in Neural Information Processing Systems (NeurIPS), 2023. arXiv:2306.03341.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In International Conference on Machine Learning (ICML), 2024. arXiv:2403.03218.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for AI safety. Transactions on Machine Learning Research (TMLR), 2025. arXiv:2409.18025.

David T. Lykken. The GSR in the detection of guilt. Journal of Applied Psychology, 43(6): 385–388, 1959. DOI:10.1037/h0046060.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in LLMs. arXiv preprint arXiv:2402.16835, 2024.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, et al. Simple probes can catch sleeper agents. Anthropic Alignment Science Blog, 2024. https://www.anthropic.com/research/probes-catch-sleeper-agents.

Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. Eliciting latent knowledge from quirky language models. In Conference on Language Modeling (COLM), 2024. arXiv:2312.01037.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In Conference on Language Modeling (COLM), 2024. arXiv:2310.06824.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. arXiv preprint arXiv:2412.04984, 2024.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In International Conference on Learning Representations (ICLR), 2025. arXiv:2410.02707.

Lorenzo Pacchiardi, Alex J. Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y. Pan, Yarin Gal, Owain Evans, and Jan Brauner. How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions. In International Conference on Learning Representations (ICLR), 2024. arXiv:2309.15840.

Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Geralnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. The MASK benchmark: Disentangling honesty from accuracy in AI systems. arXiv preprint arXiv:2503.03750, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In Annual Meeting of the Association for Computational Linguistics (ACL), 2024. arXiv:2312.06681.

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. Noise injection reveals hidden capabilities of sandbagging language models. In Advances in Neural Information Processing Systems (NeurIPS), 2025. arXiv:2412.01784.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. arXiv preprint arXiv:2308.10248, 2023.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. In International Conference on Learning Representations (ICLR), 2025. arXiv:2406.07358.

Bruno Verschuere, Gershon Ben-Shakhar, and Ewout Meijer (eds.). Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, 2011. DOI:10.1017/CBO9780511975196.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In Advances in Neural Information Processing Systems (NeurIPS), 2024. arXiv:2310.10683.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In Conference on Language Modeling (COLM), 2024. arXiv:2404.05868.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, et al. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023.

Andy Zou, Long Phan, Justin Wang, Daniel Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, J. Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In Advances in Neural Information Processing Systems (NeurIPS), 2024. arXiv:2406.04313.

A Standardizing activations for massive-activation models

Some models place a few dimensions of the residual stream at very large magnitudes, which dominate a raw dot product and make the mass-mean readout fragile. Before forming the item-contrast direction and the projections, we standardize each activation dimension using statistics computed over the option spans of the fit set,

$$\tilde{x}_i \;=\; (x_i - \mu_i)/\sigma_i, \tag{5}$$

where μ_i and σ_i are the mean and standard deviation of dimension i over the fit-set option spans, then compute w^ℓ and the scores on the standardized spans $\tilde{x}$. An equivalent logistic readout, trained on the standardized correct-option and distractor spans, gives similar results. On Gemma-3 this raises recognition from 0.389 with the raw readout to 0.735 with the standardized readout.

B Readout ablations

Where the signal lives. We run the readout at each layer of gemma-2-9b under prompt deception (Figure 4). Recognition of the concealed answer sits at the unknown-item baseline in the early layers, rises to 0.85 at layer 24 of 43, and falls back toward the baseline in the late layers, while the unknown-item baseline stays near chance throughout. The signal is a middle-layer phenomenon and is weak at the layers nearest the output, which is why we select the layer by cross-validation rather than reading a fixed depth.

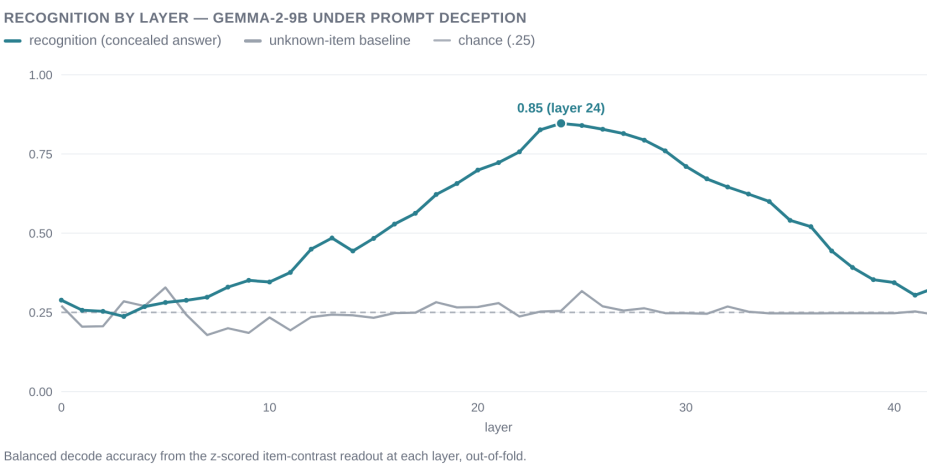

Figure 4: Recognition by layer under prompt deception. The concealed answer is decodable only in the middle layers; the unknown-item baseline stays near chance everywhere.

How little supervision it needs. Fitting the item-contrast direction at the best layer on as few as 10 of the model's own known questions already decodes the concealed answer at 0.82 (mean over five seeds), and accuracy is flat through 160 items (0.80 to 0.83). The direction is highly data-efficient, so the readout needs only a small labeled set of known questions.

Where in the option to read. PIR reads each option's last token. Comparing read positions on gemma-2-9b under prompt deception, the last-token span decodes the concealed answer at 0.85, a mean over the option's tokens at 0.79, and the first token at 0.53, against an unknown-item baseline of 0.25 to 0.32. The recognition signal concentrates toward the end of the option, so the last-token choice is the strongest of the three, with mean-pooling a close second.

C Confound controls and statistics

Belief-versus-output test. A within-condition probe can score well by reading the output and negating it, since with two options the correct answer is the negation of a lie. On a binary task such a probe reached 0.985. We control for this in two ways. We use $k > 2$ counterbalanced multiple choice, so the correct answer is not the negation of the output. And we fit the direction on trials where the model answers honestly and test it on trials where it conceals, so a signal that survives reads the belief. The binary probe drops to chance under this test, while the item-contrast readout survives it.

Statistics. Every headline claim is re-analyzed with a nested cross-validation point estimate (which removes best-over-layers selection), a permutation null taken as the max over layers (so the null pays the multiple-comparison cost), bootstrap resampling, and five fold seeds. Table 3 gives three regimes. The prompt-deception row uses gemma-3 with the raw readout, whose few very large activation dimensions depress every raw estimate for this model, the honest control included, so that row reads low across the board; the standardized readout raises gemma-3 recognition to 0.735 (Appendix A), and we keep the raw nested-CV values here only to exercise the permutation and selection controls. The boundary holds. RMU is the decisive row. The readout works (control 0.751), yet recognition sits at the unknown-item baseline (nested-CV 0.300 against a baseline of 0.339, the nested-CV estimate of the 0.39 reported in the main text, permutation $p = 0.086$), so the drop reflects a genuine absence while the readout still works. A two-one-sided-tests check places the RMU recognition-minus-baseline gap within a small margin of zero.

Table 3: Nested-CV recognition by regime (raw readout), with a permutation-null p-value and a positive control.

regime	recognition	baseline	perm-p	control
prompt deception (gemma-3, MMLU)	0.420	0.200	<0.001	0.340
output-gating (sandbag, WMDP)	0.891	0.360	<0.001	0.881
representation erasure (RMU, WMDP)	0.300	0.339	0.086	0.751

A label-free direction. The item-contrast direction can be found with no labels, as the top principal component of the per-question centered option spans (labels are used only to pick the component by cross-validation and to evaluate). It nearly matches the supervised direction, with MMLU 0.812 against 0.825, ARC 0.889 against 0.898, and WMDP-cyber 0.788 against 0.854 (chance 0.25).

Cross-dataset transfer. Fitting the direction on one dataset and applying it to another without refitting is near-lossless. Every off-diagonal is within 0.01 to 0.05 of the target's own within-dataset accuracy and far above the unknown-item baseline. A WMDP-fit direction reads MMLU at 0.815 and ARC at 0.900, and an MMLU-fit direction reads WMDP at 0.803. The recognition direction is a single model-level correct-option direction that transfers across datasets.

D Confident-concealment frontier

We test directly the claim that internals dominate black-box detection when a model conceals confidently. On gemma-2-9b under prompt deception, we compare a concealment run to an honest run on the same known questions, so the black-box negative is a fixed confident reference rather than the variable unknown set. We then bin the concealed questions by the model's output confidence on its emitted wrong answer. A black-box detector (logistic regression on output-distribution shape, entropy, max-probability, and margin) and the internal divergence detector each separate concealment from honest.

As concealment becomes more confident, the concealed output looks like an honest answer and the black-box AUC falls from 0.99 to 0.68, while the internal readout stays near 0.95

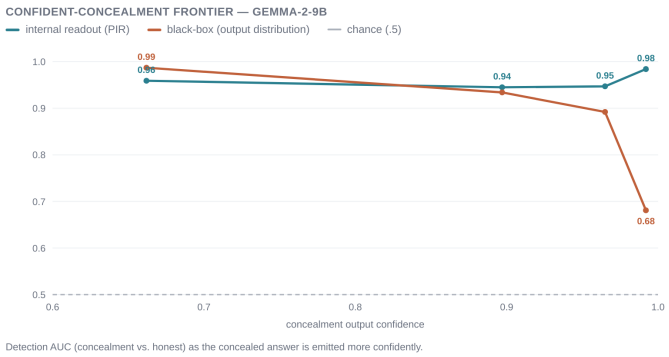

Figure 5: Confident-concealment frontier. As the concealed answer is emitted more confidently, black-box detection falls toward chance while the internal readout holds.

across the range (Figure 5). The internal signal reads the recognized answer from activations, so it does not depend on how confidently the model speaks. This is one organism on one model, and the black-box detector degrades but does not reach chance.

E The adversarial anti-probe organism

We train a LoRA organism with three objectives at once. It keeps the capability under the trigger, sandbags without it, and drives the item-contrast recognition margin to zero on the untriggered spans. We sweep the weight λ on the evasion term (Table 4).

Table 4: Anti-probe dose sweep. Recognition is on the untriggered known spans against a fresh question-disjoint probe.

λ	triggered capability	recognition (untriggered)	baseline
0.5	0.736	0.77	0.38
1.0	0.714	0.68	0.31
2.0	0.714	0.34	0.31
3.0	0.24 (collapsed)	–	–

At $\lambda = 2$ the model keeps its capability (triggered 0.714, triggered spans still readable at 0.83) and drives the untriggered recognition to the unknown-item baseline. We verified this under question-disjoint cross-validation, which corrected an earlier in-sample result. A fixed honest-calibrated monitor read 0.681 in-sample but 0.367 under proper cross-validation, at the unknown-item baseline, and an adversary trained directly against that fixed direction reached 0.407. The in-sample inflation appears only in this collapsed-span regime, where the option spans carry a large per-question self-correlation (cosine 0.997).

The attack removes the signal from the locus the monitor reads. It does not merely move the signal off one direction. On the untriggered known spans a stronger linear probe, an unsupervised principal-component direction, and a nonlinear multilayer probe all sit at the unknown-item baseline (0.30 to 0.35), while all read the triggered spans at 0.79 to 0.85. Below-chance untriggered accuracy ($0.11 < 0.25$) shows the model still computes the answer somewhere. The collapse is a plausible tamper signal, but a detector for it needs a base-model anomaly reference and an anomaly-aware adversary, which we leave open.