

Where the Transcript Runs Out: When Model Internals Beat Behavioural Monitors, Why the Strongest Generalisation Evidence Comes From Targets the Text Already Shows, and the Comparisons No Study Has Yet Run

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Linear probes and other readers of a language model's activations are now deployed as safety monitors, and two published literatures disagree about what they buy. One reports probes that match prompted and fine-tuned language-model monitors on real-world data after training only on synthetic examples, at a tiny fraction of the cost, and probes that flag deception which the transcript does not reveal. The other reports probes that fall to chance out of distribution, that lose much of their signal once the text that states the behaviour is removed, and that can be evaded by a model trained against them. In this analysis the author reviews and synthesises published results to show that the two literatures are measuring different premiums, and that the premium the case for white-box monitoring most needs is the one least often measured. The paper separates five sources of information a monitor can draw on - the public transcript, private context such as a system prompt or hidden reasoning, text the model has not yet produced, further queries to the monitored model, and its activations - and argues four points. First, where the label is a function of the public transcript, the measured premium is a cost premium plus a decorrelation of errors, and accuracy parity is what the evidence shows. Second, the generalisation evidence inverts the expected pattern: probes transfer best on targets the text already shows and worst on intent-like targets where only internal state would carry the label. Third, when evaluation labels are produced by a model reading the same transcript, an information premium on text-invisible cases cannot appear in the numbers at all. Fourth, "black-box" names two baselines, and against the one that can re-query the monitored model, white-box tools have so far added little. The paper consolidates the measured values in a table and a figure, states a comparison protocol as an algorithm, and names five studies that would settle what is open. On the record, the cost premium is established, the information premium is supported in a few designs with known ground truth, and the robustness premium under a matched adversary has not been tested.

Keywords: activation probes; model monitoring; white-box monitoring; deception detection; chain-of-thought monitoring; alignment auditing; distribution shift; AI control

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

A language model that is being served already computes, for every token, a stack of hidden states. A linear classifier reading one layer of those states costs almost nothing beside the forward pass that produced them. That simple fact is behind a fast-moving body of work on activation monitoring. McKenzie et al. (2025) train probes on synthetic examples of high-stakes interactions, test them on real conversations from several sources, and report performance in the range of prompted or fine-tuned language-model monitors of medium size, at roughly six orders of magnitude less compute. Kramar et al. (2026) describe probes for cyber-offensive requests that informed a deployment in a frontier production model. Cunningham et al. (2026) use a linear probe as the first stage of a jailbreak classifier cascade and report a cost reduction above forty-fold against their earlier exchange classifier. On the safety side, Goldowsky-Dill et al. (2025) train deception probes on simple contrastive data and report AUROCs between 0.96 and 0.999 on scenarios in which a 70B model conceals insider trading

or underperforms on purpose.

The argument for internals goes beyond cost. A model may say one thing and represent another. Reasoning traces omit influences that shaped the answer (Chen et al., 2025), produce coherent arguments for contradictory conclusions (Arcuschin et al., 2025), and become obfuscated when optimised against a monitor that reads them (Baker et al., 2025). A strategic model could fake alignment in its outputs (Greenblatt et al., 2024). If the monitor reads only what the model writes, a model that controls what it writes controls the monitor. Activations, on this view, are a channel the model does not author in the same way.

Against this sits a second literature. Dubanowska et al. (2025) find that representation-based hallucination detectors, once a spurious task-type correlation in the benchmark is controlled, do no better than linear probes in distribution, and that every method they test performs close to random out of distribution. Boxo et al. (2025) remove the tokens in which a behaviour is stated - the instruction to sandbag, the reasoning that reveals it - and watch probe AUROC fall by 10 to 30 points; on fine-

tuned model organisms that behave without verbalising the behaviour, the probes score 0.50 and 0.57. Kretschmar et al. (2025) evaluate lie detectors on seven datasets of qualitatively different lies and find a mean probe near chance on several, below an LLM judge reading the transcript. Bailey et al. (2024) show that harmfulness probes can be driven from full recall to none while the model keeps producing the harmful output. Parrack et al. (2025), who define the comparison most directly, find the gain from adding probe scores to a black-box monitor to be real but modest.

These results are not obviously about the same thing. The high-stakes probes and the hallucination probes target different properties; the deception probes were evaluated on data in which, by the authors' own account, the deception was usually explicit in the reasoning; the obfuscation results describe an adversary that the generalisation results do not include. The disagreement about whether internals help is, this paper argues, largely a disagreement about which comparison was run. The question in the title - under what conditions do model internals beat behavioural baselines - has an answer only once the conditions are named.

The paper names them. Section 2 fixes terms and defines the premium as a contrast between two monitors that differ only in whether they read activations. Section 3 separates five sources of information a monitor can draw on and derives what follows for each kind of target. Section 4 reads the deployment literature and argues that its premium is cost and decorrelation, not information. Section 5 assembles the evidence on targets that the transcript underdetermines, where an information premium would have to show. Section 6 argues that the generalisation evidence inverts the pattern the case for internals needs. Section 7 shows that evaluation labels produced by a transcript reader cap the measurable premium. Section 8 separates two black-box baselines and reads the auditing literature against the stronger one. Section 9 turns to adversarial pressure. Section 10 consolidates the measured values in Table 1 and Figure 1. Section 11 states a comparison protocol as Algorithm 1 and places the published studies against it in Table 2. Section 12 names five studies that would settle the open part, Section 13 states what is not known, and Section 14 gives the case against this paper's own argument.

Three precursors set the vocabulary, and this paper builds on each. Parrack et al. (2025) define the black-to-white performance boost: the improvement of a monitor given probe scores over the same monitor without them. Kirch et al. (2025) separate text-evident behaviours, whose definition lies in the response, from text-ambiguous ones, whose definition depends on why the response was produced, and show that probes trained off-policy fail mostly on the second kind. Zhong et al. (2026) name the elicitation confounder: an apparent gain from a white-box tool may be a gain that prompting

alone would have delivered. What this paper adds is the joint reading: an information-source decomposition that predicts where each published result should land, the observation that the generalisation evidence and the motivation for white-box monitoring point at different targets, the label-provenance ceiling, and a protocol under which the premium could be measured.

2. Terms and scope

Monitor. A function that takes some record of an interaction with a model and returns a score used to flag, block, escalate or audit. Runtime monitoring scores individual interactions as they occur; auditing scores a model, usually offline and with as many queries as the auditor chooses to spend. The distinction matters because the information a monitor can afford differs between the two.

Target. The property Y the monitor is meant to detect: a harmful request, a lie, sandbagging, a hidden objective, a hallucination, evaluation awareness. Targets differ in what they are a function of, and Section 3 makes that difference the organising variable.

White-box and black-box. A white-box monitor reads activations, gradients or weights of the monitored model. A black-box monitor does not. Section 8 argues that the black-box category hides two quite different baselines.

Premium. For a performance measure M, a black-box information set X and activations A, the white-box premium is

$$\text{Delta} = M(\text{monitor with X and A}) - M(\text{monitor with X alone}),$$

with the two monitors matched on everything else: training data, label source, compute budget, the threshold rule and the adversary they face. The definition is Parrack et al.'s black-to-white boost with the matching conditions made explicit. Each published comparison relaxes at least one of them, and which one it relaxes predicts much of what it finds.

Three premiums. The quantity Delta can be positive for three different reasons, and the paper keeps them apart throughout. A cost premium is equal accuracy at lower cost, or higher accuracy at equal cost when cost is binding. An information premium is higher accuracy at matched cost because A carries something about Y that X does not. A robustness premium is a smaller loss of accuracy under shift or under an adversary.

Out of scope. Interpretability used to understand a model rather than to flag its behaviour; steering and editing, except where a paper uses steering as an auditing tool; training against monitors, except as an adversary model (Section 9). The paper covers language models only.

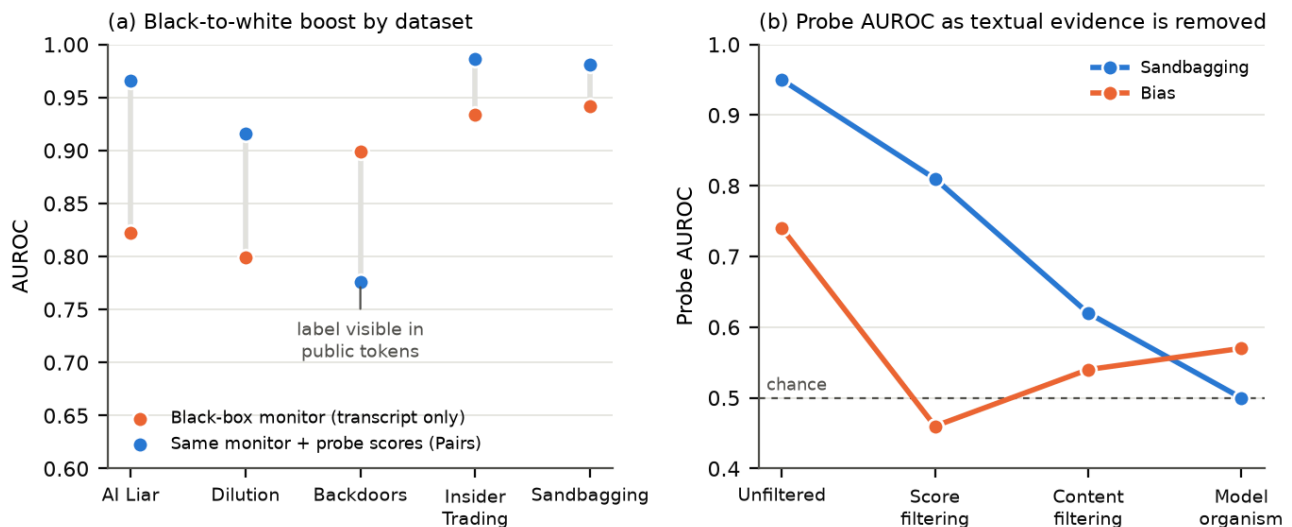


Figure 1. (a) AUROC of a black-box reasoning monitor (Claude 3.5 Sonnet reading the public transcript) and the same monitor given per-token deception-probe scores (instruction-pair probe), on five deception datasets with Llama-3.3-70B as the monitored model; on Backdoors the falsity of the response is determinable from public tokens and the probe scores make the monitor worse, though Backdoors is also the one code dataset, unlike the probe's training text (Parrack et al., 2025, Table 3 and Section 6). (b) Probe AUROC on sandbagging and grading-bias tasks with the text that states the behaviour kept, excluded from the probe's aggregate (score filtering), or deleted before the forward pass (content filtering) (Boxo et al., 2025, Figures 3-5), and for probes retrained on fine-tuned model organisms that behave without verbalising (Boxo et al., 2025, Section 6 and Conclusion, where the unfiltered sandbagging comparison is 0.94; the abstract gives 0.57 rather than 0.50 for the sandbagging organism). Points are joined to show the sequence of conditions; the organism condition uses a different model and probe.

3. Five sources of information, and what each predicts

A monitor can draw on five sources.

T, the public transcript. The user's input and the model's visible output, which any third party monitor could read.

P, private context. Tokens the model processed but a deployed transcript monitor may not see: a system prompt, a hidden scratchpad, retrieved documents, earlier turns withheld from the monitor.

F, the future. Text the model will produce but has not yet produced: the final answer while the reasoning is still running, the next action of an agent.

S, subject access. The ability to query the monitored model again: follow-up questions, prefilled responses, sampled personas, paraphrased prompts.

A, activations. The hidden states of the monitored model while it processed T and P.

Two observations organise the rest of the paper. First, A is not an independent source. The activations at a public token are a function of T, of P through attention, and of the weights W. They therefore carry information about P whenever P influenced the computation, information about F to the extent that the future is already determined by the present state, and information about the model's dispositions, encoded in W, that never reach the text. Parrack et al. (2025) make the dependence on P explicit in their design: they withhold private tokens

from both monitors and remove the probe scores at private positions, but the probe scores at public positions were computed with the private context in view. Boxo et al. (2025) separate the two routes by score filtering, which keeps the private text in the forward pass but excludes its positions from the probe's aggregate, and content filtering, which deletes the private text before the forward pass.

Second, which source a target depends on fixes where a premium can arise.

Case 1: Y is a function of T. A harmful request is harmful because of what it says; a list is a list because of its format. An ideal transcript reader then has all the information there is, and no premium in information is possible. Any measured Delta is a cost premium, a difference in the capacity of the two readers, or a difference in their training. This is not a weakness of probes. It is a statement about what a positive result in this case can mean.

Case 2: Y depends on P or F. A response to a sandbagging instruction is sandbagging because of the instruction; a reasoning trace is heading for an unsafe answer because of where it ends. A transcript reader denied P and F is structurally disadvantaged, and A can carry what it lacks. An information premium is possible here, and the size of the premium is bounded by how much of P and F survives into the activations at the positions the probe reads.

Case 3: Y depends on W. Whether the model knows the answer it denies, whether it would act differently if it

believed it were deployed, whether a trigger is present in its weights: these are facts about the model, not about any one transcript. A transcript reader can only infer them. Activations can carry them. So can S, since requerying the model samples its dispositions directly. In this case the right black-box baseline is not a transcript reader but a monitor with subject access, and the premium of A over S is a separate question from its premium over T.

The cases are not a taxonomy of behaviours; many targets mix them. Deception in an insider-trading scenario depends on T (the report), P (the earlier trades, visible in a scratchpad) and W (whether the model believes its report). What the decomposition does is predict, for a given experimental design, which premium the design can detect. A design in which the label is recoverable from T can show a cost premium and nothing else. A design that withholds P from the black-box monitor can show an information premium, but only about P. A design that compares A with T on a Case 3 target can show a premium over the weaker baseline only. The rest of the paper reads the literature through these predictions.

4. The premium that is deployed is a cost premium

The strongest positive results for activation monitoring come from misuse and harm detection on inputs and exchanges. In every one of them the target is close to Case 1, and in every one of them the result is parity in accuracy, or a probe advantage that tracks unmatched training data, at a large saving in cost.

4.1 High-stakes interactions

McKenzie et al. (2025) probe layer 31 of Llama-3.3-70B for interactions whose text suggests significant harm could follow. Probes trained on GPT-4o-generated synthetic data are evaluated on mostly natural conversations from six external datasets, labelled high- or low-stakes by GPT-4o with manual spot checks. Attention and softmax probes outperform 1B-parameter models and some medium baselines and land in the range of fine-tuned and prompted 8B-12B monitors; the larger prompted baselines, Gemma-3-27B and Llama-3.3-70B, have the best AUROC. In a same-family comparison reported in the appendix, the probe on Llama-3.3-70B scores 0.916 against 0.929 for the same model prompted, and on Llama-3.1-8B 0.887 against 0.877. The source reports the same prompted baselines twice, however: its chain-of-thought table, also a mean AUROC over the six test datasets, gives 0.970 for Llama-3.3-70B and 0.916 for Llama-3.1-8B, on which figures both probes fall below their prompted models. At a 1% false-positive rate the best probe catches 43% of high-stakes cases against 55% for the best language-model baseline. Probes surpass the best prompted baseline only after samples from the development split of

the evaluation datasets are added to their training data, and the authors note that they did not give the fine-tuned language-model baselines the same samples. A cascade that sends the probe's uncertain cases to an LLM matches the strongest baseline at a fraction of its cost.

The authors state the cost premium as their main finding, and the numbers bear them out. They also state, in their limitations, that because GPT-4o labelled both training and evaluation data the probes may be learning GPT-4o's interpretation of high stakes. Section 7 returns to that point, because it applies more widely than to this paper.

4.2 Cyber-offensive requests in a production model

Kramar et al. (2026) report the engineering of probes for cyber-offensive prompts to a frontier production model. Their main finding is a failure of generalisation from short to long contexts, which they address with new aggregation architectures; their main comparative table puts the test error of Gemini 2.5 Flash and Pro classifiers at 2.04% and 2.21% and of their best probes at 2.38% to 2.64%, with the probes ahead of Flash Lite. No single method is best on all nine of their evaluation datasets, and every method leaves an attack success rate above 1% on pre-existing jailbreaks. A probe-Flash cascade reaches lower test loss than Flash alone at roughly a fiftieth of its inference cost, and a cascade that calls the language model on fewer than one input in ten has a lower false-negative rate than the language model used alone. Again, parity in accuracy and a large cost advantage.

4.3 Jailbreak classifiers

Cunningham et al. (2026) train a linear probe on all layers of the model they protect and compare it with external classifiers in a jailbreak-defence system. The probe's attack success rate is better than that of their extra-small external classifier and close to that of a classifier built on a small production model; its flag rate on benign scientific requests is higher. Ensembling the probe with the small external classifier gives the lowest attack success rate of any pair they test, beating an ensemble of two external classifiers. Their explanation is the most interesting number in this literature for the present argument: the rank correlation between probe and external-classifier predictions is lower than between the two external classifiers. The probe errs differently. That is a decorrelation premium, and it is valuable for the same reason that independent verification is valuable (Mahendrakar, 2026), but it is not evidence that the probe knows something the text does not. Two readers of the same text trained differently also err differently.

4.4 The rest of the deployment record

The pattern repeats. Tillman and Mossing (2025), comparing prompted and unprompted probes, sparse-autoencoder probes and zero-shot prompting on moderation, question-answering and sentiment tasks,

find zero-shot prompting a strong baseline when inference compute is not limited and activation probes ahead only given enough training data; combining the model's prompted output with the probe roughly upper-bounds both but does not substantially exceed either. Kantamneni et al. (2025) find that sparse-autoencoder features do not consistently improve probing over baselines in four hard regimes. Oldfield et al. (2025) build truncated polynomial classifiers that trade compute for accuracy term by term and match or beat MLP probes of the same size on harmful-prompt classification. Hua et al. (2025) treat monitors as given and optimise how to combine them under a budget, more than doubling recall against a naive baseline in a code-review setting and showing that two monitors combined can Pareto-dominate either alone. Abdelnabi et al. (2024) detect task drift caused by injected instructions from the change in activations before and after external data, reaching near-perfect ROC AUC out of distribution. David and Fomin (2026) append a short classification instruction after the user turn and read the probe there, gaining up to about four AUC points out of distribution, and find that a content-free classification suffix works about as well as one naming the real criterion. Zhou et al. (2026) train probes for eighteen fine-grained misalignment indicators and match strong LLM judges at 0.935 AUROC on out-of-distribution transcripts, then use the probes as a pre-filter for an LLM adjudicator.

Every one of these is a credible result, and together they make the cost premium the best-established fact in the area. A probe that reuses activations the served model has already computed is several orders of magnitude cheaper than a second model reading the same text, and cascades that exploit this are a sound design. The targets, though, are inputs and exchanges whose label is fixed by their content - harm, cyber-offence, jailbreak, policy category, injected instructions - and the labels are produced by models or people reading that content. None of these results was designed to show, and none shows, an information premium.

5. Where the text runs out: the evidence on Cases 2 and 3

If internals carry information the text does not, the premium should appear where the label depends on private context, the future or the model's own state. The evidence here is thinner, more often from small model organisms, and mixed.

5.1 Private context

Parrack et al. (2025) give both their black-box and white-box monitors the public transcript only: system prompts and hidden reasoning are removed, on the reasoning that they stand in for unobservable cognition and external incentives. The monitored model is Llama-3.3-70B; the monitors are Claude 3.5 Haiku and Sonnet, given either the transcript (black-box reasoning)

or the transcript plus per-token probe scores (white-box reasoning). With the Sonnet monitor and the instruction-pair probe, mean AUROC across five deception datasets is 0.879 black-box and 0.925 white-box. The largest gains are on AI Liar (0.822 to 0.966) and Deception Dilution (0.799 to 0.916). On Backdoors the white-box monitor is worse, 0.776 against 0.899, and the authors give two reasons. The first fits the decomposition of Section 3: on that dataset the falsity of the response can in principle be determined from the public tokens alone, whereas on the others it depends on private tokens. The second is a confound: "the dataset consists in code, whereas other datasets are much more similar to the text on the probes were originally trained on." The Backdoors drop is therefore consistent with the probe adding noise where the label is in T, but it is also a drop on the one dataset where the probe is out of domain; where the label was in P, the probe carried some of P. Simply averaging the probe scores, with no language-model monitor at all, reaches 0.941.

Two features of the design bound what it shows. The monitors are stronger models than the one they monitor, and the authors note that this may overstate methods that rely on LLM reasoning. And the black-box monitor is denied P by construction, so the boost is a premium over a monitor that has been told less, not over the best monitor the deployer could build. For a deployment in which the system prompt and scratchpad are available to the monitor - most deployments of today's models - the comparison is with a stronger black-box baseline, and Parrack et al. observe that with access to private tokens such examples become trivially classifiable.

Boxo et al. (2025) attack the same question from the probe side. With the elicitation prompt and reasoning in context, their probes reach 0.95 AUROC on sandbagging and 0.74 on a grading-bias task. Excluding those token positions from the probe's aggregate drops sandbagging to 0.81 and bias to 0.46; deleting the text before the forward pass drops them to 0.62 and 0.54. Random deletion of the same number of tokens costs much less. On model organisms fine-tuned to sandbag or to grade with bias without saying so, probes retrained on contrastive pairs score 0.50 on sandbagging and 0.57 on bias, against 0.94 and 0.74 in the unfiltered evaluations (Figure 1b); the sandbagging value is 0.50 in the body of the source and 0.57 in its abstract. The authors read this as reliance on textual evidence. On the decomposition here, score filtering is the cleaner test of an information premium about P, since it keeps P in the forward pass and asks whether its trace in the public positions is enough; the answer is that some of it is, for sandbagging, and none of it is, for bias.

5.2 The future

The clearest positive result for a Case 2 target is about F. Chan et al. (2025) ask whether the reasoning trace of a reasoning model predicts whether its final response will

be safe. A linear probe on the trace's activations beats every text-based baseline - capable language models, fine-tuned classifiers and human readers - by an average of 13 absolute points of F1, and works on early segments before the answer exists. The gap comes mostly from what they call performative traces, in which the reasoning argues one way while the model heads the other. This is a premium of A over T about F: the label is in text that has not been written, and the activations already carry it. It is also the regime in which runtime monitoring is most valuable, since an alarm before the answer can stop it.

5.3 The model's own state

For Case 3 targets the evidence comes from designs that build the ground truth in. Mallen et al. (2023) fine-tune "quirky" models to answer wrongly whenever a keyword appears, and find that probes in middle layers usually report what the model knows rather than what it says, recovering 89% of the AUROC gap between truthful and untruthful contexts and 75% on questions harder than those used for training; an anomaly detector flags the untruthful behaviour at 0.95 AUROC. Orgad et al. (2024) find that models can encode the correct answer while consistently generating a wrong one, though their error detectors do not generalise across datasets. Dingeto (2026) adapts the forensic concealed-information test, presenting the true answer among decoys and reading which one the model's internal state recognises; the reported recognition holds between 0.85 and 0.93 under prompted deception, trained sandbagging and password-locked and circuit-broken checkpoints, and drops to the level of never-known questions after unlearning. That last result is a single-author preprint from September 2026 and has not been replicated.

Evaluation awareness gives a further set of Case 3 results, and a warning. Nguyen et al. (2025) separate evaluation from deployment prompts with probes on Llama-3.3-70B. Manek (2026), across eleven models, finds probe signal consistently stronger than the behavioural expression of evaluation awareness, with a family-dependent relation between the two that probe AUROC does not predict; Heidari et al. (2026) find awareness linearly decodable in every model they test (best AUROC at least 0.7) but only partly aligned with what the model verbalises. But Devbunova (2026), controlling prompt format in a 2x2 design, finds that such probes mainly track the canonical structure of benchmark prompts and fail on free-form prompts, and Singh and Chaudhary (2026) report that smaller models detect evaluation largely from format. A probe that "reads evaluation awareness" may be reading the benchmark template, which is in T. The power-law scaling of evaluation awareness reported by Chaudhary et al. (2025) is not supported under denser within-family sampling in Manek (2026).

5.4 Deception, where the two sides meet

Deception is the target the case for internals is most often made about, and it mixes all three cases. Goldowsky-Dill et al. (2025), evaluating on scenarios that include concealed insider trading (Scheurer et al., 2023) and deliberate underperformance on safety evaluations (Benton et al., 2024), report AUROCs of 0.96 to 0.999 and 95% to 99% recall at a 1% false-positive rate on chat data, and say plainly that their probes perform roughly on par with a GPT-4o black-box baseline and that in all their settings the deception is fairly blatant and often stated in the chain of thought. Golechha and Garriga-Alonso (2025) find probes trained on a simple instruction contrast generalising to deception in a social game at AUROC above 0.95, even when scored on the deceptive statement without the reasoning. In the sandbagging auditing game of Taylor et al. (2025), linear probes flagged the broadly sandbagging model across many tasks and were "arguably the most successful method pursued during phase one", though "the blue team did not trust probes sufficiently to rely on them for their predictions"; this was the phase in which black-box transcript analysis was defeated by imitation of a weaker model; the same probes failed to separate a narrowly sandbagging model from its unlocked version.

Against these, Kretschmar et al. (2025) evaluate on seven datasets designed so that the lie is verified against the model's established beliefs. At a threshold fixed for a 1% false-positive rate on control data, an LLM judge reading the transcript reaches 0.73 balanced accuracy, self-evaluation by follow-up question 0.66, a classifier on answers to unrelated follow-up questions 0.60 and the mean probe 0.54, with near-zero recall on several datasets. A probe trained with privileged examples from five of the datasets reaches 0.73 balanced accuracy and 0.91 AUROC on held-out examples of the same datasets. The judge, for its part, has almost no recall on the dataset where the lie concerns whether the model knows the answer at all, the purest Case 3 target in the benchmark. Thormann (2026) finds truth probes far better at catching false statements than at catching deception carried by misleading true ones. Natarajan et al. (2026) find deception heterogeneous: a single best probe gains 0.032 AUC over the baseline, matching probes to deception types post hoc gains 0.108, and the choice of system prompt used to build the training contrast explains 70.6% of the variance in probe performance across 2,000 configurations.

The deception evidence, then, supports an information premium in specific designs - private context withheld, a model organism with a known trigger, a broad and persistent behaviour - and does not support a general one. In the benchmark built to test lies that cannot be read off the transcript, the probe was the weakest method.

6. The inversion: probes generalise best where the text already shows the label

The live disagreement in this literature is a direct one. McKenzie et al. (2025) report probes trained on synthetic data generalising robustly to diverse real-world data; Dubanowska et al. (2025) report representation-based detectors performing close to random out of distribution. Both concern linear readers of hidden states, both are recent, and neither resolves the other. This section argues that the conflict is real but is not about probes. It is about targets, and the way it resolves is bad news for the motivation behind white-box monitoring.

6.1 Two targets

High stakes, in McKenzie et al., is a property of what the interaction says: a request that mentions a dangerous situation is high-stakes whoever answers it. Hallucination, in Dubanowska et al., is a property of the relation between the model's output and the world: the same sentence is a hallucination or not depending on facts the text does not contain. The first is Case 1; the second depends on W, since a probe for hallucination is in effect a probe for the model's own uncertainty or knowledge. Dubanowska et al. also show how a Case 1 shortcut can masquerade as a Case 3 signal: on the RAGTruth benchmark, one task type has a much higher hallucination rate than the others, and a classifier that uses nothing but the task type matches the AUC of the state-of-the-art detector. A probe can read the task type from the prompt's format.

Kossen et al. (2024) offer a partial counterpoint from the same target family: probes trained to predict semantic entropy, a measure computed from sampled generations, generalise better out of distribution than probes trained to predict accuracy directly. That is consistent with the reading here. Semantic entropy is a property of the model's output distribution, which the activations compute directly; accuracy depends on the world.

6.2 Text-evident and text-ambiguous

Kirch et al. (2025) supply the systematic version. Across eight behaviours and several models, probes trained on data from a different domain lose on average 0.162 AUROC on natural on-policy test data, a larger effect than switching from on-policy to off-policy training responses. Same-domain on-policy training gives 0.90 AUROC on average, same-domain off-policy 0.85, different-domain on-policy 0.74. The losses are concentrated in what they call text-ambiguous behaviours - sycophancy, deferral to authority, deception and sandbagging - while text-evident ones, such as refusal or using lists, transfer with small penalties. They add an observation that matters here: in their deception and sandbagging data the model was told its reasoning was private and stated its deceptive intent openly in the chain of thought, which in principle makes the

behaviour text-classifiable, and yet probes trained on that data still failed to generalise. Their conclusion is that current deception probes may fail to generalise to real monitoring.

The robustness results on deception probes point the same way. Kumar (2026) finds near-perfect AUROC on clean data collapsing under stylistic shifts in the Gemma 3 family, recovered when the probes are trained on style-augmented data, and concludes that the fragility reflects the narrowness of the training distribution. Moustafa et al. (2026) find that the type of lie in the training data changes detectability substantially and that sparse features do no better than dense ones. Yoo and Skapars (2026) find that out-of-distribution transfer depends on which principal components of the training activations the probe uses: projecting onto a small subset of the components gives cross-domain transfer that nearly matches probes trained on the target data, and the directions a source probe weights most heavily appear to encode surface features of the source. Mahadik and Skapars (2026) obtain better transfer by projecting onto directions built from contrasting harmful and harmless personas. Each of these offers a fix, and each fix is a way of removing a text-level correlate that the probe had latched onto.

The same weakness appears where the target is a rule rather than a behaviour. Sadhu et al. (2026) test compliance detectors - guard models and activation probes - by deleting, permuting or swapping the written rule a scenario is to be judged against, and find detection accuracy unchanged for every detector they test: the verdict follows the scenario, not the rule. Their own training-free activation readout fails a pre-registered test against trivial baselines, and a bag-of-words model matches its pooled generalisation. A probe can be blind to exactly the part of the input that is supposed to fix the label.

6.3 The inversion stated

Put together, the record supports a conclusion that the case for white-box monitoring has not absorbed. Probes tend to generalise well on targets defined by the content of the text, and badly on targets defined by why the text was produced. The first are the targets on which a transcript reader has all the information there is. The second are the targets on which white-box monitoring is supposed to earn its keep. Where the information premium would matter most, the probes with the best generalisation evidence are least likely to carry it across a shift.

The probing lineage that deception probes descend from was built on Case 1-like supervision. Azaria and Mitchell (2023) trained a classifier on hidden states to separate true from false statements; Burns et al. (2022) looked for a consistency-respecting direction without labels; Marks and Tegmark (2023) found linear structure separating true and false statements in curated datasets; Zou et al.

(2023) framed reading such directions as representation engineering, using contrastive instructions. Belinkov (2021) had already surveyed what probing classifiers can and cannot establish. In each of these, the label is a property of a statement or of an instruction present in the text.

There is a mechanism behind this, and it is not specific to linear probes. A supervised probe learns whatever direction best separates its training labels in its training data. If the labels are produced by text that states the behaviour - an instruction to lie, a contrastive persona prompt, a chain of thought that announces sandbagging - the most separable direction is often one that encodes the statement rather than the behaviour. Farquhar et al. (2023) make the unsupervised version of this point for contrast-consistent search: the method finds whatever feature is most prominent, which need not be knowledge. Levinstein and Herrmann (2024) find supervised truth probes falling below chance on negated versions of their test statements, and conclude that such probes are "learning some other feature that correlates well with truth on the training sets"; a sentence being true and containing no negation is one of the candidate features they name. Hewitt and Liang (2019) give the general caution for probes of any kind: a probe's accuracy is a fact about the probe as well as about the representation, and selectivity against a control task is needed to tell them apart. Natarajan et al. (2026) observe that instruction-pair probes appear to capture deceptive intent rather than content-specific patterns, which is the hopeful reading; their own finding that the system prompt explains 70.6% of performance variance says the intent captured is the one the prompt installed.

The inversion is not a proof that probes cannot carry Case 2 or Case 3 information across a shift. The quirky-model results of Mallen et al. (2023) and the forecasting result of Chan et al. (2025) show that they can in suitable designs. Two deception results also cut against it. Golechha and Garriga-Alonso (2025) report that probes trained on a "pretend you're a dishonest model" contrast, the kind of supervision blamed above, "generalize extremely well out-of-distribution" to deception in the Among Us game, above 0.95 AUROC even without the chain of thought; and Goldowsky-Dill et al. (2025) transfer instruction-pair probes to insider trading and sandbagging at 0.96 to 0.999. Neither fully undercuts the inversion, because in both the label is within reach of a reader of the text: the Among Us ground truth was produced "by passing the context and agent output to 'GPT-4o-mini' with a custom prompt", and Goldowsky-Dill et al. say the deception in their settings is often stated in the chain of thought. But both are generalisation successes on intent-like targets. The claim is therefore about a tendency in the current evidence, with stated exceptions: more of the published generalisation successes fall on Case 1 targets, and more of the published generalisation failures on the targets the

white-box argument is about.

7. The label-provenance ceiling

Section 4 noted that McKenzie et al. labelled their evaluation data with GPT-4o and flagged the risk that their probes learn GPT-4o's notion of high stakes. The issue is general, and it limits what the bulk of the literature can show.

Suppose evaluation labels are produced by a reader L of the transcript: $L = g(T)$. A monitor's measured accuracy is its agreement with L . On instances where T determines Y , a good L is right and any good monitor agrees with it. On instances where T underdetermines Y - the instances where a white-box monitor would beat a transcript reader - L is either wrong, in which case a monitor that tracks Y is scored as making an error, or uncertain, in which case curation pipelines often drop the instance. Either way, an information premium of A over T on exactly the text-invisible instances cannot appear in the measured numbers. What can appear is a cost premium, a decorrelation premium, and a premium of the probe over transcript readers weaker than L .

This is not a hypothetical concern. Goldowsky-Dill et al. (2025) label most of their evaluation responses with GPT-4o under dataset-specific prompts (sandbagging responses are labelled programmatically from the model's chain of thought), and compare their probe with a GPT-4o black-box baseline under a generic prompt; they report the two as roughly equivalent, which under the ceiling is close to the most a probe could show against a near-twin of the labeller. The parity is not only a ceiling effect, though: the authors gave the baseline "access to the entire conversation", which they describe as "slightly more information compared to what we provide the probe", since the probe reads the final output without the chain-of-thought reasoning the baseline saw. McKenzie et al. (2025) filter their synthetic training data by GPT-4o's stakes and confidence scores, removing ambiguous examples, and label evaluation data with GPT-4o. Zhou et al. (2026) produce per-turn labels with a strong LLM judge and exclude borderline transcripts by an ambiguity audit before evaluating probes against weaker LLM judges. Cunningham et al. (2026) train their probe on soft labels from a fine-tuned constitutional classifier. Each choice is reasonable for the paper's own purpose, and each one places the measurable premium inside the reach of a transcript reader. Shiromani and Richter (2026) are a partial exception: the LLM judge that labels hollow hacks in their live-agent study compares each trajectory "against the task's reference solution", information that a transcript reader does not have.

Shiromani and Richter (2026) also supply a check that the field could adopt widely. A probe on the activations of an open-weight surrogate monitor separates obfuscated reward hacks from legitimate solutions at

AUC 0.94 to 0.98, while the surrogate's own verdict manages 0.57 to 0.69. The tempting reading is that the monitor recognised the hack and its narration suppressed the recognition. A behavioural control argues against it: when the action itself carries a tell, the verbal verdict tracks the action, and the probe advantage on the remaining cases cannot be separated from a re-judgement of the action that a better-calibrated verdict would also make. The probe is also cross-validated on the evaluation population while the verdict is zero-shot, an asymmetry in supervision that Section 11 lists among the matching conditions. The authors' recommendation is to run such a control before reading a probe-over-verdict gap as concealed knowledge.

The escape from the ceiling is ground truth by construction. Model organisms whose trigger or hidden behaviour was trained in (Mallen et al., 2023; Boxo et al., 2025; Taylor et al., 2025; Sheshadri et al., 2026; Zhong et al., 2026; Cywinski et al., 2025), lies verified against beliefs established by separate queries (Kretschmar et al., 2025) and forecasting targets whose label is the model's own later output (Chan et al., 2025) all fix Y independently of what a transcript reader would say. It is in these designs that the information premium can be measured, and Sections 5 and 8 show that in them it is sometimes present, sometimes absent, and smaller against the stronger black-box baseline.

8. Two black-box baselines

"Black-box" is used for two different monitors. One reads the transcript, possibly with private context: an LLM judge, a fine-tuned classifier, a guard model such as Llama Guard (Inan et al., 2023), a constitutional classifier (Sharma et al., 2025). The other can query the monitored model again. Pacchiardi et al. (2023) showed early that a logistic regression on a model's yes/no answers to a fixed set of unrelated follow-up questions detects lies, generalises across architectures and settings, and needs neither activations nor knowledge of the truth. That is subject access without activations, and for Case 3 targets it reaches the same thing activations reach: the model's own state.

The auditing literature compares white-box tools mostly against this second baseline, and the results are sobering for the white-box side.

Cywinski et al. (2025) train models to apply secret knowledge - for example, that the user is female - while denying it, and test elicitation techniques by whether they help an auditor guess the secret. Their best techniques in every setting are black-box prefill attacks, in which the model is made to continue a response from a chosen prefix, and sampling from the user persona; these exceed 90% success on two of their three settings. White-box techniques based on the logit lens and sparse autoencoders consistently improve on the bare input-output baseline but are less effective, and the same

information extracted from the untuned base model does not help, which shows the white-box signal is real.

Sheshadri et al. (2026) build AuditBench, 56 models with 14 implanted hidden behaviours trained not to confess, and an investigator agent that can be given one of thirteen tool configurations. Scaffolded black-box tools - text completion, user-persona sampling and prefilling, each driven by an auxiliary model that generates many inputs - are the three best, at 58.0%, 53.2% and 51.6% success. White-box tools help on some model types: averaged over five white-box tools, success on models implanted by synthetic-document training with SFT adversarial training rises from 40.2% to 46.9%, but on models trained on transcripts it stays flat (34.4% against 33.5%), and on models adversarially trained with KTO the white-box tools are on average a few points worse than the default. Honesty steering is harmful across the board. The authors also report a gap between a tool's standalone usefulness and its usefulness to an agent.

Zhong et al. (2026) design Pando to break the elicitation confounder. Across 720 fine-tuned models implementing hidden decision rules, models are trained to give faithful explanations, no explanation, or confident explanations of a distractor rule. When explanations are faithful, black-box elicitation matches or beats every white-box method. When they are absent or misleading, gradient-based attribution adds 3 to 5 percentage points and relevance patching adds most, while logit lens, sparse autoencoders and circuit tracing give no reliable benefit. This is the cleanest published estimate of an information premium under ground truth, and it is positive, small in held-out accuracy (+2.9 to 4.7 points for the gradient methods), larger on field-level F1 (relevance patching +16.8 points without explanations) and at low query budgets (10.1 to 10.3 points at the smallest), and specific to methods that track the computation of the decision rather than its representation.

Marks et al. (2025), in a blind auditing game against a model trained with a hidden objective, report three of four teams succeeding with techniques that included sparse autoencoders, behavioural attacks and training-data analysis; the white-box tools were among several routes to the answer, not a necessary one. Tice et al. (2024) detect sandbagging by injecting noise into weights and watching performance rise rather than fall, a white-box intervention that reads neither activations nor text. Daniels et al. (2026) generate system prompts that deceive both black-box and white-box auditing methods into confident wrong answers, the first documented case, in their account, of activation-based strategic deception.

For runtime monitoring the comparison is different again. Subject access costs extra queries per interaction and may alter what the model does next, so the baseline that matters at runtime is usually the transcript reader.

For auditing, where queries are cheap and the target is a property of the model, the subject-access baseline is available and, on the record, strong. A claim that internals beat behaviour should say which of the two it beat. Kretschmar et al. (2025) illustrate the point inside one benchmark: on the dataset where a transcript judge has almost no recall because the lie concerns the model's own knowledge, the unrelated-question classifier, a subject-access method, achieves high AUROC for some models, and the authors attribute this to its indirect access to the model's private knowledge.

9. Under pressure

The motivating threat for white-box monitoring is a model that manages its outputs. If the same model can manage its activations, the premium in that threat model is unclear. The record has strong results on both sides and almost no direct comparison.

9.1 Activations can be shaped

Bailey et al. (2024) find embedding and fine-tuning attacks that keep a model's harmful behaviour while making its activations inconspicuous to probes, sparse autoencoders and out-of-distribution detectors; against harmfulness probes, recall can fall from 100% to 0% while the jailbreak succeeds 90% of the time. Obfuscation has a cost on hard tasks: when the monitored behaviour is writing SQL, obfuscating against an SQL probe makes the code less often correct. McGuinness et al. (2025) fine-tune models to evade probes for benign concepts when told which concept is being probed, and find that the evasion transfers zero-shot to safety probes the model never saw, including probes trained after the fact on the frozen weights; ensembles and non-linear probes are more resilient but do not close the gap. Gupta and Jenner (2025) train models with reinforcement learning to evade probes they get no gradients from, and find that monitors which average per-token scores are highly vulnerable, while monitors that max-pool or use attention over tokens hold up. Their decomposition adds a detail that matters: most of the evasion came from the model changing what it wrote, with a smaller but significant part from changed activations for the same text.

9.2 Training against a detector

Cundy and Gleave (2025) put a lie detector into the labelling step of preference learning. With GRPO the learned policy can become deceptive and evade the detector, at deception rates above 85%; with a sufficiently accurate detector or strong KL regularisation it becomes honest, and with DPO deception stays under 25% at realistic detector accuracy. Hollinsworth et al. (2026) report that undetected deception falls from 34% at 1B parameters to 14% at 405B at a 99% detector true-positive rate, while distribution shift between detector training and preference data can push false-positive rates to impractical levels. Taufeeque et al. (2026) build a

coding environment where reward hacking arises naturally and train against a deception probe. They separate two failure modes: an obfuscated policy, in which the model writes deceptive text that the detector does not flag, typically by adding a justification for the hack, and obfuscated activations, which they find arise from representation drift during RL whether or not the detector is penalised. High KL regularisation together with a high detector penalty yields honest policies.

9.3 The text channel under the same pressure

The text channel fails under pressure too, which is why Korbak et al. (2025) describe chain-of-thought monitorability as a fragile opportunity that development choices could remove. Baker et al. (2025) find chain-of-thought monitoring much more effective than action monitoring for reward hacking, and find that optimising the policy against the monitor produces obfuscated reward hacking. Shiromani and Richter (2026) rewrite only the reasoning of agents that hacked a coding benchmark, leaving every command identical, and drop a held-out monitor's catch rate on the reasoning-dependent subset from about 95% to between 4% and 11%. Kale et al. (2025) find that an agent's awareness of being monitored substantially degrades monitor reliability, while giving the monitor more information about the agent helps less than expected. Emmons et al. (2025) find the converse case for the text side: when the harmful task genuinely needs multi-step reasoning, models struggle to hide it from a CoT monitor without substantial help. Arnav et al. (2025) find that a hybrid that scores reasoning and actions separately beats either alone, with detection rates about twice those of action-only monitoring for subtle deception. Kutasov et al. (2025) report a best monitor at AUC 0.87 on SHADE-Arena, and that sabotage success depended heavily on a hidden scratchpad the monitor could not see.

9.4 What has not been compared

The robustness premium would be measured by giving one adversary one budget against a white-box monitor and a black-box monitor of matched accuracy, and comparing the loss. No study found in this review does that. The adversaries differ in access (gradients through the probe for Bailey et al., reward only for Gupta and Jenner, a trigger-conditioned fine-tune for McGuinness et al.), in objective, and in what the black-box monitor would have seen. Two results suggest the comparison could favour combinations rather than either channel. The probe errors in Cunningham et al. (2026) are less correlated with external classifiers than the classifiers are with each other, and Gupta and Jenner (2025) find most evasion running through changed outputs, which a text monitor could in principle catch. Hubinger et al. (2024), whose backdoored models kept their behaviour through safety training and whose adversarial training taught models to recognise their triggers better, are a reminder that a detector used as a training signal can teach

concealment. That applies to probes and transcript readers alike. Mechanistic anomaly detection offers a route that does not depend on a probe direction at all. Johnston et al. (2025) find detectors on quirky models that discriminate well on some tasks but none that works across all models and tasks, and Lyons Keenan et al. (2026) recast the problem as functional attribution through influence functions and report gains over baselines on backdoored language models, including models explicitly trained to obfuscate. Whether such methods keep their edge against an adaptive adversary with a matched budget is, again, untested. Benign shift matters as well: Duan (2026) finds that quantisation mostly preserves frozen probes while fine-tuning frequently makes them stale, that the degradation is predictable, and that label-free realignment repairs it. The AI control framing of Greenblatt et al. (2023) asks for protocols evaluated against a model actively trying to subvert them. The white-box premium has not yet been evaluated that way.

10. The numbers in one place

Table 1 collects the published measurements the argument rests on, one cited result per row, grouped by the case of Section 3 that the target most closely fits. The case assignment is the author's; every value is the source's own.

Two cautions apply to Table 1. The rows come from different models, metrics and thresholds, and invite comparison within rows rather than across them. And many rows are 2025-2026 preprints, several from single groups; Section 14 says which conclusions would weaken if they fail to replicate.

Figure 1 shows the two patterns the argument turns on. In panel (a), the probe helps where the label sits in withheld context and hurts where it sits in the transcript, a drop that is consistent with the decomposition but confounded by domain shift, since the Backdoors dataset is code. In panel (b), as the stated behaviour is removed from what the probe can draw on, its signal falls toward chance.

11. A comparison protocol

The analysis implies a protocol for any claim that internals beat behaviour. Its organising rule is that a premium is reported only against the strongest black-box baseline the deployment could actually use, under matched supervision and label provenance, and separately for each of the three premiums.

Table 2 places a selection of the published studies against the decomposition. It does not grade them; each was designed for its own question. It records which premium each design could reveal, given its target, its comparator and its label source.

Read down the last column, the table states the paper's argument compactly. The designs that can show only a

cost premium report one. The designs that can show an information premium against a transcript reader denied private context or the future report one. The designs that compare against subject access report a small premium or none.

12. What would settle the open part

Five studies would resolve most of what this analysis leaves open. Each is a design proposal for future work.

S1. A premium ladder on one target set. Take a set of behaviours spanning the three cases, each with ground truth fixed by construction - for example, model organisms in the style of Boxo et al. (2025) and Taylor et al. (2025), a quirky-model knowledge task (Mallen et al., 2023) and a belief-verified lie set (Kretschmar et al., 2025). Run a transcript reader without private context, one with it, a subject-access method, a probe, and the probe added to each reader, all trained on the same labelled data and reported at matched compute. The prediction from Section 3 is a premium near zero in information for Case 1, positive against the transcript reader without private context in Case 2, and small against subject access in Case 3. A different pattern would revise the decomposition.

S2. Silent-behaviour transfer. Train probes in the standard way, on text that states the behaviour, and evaluate on organisms that perform the same behaviour without stating it, for every behaviour in Kirch et al. (2025). Report selectivity against control probes in the sense of Hewitt and Liang (2019). The inversion of Section 6 predicts that transfer is poor for intent-like behaviours. Training methods that pass this test - style augmentation, subspace selection, persona coordinates - would then be evaluated on the test that matters rather than on held-out text-evident data.

S3. A label-provenance audit. Take an existing probe benchmark labelled by an LLM reading the transcript and relabel a sample with ground truth obtained independently: verified beliefs, a known trigger, or the model's later behaviour. Estimate the fraction of items on which the transcript underdetermines the label, and recompute the probe-versus-judge comparison on that fraction alone. This would put a number on the ceiling of Section 7.

S4. A matched-adversary comparison. Apply one optimisation budget - for example, reinforcement learning against the monitor's score with a fixed KL constraint - to a max-pooled or attention probe, a transcript monitor of matched clean accuracy, and their ensemble. Report evasion success, the task performance cost of evasion (Bailey et al., 2024), and whether the evasion runs through changed text or changed activations (Gupta and Jenner, 2025). This is the robustness premium, and it has not been measured.

Table 1. Published measurements behind the argument. "WB" is the white-box monitor, "BB" the black-box comparator. Each value is taken from the source named on its row; the grouping by case is the author's.

#	Source	Target (case)	WB	BB comparator	Reported value
1	McKenzie et al. 2025	High-stakes (1)	Probe, Llama-3.3-70B	Same model prompted	AUROC 0.916 vs 0.929 (0.970 in App. B.8)
2	McKenzie et al. 2025	High-stakes (1)	Probe, Llama-3.1-8B	Same model prompted	AUROC 0.887 vs 0.877 (0.916 in App. B.8)
3	McKenzie et al. 2025	High-stakes (1)	Best probe	Best LLM monitor	TPR at 1% FPR 43% vs 55%
4	Kramar et al. 2026	Cyber misuse (1)	Best probes	Gemini 2.5 Flash, Pro	Test error 2.38-2.64% vs 2.04%, 2.21%
5	Cunningham et al. 2026	Jailbreak (1)	Probe in cascade	Exchange classifier	Compute cut by over 40x
6	Zhou et al. 2026	Misalignment (mixed)	Indicator probes	Strong LLM judge	Probes match judge, AUROC 0.935
7	Parrack et al. 2025	Deception, P withheld (2)	Sonnet + probe	Sonnet, transcript	Mean AUROC 0.925 vs 0.879
8	Parrack et al. 2025	Backdoors, label in T (1)	Sonnet + probe	Sonnet, transcript	AUROC 0.776 vs 0.899
9	Boxo et al. 2025	Sandbagging (2)	Probe, text kept vs removed	none	AUROC 0.95, 0.81, 0.62
10	Boxo et al. 2025	Silent sandbagging (3)	Probe on organism	none	AUROC 0.50 vs 0.94 unfiltered (0.57 in abstract)
11	Chan et al. 2025	Unsafe final answer (F)	CoT activation probe	Best text baseline	+ 13 F1 on average
12	Mallen et al. 2023	Knowledge vs output (3)	Contrast-pair probe	Model output	89% of AUROC gap recovered
13	Goldowsky-Dill et al. 2025	Deception (mixed)	Instruction-pair probe	GPT-4o judge	AUROC 0.96-0.999; roughly on par
14	Kretschmar et al. 2025	Belief-checked lies (3)	Mean probe	LLM judge	Balanced acc. 0.54 vs 0.73
15	Kirch et al. 2025	Eight behaviours	Probes, domain shift	none	AUROC 0.74 vs 0.90 same-domain
16	Dubanowska et al. 2025	Hallucination (3)	Probes and SOTA	none	Near random out of distribution
17	Natarajan et al. 2026	Deception (mixed)	Single vs matched probes	Baseline probe	+ 0.032 vs + 0.108 AUC
18	Cywinski et al. 2025	Secret knowledge (3)	Logit lens, SAE	Prefill (S)	Prefill over 90% on 2 of 3; WB lower
19	Sheshadri et al. 2026	Hidden behaviours (3)	Five WB tools	Scaffolded BB (S)	Best BB 58.0% overall; WB avg 46.9% vs 40.2% default on SDF + SFT models only, flat on TD, below default on KTO
20	Zhong et al. 2026	Hidden rule (3)	Gradient attribution	Elicitation (S)	+ 3 to 5 pp only if no faithful explanation
21	Shiromani and Richter 2026	Obfuscated hacks	Surrogate probe	Surrogate verdict	AUC 0.94-0.98 vs 0.57-0.69
22	Bailey et al. 2024	Harmfulness, attacked	Probe	n.a.	Recall 100% to 0% at 90% jailbreak
23	Cundy and Gleave 2025	Lie detector in training	Detector as label	n.a.	GRPO deception over 85%; DPO under 25%
24	Hollinsworth et al. 2026	Lie detector in training	Detector at 99% TPR	n.a.	Undetected deception 34% (1B) to 14% (405B)

S5. A forecasting premium for agents. Extend Chan et al. (2025) from final answers to agent actions: predict, from activations during reasoning, whether the next tool call will be harmful, against a chain-of-thought monitor reading the same reasoning, and apply the behavioural control of Shiromani and Richter (2026) to the gaps. If the premium about the future holds for actions, it is the strongest practical argument for white-box runtime monitoring.

13. What is not known

No published study measures the white-box premium for one target against a transcript reader, a transcript reader with private context, and a subject-access method under matched supervision and matched compute.

No published study measures the robustness premium: the loss of a white-box monitor and of a black-box monitor of matched accuracy under the same adversary and budget.

The fraction of deployed monitoring traffic on which the transcript underdetermines the label is not known for any target. It sets the ceiling on the information premium in deployment, and nobody has estimated it.

Algorithm 1. Measuring a white-box monitoring premium

```

Input: target Y; monitored model; deployment mode (runtime | audit);
       probe family; candidate black-box monitors; cost budget C
1 CLASSIFY Y by what fixes its label: T only (case 1), P or F (case 2), weights (case 3)
2 CHOOSE baselines:
   runtime -> strongest transcript reader allowed to see what the deployment shows it
   audit   -> transcript reader AND a subject-access method (re-query, prefill, persona)
3 LABELS: if labels come from a reader of T, mark the run "ceiling-bound" and
   report only cost and decorrelation premiums; for an information premium,
   use ground truth fixed independently of T (organism, verified belief, later output)
4 MATCH supervision: give every monitor the same labelled data and the same
   in-distribution samples; or report zero-shot and trained results separately
5 MATCH cost: report each monitor at equal compute and at equal accuracy
6 STRATIFY the test set by text-evidence: items whose label is recoverable from T
   vs items where it is not; report Delta on each stratum
7 SHIFT: evaluate on a held-out domain and on silent model organisms;
   include a selectivity control (random or control-task probe)
8 CONTROL: where WB beats a BB verdict, vary the blatancy of the action
   and check whether the verdict tracks it before reading the gap as hidden knowledge
9 ADVERSARY: apply one attack budget to WB, BB and WB+BB monitors of matched accuracy;
   report evasion rate and the attacker's capability cost
10 REPORT Delta_cost, Delta_info (per stratum, per baseline), Delta_robust, and the ensemble

```

Table 2. What each design could reveal. "T" is the transcript, "P" private context, "S" subject access. The classification is the author's reading of each source's stated design.

Source	Target case	Comparator	Label source	Premium the design can show
McKenzie et al. 2025	1	T readers	GPT-4o on T	Cost, decorrelation
Kramar et al. 2026	1	T readers	Curated, partly expert	Cost
Cunningham et al. 2026	1	T classifiers	Classifier soft labels	Cost, decorrelation
Goldowsky-Dill et al. 2025	mixed	T + P reader (full conversation incl. CoT)	GPT-4o, per-dataset prompts; Sandbagging from CoT	Bounded by labeller
Parrack et al. 2025	2	T without P	Scenario design, uses P	Information about P
Chan et al. 2025	F	T so far	Later output	Information about F
Kretschmar et al. 2025	3	T reader and S	Verified beliefs	Information vs T and S
Sheshadri et al. 2026	3	S, scaffolded	Implanted behaviour	Information vs S
Zhong et al. 2026	3	S, elicitation	Implanted rule	Information vs S
Boxo et al. 2025	2, 3	none	Constructed, organisms	Dependence on text

Whether a probe trained on behaviour that is stated in text transfers to the same behaviour performed silently has been tested for three behaviours, two in fine-tuned model organisms and one as natural sycophancy, on 8B and 30B-A3B models (Boxo et al., 2025) and for sandbagging in an auditing game (Taylor et al., 2025). For the other text-ambiguous behaviours it is not known.

Whether the forecasting premium reported by Chan et al. (2025) survives an adversary that knows it is being probed is not known.

How much of the auditing advantage of black-box methods reflects the tools themselves rather than how well an agent uses them is not known; Sheshadri et al. (2026) report that standalone and agentic usefulness diverge.

14. The case against this analysis

The decomposition assumes an ideal transcript reader. In Case 1 the paper says no information premium is possible against an ideal reader of T. No such reader exists, and a probe can beat the readers that do exist, which is what deployment cares about. This is

correct, and it is why the paper calls the Case 1 premium a cost premium plus a difference in reader capacity rather than calling it nothing. The claim is narrower: such a premium does not show that internals carry information the text does not, and so does not support the argument that white-box monitoring survives a model that manages its outputs.

The inversion rests on a small and recent record. Many of the results in Sections 5 and 6 are 2025-2026 preprints, several from single authors or single groups (Kumar, 2026; Dingeto, 2026; Devbunova, 2026; Thormann, 2026; Manek, 2026), and model organisms are small and may not resemble a model that behaves strategically for its own reasons; Taylor et al. (2025) note that their sandbagging organisms may lack a robust self-conception of what they are doing. If the silent-behaviour transfer of study S2 turns out well for most behaviours, the inversion would describe a stage of the field rather than a property of probes. Against that, the generalisation successes the paper relies on for the other side of the inversion (McKenzie et al., 2025; Kramar et al., 2026) come from large, careful studies, and the

failures appear across independent groups and methods.

The label ceiling can be escaped by careful labelling. Human experts with access to private context, or labellers given the model's later behaviour, can produce labels that a transcript reader could not. Kramar et al. (2026) use expert-curated sets. The ceiling binds only when the labeller reads what the black-box monitor reads, and the paper claims no more than that. It does bind on several of the most cited positive results.

Auditing comparisons favour agents that use text well. The black-box advantage in Sheshadri et al. (2026) and Cywinski et al. (2025) is measured through an LLM auditor that reads tool outputs. White-box tools whose output is a list of features or tokens may be harder for such an auditor to use than a prefilled response is. The tool-to-agent gap Sheshadri et al. report is evidence of exactly this. A better interface could raise the white-box numbers without any change in what the activations contain.

The monitored model is not the only source of activations. Shiromani and Richter (2026) probe the monitor's activations, not the agent's. That is a different white-box channel, aimed at the monitor's failure to say what it registered rather than at the agent's concealment, and this paper's decomposition does not cover it beyond the behavioural control it borrows.

The self-directed caution is this. The claim on which this paper should be judged is the information-source decomposition and the protocol in Algorithm 1, which can be checked against any new comparison. The inversion is the claim most likely to be overtaken.

15. Conclusion

The question "do model internals beat behavioural monitors?" has three answers, one for each premium. On cost the answer is yes, by orders of magnitude, and production systems already exploit it in cascades. On robustness under an adversary the answer is unknown, because nobody has given one attacker one budget against both kinds of monitor. On information - the premium that would let a monitor catch what a model does not say - the answer depends on where the label lives. When it lives in the transcript, internals add nothing a good reader lacks. When it lives in withheld context or in text not yet written, internals have been shown to carry some of it. When it lives in the model itself, internals help in designs with known ground truth, and help less than simply asking the model again in the ways the auditing literature has developed.

The uncomfortable part is the inversion. Probes transfer best on targets the text already shows and worst on the intent-like targets for which white-box monitoring is proposed, and evaluations whose labels come from a transcript reader cannot show a premium on exactly the cases that would matter. Neither fact says internals are

useless. Together they say that most of the evidence for the information premium has yet to be collected, and Algorithm 1 sets out how to collect it.

References

- Abdelnabi, S., Fay, A., Cherubin, G., Salem, A., Fritz, M., & Pavard, A. (2024). Get my drift? Catching LLM Task Drift with Activation Deltas. IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) 2025. arXiv:2406.00799. doi:10.48550/arXiv.2406.00799.
- Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., & Conmy, A. (2025). Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. International Conference on Machine Learning (ICML) 2026. arXiv:2503.08679. doi:10.48550/arXiv.2503.08679.
- Arnav, B., Bernabeu-Perez, P., Helm-Burger, N., Kostolansky, T., Whittingham, H., & Phuong, M. (2025). CoT Red-Handed: Stress Testing Chain-of-Thought Monitoring. Advances in Neural Information Processing Systems (NeurIPS) 2025. arXiv:2505.23575. doi:10.48550/arXiv.2505.23575.
- Azaria, A., & Mitchell, T. (2023). The Internal State of an LLM Knows When It's Lying. arXiv:2304.13734. doi:10.48550/arXiv.2304.13734. [PREPRINT]
- Bailey, L., Serrano, A., Sheshadri, A., Seleznyov, M., Taylor, J., Jenner, E., et al. (2024). Obfuscated Activations Bypass LLM Latent-Space Defenses. arXiv:2412.09565. doi:10.48550/arXiv.2412.09565. [PREPRINT]
- Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., et al. (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv:2503.11926. doi:10.48550/arXiv.2503.11926. [PREPRINT]
- Belinkov, Y. (2021). Probing Classifiers: Promises, Shortcomings, and Advances. Computational Linguistics. arXiv:2102.12452. doi:10.48550/arXiv.2102.12452.
- Benton, J., Wagner, M., Christiansen, E., Anil, C., Perez, E., Srivastav, J., et al. (2024). Sabotage Evaluations for Frontier Models. arXiv:2410.21514. doi:10.48550/arXiv.2410.21514. [PREPRINT]
- Boxo, G., Neelappa, A., & Raval, S. (2025). Linear probes rely on textual evidence: Results from leakage mitigation studies in language models. arXiv:2509.21344. doi:10.48550/arXiv.2509.21344. [PREPRINT]
- Burns, C., Ye, H., Klein, D., & Steinhart, J. (2022). Discovering Latent Knowledge in Language Models Without Supervision. International Conference on Learning Representations (ICLR) 2023. arXiv:2212.03827. doi:10.48550/arXiv.2212.03827.
- Chan, Y. S., Yong, Z., & Bach, S. H. (2025). Can We Predict Alignment Before Models Finish Thinking? Towards Monitoring Misaligned Reasoning Models. arXiv:2507.12428. doi:10.48550/arXiv.2507.12428. [PREPRINT]
- Chaudhary, M., Su, I., Hooda, N., Shankar, N., Tan, J., Zhu, K., et al. (2025). Evaluation Awareness Scales Predictably in Open-Weights Large Language Models. arXiv:2509.13333. doi:10.48550/arXiv.2509.13333. [PREPRINT]
- Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., et al. (2025). Reasoning Models Don't Always Say What They Think. arXiv:2505.05410. doi:10.48550/arXiv.2505.05410. [PREPRINT]
- Cundy, C., & Gleave, A. (2025). Preference Learning with Lie Detectors can Induce Honesty or Evasion. Advances in Neural Information Processing Systems (NeurIPS) 2025. arXiv:2505.13787. doi:10.48550/arXiv.2505.13787.
- Cunningham, H., Wei, J., Wang, Z., Persic, A., Peng, A., Abderrachid, J., et al. (2026). Constitutional Classifiers + +: Efficient Production-Grade Defenses against Universal Jailbreaks. arXiv:2601.04603. doi:10.48550/arXiv.2601.04603. [PREPRINT]

- Cywinski, B., Ryd, E., Wang, R., Rajamanoharan, S., Nanda, N., Conmy, A., et al. (2025). Eliciting Secret Knowledge from Language Models. arXiv:2510.01070. doi:10.48550/arXiv.2510.01070. [PREPRINT]
- Daniels, O., Moodley, P., Marlin, B. M., & Lindner, D. (2026). Stress-Testing Alignment Audits With Prompt-Level Strategic Deception. ICLR 2026 Workshop on Principled Design for Trustworthy AI. arXiv:2602.08877. doi:10.48550/arXiv.2602.08877. [WORKSHOP]
- David, E., & Fomin, M. (2026). Prompted to Discriminate: Generalizing Malicious-Input Probes in the Wild. NeurIPS 2026 Workshop on Agents in the Wild. arXiv:2610.02413. doi:10.48550/arXiv.2610.02413. [WORKSHOP]
- Devbunova, V. (2026). Is Evaluation Awareness Just Format Sensitivity? Limitations of Probe-Based Evidence under Controlled Prompt Structure. ICLR 2026 Workshop I Can't Believe It's Not Better. arXiv:2603.19426. doi:10.48550/arXiv.2603.19426. [WORKSHOP]
- Dingeto, H. (2026). A Lie Detector Test for Language Models: Reading Knowledge a Model Won't Reveal. arXiv:2609.21996. doi:10.48550/arXiv.2609.21996. [PREPRINT]
- Duan, E. (2026). Do Activation Monitors Survive Model Updates? Benchmarking, Predicting, and Repairing Activation-Monitor Staleness. arXiv:2606.15980. doi:10.48550/arXiv.2606.15980. [PREPRINT]
- Dubanowska, Z., Zelazczyk, M., Brzozowski, M., Mandica, P., & Karpowicz, M. (2025). Representation-based Broad Hallucination Detectors Fail to Generalize Out of Distribution. Findings of the Association for Computational Linguistics: EMNLP 2025. arXiv:2509.19372. doi:10.48550/arXiv.2509.19372.
- Emmons, S., Jenner, E., Elson, D. K., Saurous, R. A., Rajamanoharan, S., Chen, H., et al. (2025). When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors. arXiv:2507.05246. doi:10.48550/arXiv.2507.05246. [PREPRINT]
- Farquhar, S., Varma, V., Kenton, Z., Gasteiger, J., Mikulik, V., & Shah, R. (2023). Challenges with unsupervised LLM knowledge discovery. arXiv:2312.10029. doi:10.48550/arXiv.2312.10029. [PREPRINT]
- Goldowsky-Dill, N., Chughtai, B., Heimersheim, S., & Hobbhahn, M. (2025). Detecting Strategic Deception Using Linear Probes. arXiv:2502.03407. doi:10.48550/arXiv.2502.03407. [PREPRINT]
- Golechha, S., & Garriga-Alonso, A. (2025). Among Us: A Sandbox for Measuring and Detecting Agentic Deception. Advances in Neural Information Processing Systems (NeurIPS) 2025. arXiv:2504.04072. doi:10.48550/arXiv.2504.04072.
- Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., et al. (2024). Alignment faking in large language models. arXiv:2412.14093. doi:10.48550/arXiv.2412.14093. [PREPRINT]
- Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2023). AI Control: Improving Safety Despite Intentional Subversion. International Conference on Machine Learning (ICML) 2024. arXiv:2312.06942. doi:10.48550/arXiv.2312.06942.
- Gupta, R., & Jenner, E. (2025). RL-Obfuscation: Can Language Models Learn to Evade Latent-Space Monitors? arXiv:2506.14261. doi:10.48550/arXiv.2506.14261. [PREPRINT]
- Heidari, F., Memarian, A., & Rabusseau, G. (2026). Evaluation Awareness in Language Models: Representation, Verbalization, and Control. arXiv:2608.21766. doi:10.48550/arXiv.2608.21766. [PREPRINT]
- Hewitt, J., & Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019. arXiv:1909.03368. doi:10.48550/arXiv.1909.03368.
- Hollinsworth, O. J., Dombrowski, A., Adam-Day, S., Gleave, A., & Cundy, C. (2026). Scaling Trends for Lie Detector Oversight in Preference Learning. arXiv:2607.01567. doi:10.48550/arXiv.2607.01567. [PREPRINT]
- Hua, T. T., Baskerville, J., Lemoine, H., Hopman, M., Bhatt, A., & Tracy, T. (2025). Combining Cost-Constrained Runtime Monitors for AI Safety. arXiv:2507.15886. doi:10.48550/arXiv.2507.15886. [PREPRINT]
- Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566. doi:10.48550/arXiv.2401.05566. [PREPRINT]
- Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., et al. (2023). Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674. doi:10.48550/arXiv.2312.06674. [PREPRINT]
- Johnston, D. O., Chakraborty, A., & Belrose, N. (2025). Mechanistic Anomaly Detection for "Quirky" Language Models. ICLR 2025 Building Trust Workshop. arXiv:2504.08812. doi:10.48550/arXiv.2504.08812. [WORKSHOP]
- Kale, N., Zhang, C. B. C., Zhu, K., Aich, A., Rodriguez, P., Scale Red Team, et al. (2025). Reliable Weak-to-Strong Monitoring of LLM Agents. arXiv:2508.19461. doi:10.48550/arXiv.2508.19461. [PREPRINT]
- Kantamneni, S., Engels, J., Rajamanoharan, S., Tegmark, M., & Nanda, N. (2025). Are Sparse Autoencoders Useful? A Case Study in Sparse Probing. arXiv:2502.16681. doi:10.48550/arXiv.2502.16681. [PREPRINT]
- Kirch, N., Dower, S., Skapars, A., Yannakoudakis, H., Lubana, E. S., & Krashennikov, D. (2025). The Impact of Off-Policy Training Data on Probe Generalisation. Annual Meeting of the Association for Computational Linguistics (ACL) 2026. arXiv:2511.17408. doi:10.48550/arXiv.2511.17408.
- Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., et al. (2025). Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. arXiv:2507.11473. doi:10.48550/arXiv.2507.11473. [PREPRINT]
- Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., & Gal, Y. (2024). Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. arXiv:2406.15927. doi:10.48550/arXiv.2406.15927. [PREPRINT]
- Kramar, J., Engels, J., Wang, Z., Chughtai, B., Shah, R., Nanda, N., et al. (2026). Building Production-Ready Probes For Gemini. arXiv:2601.11516. doi:10.48550/arXiv.2601.11516. [PREPRINT]
- Kretschmar, K., Laurito, W., Maiya, S., & Marks, S. (2025). Liars' Bench: Evaluating Lie Detectors for Language Models. arXiv:2511.16035. doi:10.48550/arXiv.2511.16035. [PREPRINT]
- Kumar, S. (2026). Pressure-Testing Deception Probes in LLMs: Scaling, Robustness, and the Geometry of Deceptive Representations. GEM Workshop at ACL 2026. arXiv:2605.27958. doi:10.48550/arXiv.2605.27958. [WORKSHOP]
- Kutasov, J., Sun, Y., Colognese, P., van der Weij, T., Petrini, L., Zhang, C. B. C., et al. (2025). SHADE-Arena: Evaluating Sabotage and Monitoring in LLM Agents. arXiv:2506.15740. doi:10.48550/arXiv.2506.15740. [PREPRINT]
- Levinstein, B. A., & Herrmann, D. A. (2024). Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks. Philosophical Studies. doi:10.1007/s11098-023-02094-3. arXiv:2307.00175. doi:10.48550/arXiv.2307.00175.
- Lyons Keenan, H., Leckie, C., & Erfani, S. (2026). Mechanistic Anomaly Detection via Functional Attribution. International Conference on Machine Learning (ICML) 2026. arXiv:2604.18970. doi:10.48550/arXiv.2604.18970.
- Mahadik, P., & Skapars, A. (2026). Do Linear Probes Generalize Better in Persona Coordinates? arXiv:2605.09391. doi:10.48550/arXiv.2605.09391. [PREPRINT]
- Mahendrakar, P. (2026). A Second Model Is Not a Second Opinion: Peer Verification, Correlated Failure, and the Independence Variable Nobody Reports. Life of Research. doi:10.5281/zenodo.22044420.
- Mallen, A., Brumley, M., Kharchenko, J., & Belrose, N. (2023). Eliciting Latent Knowledge from Quirky Language Models. Conference on Language Modeling (COLM) 2024.

- arXiv:2312.01037. doi:10.48550/arXiv.2312.01037.
- Manek, A. (2026). Representational Depth of Evaluation Awareness Shifts With Scale in Open-Weight Language Models. ICML 2026 Mechanistic Interpretability Workshop. arXiv:2606.29196. doi:10.48550/arXiv.2606.29196. [WORKSHOP]
- Marks, S., & Tegmark, M. (2023). The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. Conference on Language Modeling (COLM) 2024. arXiv:2310.06824. doi:10.48550/arXiv.2310.06824.
- Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., et al. (2025). Auditing language models for hidden objectives. arXiv:2503.10965. doi:10.48550/arXiv.2503.10965. [PREPRINT]
- McGuinness, M., Serrano, A., Bailey, L., & Emmons, S. (2025). Neural Chameleons: Language Models Can Learn to Hide Their Thoughts from Unseen Activation Monitors. arXiv:2512.11949. doi:10.48550/arXiv.2512.11949. [PREPRINT]
- McKenzie, A., Pawar, U., Blandfort, P., Bankes, W., Krueger, D., Lubana, E. S., & Krasheninnikov, D. (2025). Detecting High-Stakes Interactions with Activation Probes. Advances in Neural Information Processing Systems (NeurIPS) 2025. arXiv:2506.10805. doi:10.48550/arXiv.2506.10805.
- Moustafa, A., Feser, M., & Mai, F. (2026). Beyond Liars' Bench: The Impact of Lie Typology, Depth, and Sparsity on Deception Detection in LLMs. arXiv:2607.20479. doi:10.48550/arXiv.2607.20479. [PREPRINT]
- Natarajan, V., Jain, D., Arora, S., Golechha, S., & Bloom, J. (2026). One Probe Won't Catch Them All: Towards Targeted Deception Detection. International Conference on Machine Learning (ICML) 2026, PMLR 306. arXiv:2602.01425. doi:10.48550/arXiv.2602.01425.
- Nguyen, J., Hoang, K., Attubato, C. L., & Hofstadter, F. (2025). Probing and Steering Evaluation Awareness of Language Models. ICML 2025 Actionable Interpretability Workshop. arXiv:2507.01786. doi:10.48550/arXiv.2507.01786. [WORKSHOP]
- Oldfield, J., Torr, P., Patras, I., Bibi, A., & Barez, F. (2025). Beyond Linear Probes: Dynamic Safety Monitoring for Language Models. International Conference on Learning Representations (ICLR) 2026. arXiv:2509.26238. doi:10.48550/arXiv.2509.26238.
- Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., et al. (2024). LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. arXiv:2410.02707. doi:10.48550/arXiv.2410.02707. [PREPRINT]
- Pacchiardi, L., Chan, A. J., Mindermann, S., Moscovitz, I., Pan, A. Y., Gal, Y., et al. (2023). How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions. arXiv:2309.15840. doi:10.48550/arXiv.2309.15840. [PREPRINT]
- Parrack, A., Attubato, C. L., & Heimersheim, S. (2025). Benchmarking Deception Probes via Black-to-White Performance Boosts. arXiv:2507.12691. doi:10.48550/arXiv.2507.12691. [PREPRINT]
- Sadhu, S., Sengupta, A., Sankarapu, V. K., & Seth, P. (2026). What Do Compliance Detectors Read? An Audit of Activation Probes and Guard Models. arXiv:2608.16852. doi:10.48550/arXiv.2608.16852. [PREPRINT]
- Scheurer, J., Balesni, M., & Hobbhahn, M. (2023). Large Language Models can Strategically Deceive their Users when Put Under Pressure. arXiv:2311.07590. doi:10.48550/arXiv.2311.07590. [PREPRINT]
- Sharma, M., Tong, M., Mu, J., Wei, J., Kruthoff, J., Goodfriend, S., et al. (2025). Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. arXiv:2501.18837. doi:10.48550/arXiv.2501.18837. [PREPRINT]
- Sheshadri, A., Ewart, A., Kempf, E., Fronsdal, K., Gupta, I., Schrod, S., et al. (2026). AuditBench: Evaluating Alignment Auditing Techniques on Models with Hidden Behaviors. arXiv:2602.22755. doi:10.48550/arXiv.2602.22755. [PREPRINT]
- Shiromani, S., & Richter, L. (2026). A False Average: Pooled CoT-Monitor Accuracy Conceals a Reasoning-Dependent Fragility. arXiv:2608.00583. doi:10.48550/arXiv.2608.00583. [PREPRINT]
- Singh, N., & Chaudhary, M. (2026). Evaluation Awareness Shifts from Format to Context with Model Scale. arXiv:2609.22119. doi:10.48550/arXiv.2609.22119. [PREPRINT]
- Taufeeque, M., Heimersheim, S., Gleave, A., & Cundy, C. (2026). The Obfuscation Atlas: Mapping Where Honesty Emerges in RLVR with Deception Probes. International Conference on Machine Learning (ICML) 2026. arXiv:2602.15515. doi:10.48550/arXiv.2602.15515.
- Taylor, J., Black, S., Bowen, D., Read, T., Golechha, S., Zelenka-Martin, A., et al. (2025). Auditing Games for Sandbagging. arXiv:2512.07810. doi:10.48550/arXiv.2512.07810. [PREPRINT]
- Thormann, T. (2026). Probing the Limits of the Lie Detector Approach to LLM Deception. arXiv:2603.10003. doi:10.48550/arXiv.2603.10003. [PREPRINT]
- Tice, C., Kreer, P. A., Helm-Burger, N., Shahani, P. S., Ryzhenkov, F., Roger, F., et al. (2024). Noise Injection Reveals Hidden Capabilities of Sandbagging Language Models. Advances in Neural Information Processing Systems (NeurIPS) 2025. arXiv:2412.01784. doi:10.48550/arXiv.2412.01784.
- Tillman, H., & Mossing, D. (2025). Investigating task-specific prompts and sparse autoencoders for activation monitoring. arXiv:2504.20271. doi:10.48550/arXiv.2504.20271. [PREPRINT]
- Yoo, D., & Skapars, A. (2026). Probe Generalization as Subspace Selection for OOD Deception Detection. arXiv:2609.02893. doi:10.48550/arXiv.2609.02893. [PREPRINT]
- Zhong, Z., Muhamed, A., Diab, M. T., Smith, V., & Raghunathan, A. (2026). Pando: Do Interpretability Methods Work When Models Won't Explain Themselves? arXiv:2604.11061. doi:10.48550/arXiv.2604.11061. [PREPRINT]
- Zhou, K., Venhoff, C., Michala, J., Wang, X. E., & Saunders, W. (2026). Probing the Mismatched Thinking Process of Language Models. arXiv:2606.24251. doi:10.48550/arXiv.2606.24251. [PREPRINT]
- Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405. doi:10.48550/arXiv.2310.01405. [PREPRINT]