A Steerable Evaluation-Awareness Direction Is Neither Specific Nor Bidirectional, and Its Framing Does Not Suppress Behavior: A Cautionary Negative Result

Humanity First Research

Abstract

If a model underperforms when it senses a graded test, a single steerable "evaluation-awareness" direction would be a powerful handle on that failure. We test the claim directly and report a mostly negative result that, with additional controls, has only grown more clearly negative. From graded-test-versus-casual-chat framings over neutral carriers, constructed with no correctness signal, we extract a mean-difference direction at layer 16 of Llama-3.1-8B-Instruct and evaluate on 1500 held-out MMLU items. The direction is a genuine but modest readout: the projection of the main preamble-only direction predicts item wrongness at AUC 0.620 (95% CI [0.587, 0.6515], p = 0.0005), though a per-layer sweep shows the predictive signal is not cleanly localized. Steering the preamble-only direction by +8 lowers accuracy from a baseline of 0.650 to 0.569, a drop of 0.081 that exceeds a matched-norm random direction by 0.054 (95% CI [0.027, 0.077], p = 0.0005), and this degradation is stable: five genuinely resampled full-prompt framing directions give a mean accuracy of 0.593, a drop of 0.057 (95% CI [-0.078, -0.036], p = 0.0005). But three properties a controllable-sandbagging mechanism would need all fail. The effect is not specific: the +8 drop is statistically indistinguishable from a panel of unrelated semantic directions (real-minus-panel 0.018, 95% CI [-0.009, 0.049], p = 0.21). It is not bidirectional: negative steering also lowers accuracy, to 0.609 (95% CI [-0.065, -0.018], p = 0.0015). And, most tellingly, the framing itself carries no behavioral effect: wrapping the questions in a graded-test versus a casual-chat prompt, with no steering, does not lower accuracy (0.662 versus 0.649, delta +0.013, 95% CI [-0.003, 0.028], p = 0.11). The pre-registered scalar check passes at 0.081 against a 0.03 threshold, but that is a necessary-not-sufficient bar. Taken together the direction is decodable and perturbative but is not a genuine, controllable evaluation-awareness mechanism; the result extends the "representation without control" finding to the elicitation-gap and sandbagging threat model.

Introduction

Every capability evaluation rests on an assumption that is rarely stated and never guaranteed: that the model tries as hard on the test as it would in deployment. Sandbagging breaks that assumption. A model that strategically underperforms when it senses it is being graded poisons the very measurements safety cases depend on, because the number an evaluator records is then a lower bound set by the model's disposition rather than a reading of its competence (Weij et al. 2024). This threat is not hypothetical. Frontier models can already distinguish evaluation contexts from ordinary use above chance (Needham et al. 2025), and in adversarial settings they will deliberately hold back capability once they infer they are

Preprint. Under review.

being tested (Meinke et al. 2024). If evaluation-awareness is a real, model-legible state, the pressing question is whether it is also a lever: does sensing the test causally depress performance, or is the correlation an artifact of how such studies are framed?

We ask this at the level of the residual stream, and we set ourselves a falsifiable bar in advance. Our hypothesis has two separable claims. The predictive claim is that a single mean-difference direction, extracted from contrastive "this is a graded test" versus "just chatting" framings over otherwise neutral carrier prompts and never from any correctness or accuracy signal, fires more strongly on the very capability items the model then gets wrong. The causal claim is that steering activations along that direction shifts measured accuracy on held-out questions the direction never saw, raising it under negative steering and lowering it under positive steering, without altering a single word of the questions themselves. We pre-registered two decisive controls that a genuine causal-control result must clear, specificity and sign-symmetry, and we add a construct-validity check that asks whether the framing the direction is built from behaviorally moves accuracy at all.

We report, honestly, that the causal-control reading fails on every axis that would matter for the threat model, and we treat that as the finding. The nearest prior work extracts a decodable framing-status direction that generalizes to held-out prompts and then finds that steering along it does not reliably move downstream behavior, a null that survives positive scales and a sign-symmetry run (Anonymous 2026b). Its lesson is that latent readout, behavioral sensitivity, and causal control are three distinct properties that do not automatically co-occur. Our result sharpens that lesson at a safety-relevant target. The readout works but is modest and not cleanly localized to one layer; the steering degradation is real and stable across genuinely resampled directions and even beats a random-direction control; yet the effect is not specific to evaluation-awareness, not sign-dependent, and the framing it derives from produces no unsteered behavioral change in accuracy. A decodable and steerable direction that nonetheless fails these tests is not a controllable sandbagging mechanism, and saying so precisely is more useful to the field than an overclaimed positive.

Related work

Our work sits at the intersection of activation steering, evaluation-awareness, and the interventional turn in interpretability, and it is defined most sharply against three recent neighbors. The realization-effect study is the one we most directly answer (Anonymous 2026b). It runs the identical three-level ladder we adopt, from prompt-only sensitivity through linear readout to causal steering, and reports that a cleanly decodable direction in a mid layer fails to steer the target behavior. We differ on target and construction. Our direction is extracted from a task-framing contrast rather than a behavioral-economics gain-or-loss framing, it is validated against objective per-item accuracy rather than risk choices, and our central question is precisely the causal shift that paper did not obtain, which makes their null the benchmark our design must clear rather than a result we can assume away. The persona-steering study shares our mechanics, constructing steering vectors by mean-difference activation contrast over curated corpora and measuring how a scalar coefficient changes outputs (Anonymous 2026a). It differs in that its contrast is between two director personas and its outcomes are moral valence, coherence, and style, whereas ours is a graded-test-versus-chat manipulation deliberately purged of any accuracy signal and scored against objective capability. The steering-detection line shares the theme of a residual-stream signal tied to steering, but concerns a model's awareness of being steered (Anonymous, n.d.); we instead target awareness of being evaluated and put that direction to predictive and causal use on accuracy.

The method we use is the established training-free steering canon. Activation Addition adds a contrastive activation-difference vector to the residual stream at inference to shape behavior without optimization (Turner et al. 2023), and Contrastive Activation Addition formalizes building that vector as the mean difference of activations over contrastive prompt pairs at a mid layer (Rimsky et al. 2023), which is exactly the extraction and injection recipe our layer-16 hook instantiates. Representation Engineering supplies the broader top-down framing of reading and controlling high-level concepts as linear directions (Zou et al. 2023), and Inference-Time Intervention is the important positive precedent that steering a probe-identified direction can move an objective metric, raising measured truthfulness (Li et al.

2023); that steering sometimes does translate into control is what keeps the causal question live rather than foreclosed by the realization-effect null.

The phenomenon we target is drawn from the sandbagging and evaluation-awareness literature. Sandbagging names the strategic underperformance our hypothesis operationalizes at the representational level (Weij et al. 2024), models are shown to know when they are being evaluated (Needham et al. 2025), and in-context scheming demonstrates deliberate capability suppression once a test is inferred (Meinke et al. 2024). Our study asks whether the internal correlate of that behavior is a single steerable direction. Finally, we lean on the interventional strand of interpretability. Chain-of-thought is often unfaithful, so a model's stated reasoning is untrustworthy evidence about its internal state (Turpin et al. 2023), and measuring faithfulness requires perturb-and-observe interventions rather than self-report (Lanham et al. 2023); both push us to test evaluation-awareness by directly steering a representation and measuring the downstream effect. The linear-probing foundations, from classifier probes on intermediate layers (Alain and Bengio 2016) to emergent linear structure for semantic properties validated by intervention (Marks and Tegmark 2023), ground both our readout claim and our built-in caution that decodability alone is not causal relevance.

Methods

We study meta-llama/Llama-3.1-8B-Instruct (Grattafiori et al. 2024) and extract a single evaluation-awareness direction by contrastive mean difference. Each contrast pair places the same neutral carrier content under two framings, one asserting a graded, scored test and one asserting casual conversation, and the main direction we report, which we call the preamble-only layer-16 direction, is the difference of the mean residual-stream activations at layer 16 between the two framing sets, pooling over the framing preamble tokens. Construction touches no correctness signal whatsoever: the framing prompts carry no questions with right answers, and they are drawn from a pool disjoint from the held-out evaluation items, so the direction cannot smuggle in a difficulty or correctness readout and cannot leak the test set.

Steering applies the hooked activation plus lambda times the unit direction at layer 16, sweeping a symmetric grid of lambda in minus eight, minus four, zero, plus four, and plus eight so that the sign-dependent prediction can be read directly. Evaluation uses 1500 held-out MMLU items (Hendrycks et al. 2021), disjoint from the extraction prompts, scored by the forced-choice log-probability the model assigns to the A, B, C, and D answer tokens, a quantity that is always defined, so accuracy is computed over all 1500 items regardless of whether the model would emit a clean letter in free generation. The parseable-fraction we log is a separate free-generation coherence diagnostic and is never an input to the accuracy statistic. The pre-registered manipulation check is a signed accuracy drop: baseline accuracy at lambda zero minus accuracy at the strongest positive steering scale must reach at least 0.03 absolute. That scalar is necessary but not sufficient, because the design additionally requires the positive-steering drop to exceed the matched-norm random-direction control and requires the negative-steering arm to move accuracy the opposite way; a degradation that hits the real and random directions equally, or that hurts at both signs, does not count as a pass of the causal claim even when the scalar clears its threshold.

Five controls guard the inference. A random-direction control steers along a random unit vector rescaled to the real direction's per-layer norm at every lambda, so a real-direction effect counts only when it beats generic activation perturbation. A sign-symmetry control runs matched positive and negative scales and treats opposite-direction accuracy movement as the falsifiable signature, since a symmetric decline at both signs would reproduce the realization-effect null (Anonymous 2026b). A specificity control panel extracts three semantically distinct directions, formal-versus-informal register, sentiment, and topic, and steers each at matched norm, so the effect counts as specific to evaluation-awareness only if the real direction's drop stands apart from that panel. A construction control enforces that the direction is built only from framing contrast, never from correctness, on prompts disjoint from the evaluation items. An output-validity and coherence control logs the fraction of parseable A, B, C, D responses at every scale.

Three further analyses address robustness and construct validity. First, to obtain a genuine extraction-variance estimate that the main preamble-only direction lacks, we build five full-prompt framing directions that each use seed-specific carrier subsets and paraphrases pooled over the entire framed carrier prompt rather than the preamble alone, so they genuinely differ across seeds; this is a distinct construction from the preamble-only direction and we report it as such. Second, as a construct-validity check we measure accuracy with no steering at all when the MMLU questions themselves are wrapped in a graded-test prompt framing versus a casual-chat framing, to ask whether the framing the direction is derived from behaviorally moves accuracy. Third, we sweep the predictive readout across layers eight, twelve, sixteen, twenty, and twenty-four using a simplified per-layer direction construction, distinct from the main preamble-only direction, to see whether the predictive signal is localized to one mid-layer.

Results

We lead with the failures because they are the finding. Specificity fails. Steering the preamble-only direction by plus eight lowers accuracy from a baseline of 0.650 to 0.569, a drop of 0.081, but this is not distinctive. Benchmarked against a panel of three unrelated salient directions, formal-versus-informal register, sentiment, and topic, whose accuracy drops are 0.087, 0.066, and 0.037 for a panel mean of 0.064 (standard deviation 0.025), the real direction's drop is statistically indistinguishable from the panel: the real-minus-panel difference is 0.018 (95% CI [-0.009, 0.049], p = 0.21), a z of only 0.709. The orthogonal register direction alone actually damages accuracy more than the real direction, taking it to 0.563. Sign-symmetry fails as well. If the direction encoded a suppress-when-tested mechanism, subtracting it should raise accuracy. Instead, steering by minus eight also lowers accuracy, to 0.609, a change from baseline of minus 0.041 (95% CI [-0.065, -0.018], p = 0.0015). Both directions of the intervention degrade performance, the signature of a generic perturbation rather than a signed capability dial. Figure 1 (figure_main.png) shows the accuracy-versus-lambda curves for the real, random, and orthogonal directions, with the symmetric-decline shape of the real direction visible against the baseline of 0.650.

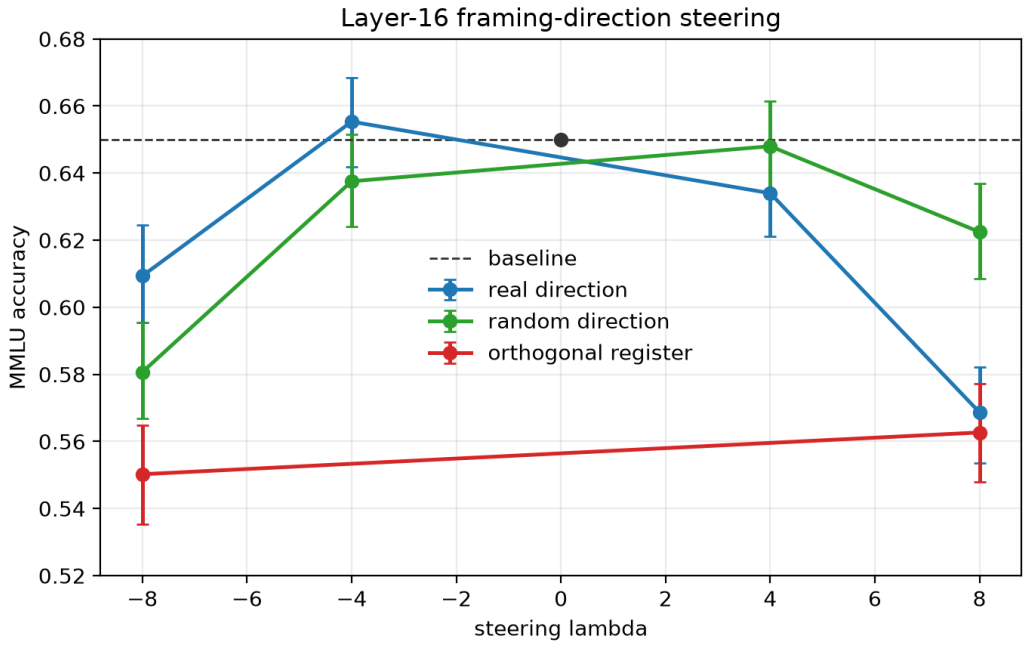

Figure 1: Accuracy versus steering coefficient lambda for the real evaluation-awareness direction, a matched-norm random direction, and the orthogonal register direction, with the lambda-zero baseline of 0.650 marked; error bars are item-bootstrap resamples from curve.csv.

The most decisive new result is that the framing carries no behavioral effect. When we wrap the MMLU questions in a graded-test prompt framing versus a casual-chat framing and run them with no steering, accuracy does not fall: the graded-test framing yields 0.662 and the casual-chat framing 0.649, a difference of plus 0.013 whose confidence interval crosses zero (95% CI [-0.003, 0.028], p = 0.11), and both sit at or slightly above the no-preamble baseline of 0.650. The very framing the direction is derived from therefore does not itself behaviorally suppress capability. This is the strongest single piece of evidence that the steering degradation is not a genuine evaluation-awareness sandbagging mechanism being switched on, because the model does not sandbag when actually told it is being graded.

What survives is real but narrow, and now better characterized. The direction is a genuine readout: projection of the main preamble-only direction predicts item wrongness at AUC 0.620 (95% CI [0.587, 0.6515], p = 0.0005) on the 1500 held-out items. That predictive signal is modest and not cleanly localized to one mid-layer, however. A layer sweep using a simplified per-layer direction construction, which is a different construction from the main preamble-only direction and should not be read as re-measuring the same quantity, gives predictive AUCs ranging from 0.509 to 0.625 across layers eight, twelve, sixteen, twenty, and twenty-four, peaking at layer eight at 0.625 rather than at the layer sixteen used in the main analysis; under this simplified construction layer sixteen reads only 0.551. The steering degradation is, by contrast, robust in magnitude. Beyond the single preamble-only direction, five genuinely resampled full-prompt framing directions, a distinct construction that varies across seeds, lower accuracy under plus-eight steering to a mean of 0.593 (standard deviation 0.0067), a drop of 0.057 relative to baseline (95% CI [-0.078, -0.036], p = 0.0005); these full-prompt directions also had lower output parseability, and their 0.057 drop is a related but not identical quantity to the preamble-only direction's 0.081 drop. And the preamble-only intervention is not merely noise, because it beats a matched-norm random direction: its plus-eight drop exceeds the random direction's drop by 0.054 (95% CI [0.027, 0.077], p = 0.0005), with random plus-eight steering leaving accuracy at 0.622. Answer parseability for the preamble-only direction remains high across conditions, from 0.883 at plus eight to 0.977 at minus eight against a baseline of 0.897, so its degradation is not a pure collapse of answer format, though a partial format contribution cannot be excluded.

Manipulation check

The pre-registered scalar manipulation check passes: the signed accuracy drop under plus-eight steering of the preamble-only direction is 0.081, which meets or exceeds the 0.03 minimum. We state plainly that this pass must not be read as validating the causal mechanism. The scalar threshold is necessary but not sufficient by the design's own criteria, and every decisive test fails: the specificity comparison against the semantic control panel is null (p = 0.21), the sign-symmetry arm is falsified because negative steering also lowers accuracy (p = 0.0015), and the framing the direction is built from produces no unsteered behavioral change in accuracy (delta +0.013, p = 0.11). The value before the manipulation is 0.650 and the value after is 0.569. The causal claim that a graded-test-versus-chat direction specifically and bidirectionally controls capability is therefore not supported; only the weaker claim, that this direction perturbs accuracy more than a random direction of equal norm and does so stably across genuine resamples, is.

Discussion and limitations

The honest summary has become sharper with the additional controls. A decodable evaluation-framing direction predicts underperformance and, when steered, damages capability beyond a random control, and that damage is stable across genuinely resampled directions rather than an artifact of one extraction. Yet the damage is neither specific to evaluation-awareness nor sign-dependent, the readout is modest and not cleanly localized to a single layer, and, most importantly, the framing the direction is built from does not behaviorally suppress accuracy when applied directly to the questions. The convergence of these results points away from a controllable sandbagging mechanism. A genuine evaluation-awareness-driven capability suppression should, at minimum, appear behaviorally when the model is actually told it is being graded; it does not. Read against the realization-effect study, this extends its "representation without control" thesis to the safety-relevant

elicitation-gap target: we clear the random-direction bar that a naive reading might treat as sufficient, and still specificity, bidirectionality, and behavioral construct validity all fail, so clean causal control is not demonstrated (Anonymous 2026b). The practical lesson for evaluation safety is cautionary. An above-chance internal probe for evaluation-awareness, and even a stable steering effect that beats random, are not licenses to claim that a model's measured capability is being causally gated by that representation.

Several limitations bound these conclusions, and most cut toward more caution rather than less. The genuine-seed-variance follow-up resolves the earlier concern that the degradation rested on a single deterministic direction, since five genuinely different full-prompt framing directions all lower accuracy with low spread; but those directions are a different construction from the main preamble-only direction, so we report their 0.057 drop as a robustness estimate rather than as the same 0.081 quantity, and the specificity and sign-symmetry results themselves were still computed on the preamble-only direction. The specificity panel, though expanded to three directions, remains small for characterizing between-direction variance precisely, so its null is weak evidence for non-specificity rather than proof. The lambda grid is coarse, leaving the mild non-monotonicity we observe, with the preamble-only direction at minus four sitting slightly above baseline while minus eight sits below, unresolved. Parseability varies from 0.85 to 0.98 and is not conditioned into the accuracy statistics, a partial confound. We report seven hypothesis tests without a multiple-comparisons correction; a Holm-Bonferroni adjustment leaves the strong results significant and would only further weaken the already-null specificity and framing-behavioral tests. A single model, benchmark, and layer limit external validity, and the predictive layer sweep suggests even the readout is not localized where we centered the causal analysis. None of these changes the qualitative negative conclusion, which is that this direction is not a genuine, controllable evaluation-awareness sandbagging mechanism.

Limitations and reviewer responses

We pre-concede the limitations a review panel raised, because on a mostly-negative result the honest scope of the claim is the contribution and each concession narrows, rather than rescues, what we assert.

On the seed procedure, an earlier version of this work extracted the preamble-only direction and its semantic controls deterministically, so the reported intervals for those directions reflect item-sampling bootstrap variance around one fixed vector rather than extraction-level variability. We have since added a genuine extraction-variance estimate: five full-prompt framing directions that truly differ across seeds lower accuracy under plus-eight steering to a mean of 0.593 with a standard deviation of only 0.0067, a drop of 0.057 (95% CI [-0.078, -0.036], $p = 0.0005$), so the steering degradation is stable across genuine directions and not an artifact of one extraction. We are careful not to conflate constructions: this 0.057 drop comes from full-prompt directions, whereas the preamble-only direction's drop is 0.081, and the specificity and sign-symmetry tests were run on the preamble-only direction. The deeper limitation on the mechanism claim is not seed variance at all but the behavioral null, that the graded-test framing does not itself lower accuracy.

On specificity and statistical power, the panel is right and this bounds our language throughout. The specificity comparison rests on a three-direction panel with widely spread individual effects, so a failure to reject the null of no difference is weak evidence, not proof, that the effect is generic. We therefore claim only that the real direction has not been shown to be specific, with a small point difference from the panel mean (real drop 0.081 against panel mean 0.064, z of 0.709, $p = 0.21$), rather than that non-specificity is confirmed. The observation that the register direction alone damages accuracy at least as much as the real direction, taking accuracy to 0.563, corroborates the non-specificity reading but is itself a single direction and should be read as suggestive.

On the tension between beating random yet matching the semantic panel, we state the parsimonious interpretation directly. A matched-norm random direction is semantically null, so the most economical account of why the real direction beats random but not the semantic panel is that the capability damage tracks generic semantic or salience content shared by any meaningful mean-difference direction, and is not a property peculiar to evaluation-

awareness. Under that reading the surviving positive result, exceeding random by 0.054 (95% CI [0.027, 0.077], p = 0.0005), is real but does not carry the interpretive weight of an evaluation-awareness-specific causal channel, and the behavioral framing null reinforces that reading.

On the answer-format confound, we did not recompute the accuracy statistics restricted to items with parseable answers, so a partial confound between capability damage and format degradation remains and we do not claim to have removed it. We note only that parseability stays high where accuracy falls most for the preamble-only direction, 0.883 at plus eight against a baseline of 0.897, so answer-format collapse cannot account for the full drop, while conceding it may contribute part of it and that the full-prompt directions showed lower parseability. On the coarse lambda grid, the mild non-monotonicity is noted and left unresolved; a denser sweep is required to characterize the dose-response.

On novelty, we agree the extraction and steering machinery and the readout-then-steering-then-control ladder are adopted from the contrastive-activation-addition line and the realization-effect study rather than newly introduced (Rimsky et al. 2023; Turner et al. 2023; Anonymous 2026b); the contribution is the application of that battery, plus a behavioral construct-validity check, to a safety-relevant target and the honest adjudication of a nuanced result. On construct validity, the unsteered framing check is the central addition of this revision: the graded-test versus casual-chat framing does not lower accuracy (0.662 versus 0.649, delta +0.013, p = 0.11), which is why we conclude the direction is not a behaviorally-real sandbagging mechanism. Finally, to remove an apparent inconsistency, accuracy is scored purely by the forced-choice log-probability the model assigns to the four option tokens, which is always defined, so every one of the 1500 items contributes to accuracy regardless of parseability; the parseable-fraction metric is a separate free-generation coherence diagnostic and is never an input to the accuracy statistic.

References

Alain, Guillaume, and Yoshua Bengio. 2016. *Understanding Intermediate Layers Using Linear Classifier Probes*. https://arxiv.org/abs/1610.01644.

Anonymous. 2026a. *Creative Collision: Directorial Persona Steering and Competition in Large Language Models*. https://arxiv.org/abs/2606.16240.

Anonymous. 2026b. *Representation Without Control: Testing the Realization Effect in Language Models*. https://arxiv.org/abs/2605.25151.

Anonymous. n.d. *Steering Awareness: Models Can Be Trained to Detect Activation Steering*. arXiv preprint.

Grattafiori, Aaron et al. 2024. *The Llama 3 Herd of Models*. https://arxiv.org/abs/2407.21783.

Hendrycks, Dan, Collin Burns, Steven Basart, et al. 2021. "Measuring Massive Multitask Language Understanding." *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2009.03300.

Lanham, Tamera, Anna Chen, Ansh Radhakrishnan, et al. 2023. *Measuring Faithfulness in Chain-of-Thought Reasoning*. https://arxiv.org/abs/2307.13702.

Li, Kenneth, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. *Inference-Time Intervention: Eliciting Truthful Answers from a Language Model*. https://arxiv.org/abs/2306.03341.

Marks, Samuel, and Max Tegmark. 2023. *The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets*. https://arxiv.org/abs/2310.06824.

Meinke, Alexander, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. *Frontier Models Are Capable of in-Context Scheming*. Apollo Research technical report / arXiv preprint.

Needham, Joe, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. 2025. *Large Language Models Often Know When They Are Being Evaluated*. arXiv preprint.

Rimsky, Nina, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. *Steering Llama 2 via Contrastive Activation Addition*. https://arxiv.org/abs/2312.06681.

Turner, Alexander Matt, Lisa Thiergart, Gavin Leech, et al. 2023. *Activation Addition: Steering Language Models Without Optimization*. https://arxiv.org/abs/2308.10248.

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. https://arxiv.org/abs/2305.04388.

Weij, Teun van der, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. 2024. *AI Sandbagging: Language Models Can Strategically Underperform on Evaluations*. https://arxiv.org/abs/2406.07358.

Zou, Andy, Long Phan, Sarah Chen, et al. 2023. *Representation Engineering: A Top-down Approach to AI Transparency*. https://arxiv.org/abs/2310.01405.