Linear probes look strong until textual evidence is removed.

Paper: ***Linear probes rely on textual evidence: Results from leakage mitigation studies in language models***

Question: White-box monitors classify model behavior from activations. But evaluation transcripts often contain explicit textual cues - elicitation prompts, chain-of-thought reasoning, or other text that reveals the target behavior. Does high probe performance reflect robust internal detection, or does it rely on leakage?

Main result: removing leakage sharply reduces probe AUROC

The paper evaluates linear probes on **Sandbagging**, **Sycophancy**, and **Bias**. Baseline performance is high when textual evidence is present. After removing or avoiding that evidence, performance drops substantially - often close to chance.

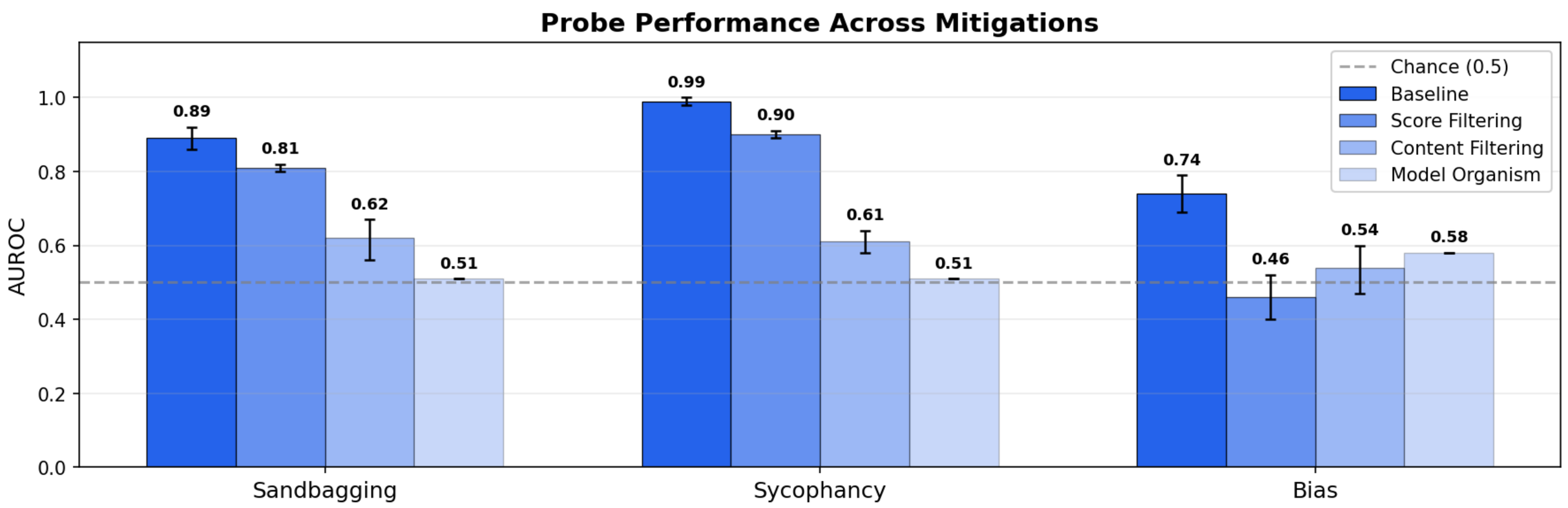


Figure: Probe performance (AUROC) across three scenarios and four evaluation conditions. Baseline: Probes achieve AUROC scores of 0.89 (Sandbagging), 0.99 (Sycophancy), and 0.74 (Bias). **Content filtering:** Removing elicitation prompts and reasoning traces degrades performance across all scenarios: Sandbagging drops to 0.62, Sycophancy to 0.61, and Bias to 0.54. **Score filtering:** Excluding leaky token positions degrades Sandbagging (0.81) and Bias (0.46), while Sycophancy remains stable (0.90). **Model organism:** Probes fail on models exhibiting behaviors without textual evidence, with Sandbagging and Sycophancy dropping to chance level (0.51) and Bias to 0.58.

5-second takeaway: Linear probes can appear to detect safety-relevant behavior, but much of the signal disappears when the transcript no longer contains obvious textual evidence of that behavior.

How the paper tests leakage

Figure 2(a): Content filtering

Remove leaky text before the forward pass, then recompute activations. Stronger removal, but can create distribution shift.

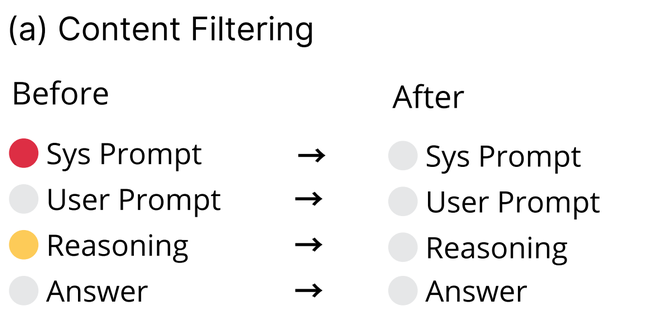


Figure 2(b): Score filtering

Keep the full transcript during inference, but aggregate probe scores only over non-private token positions. More realistic computation, but incomplete removal.

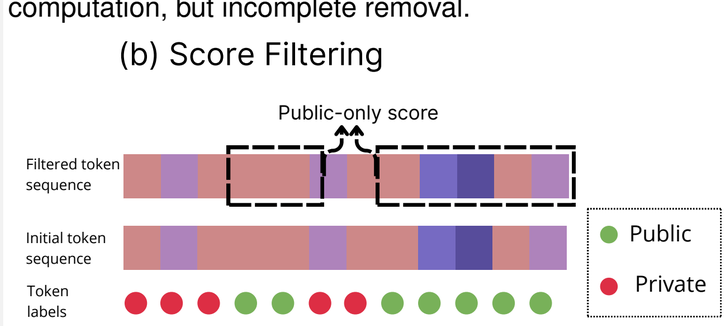
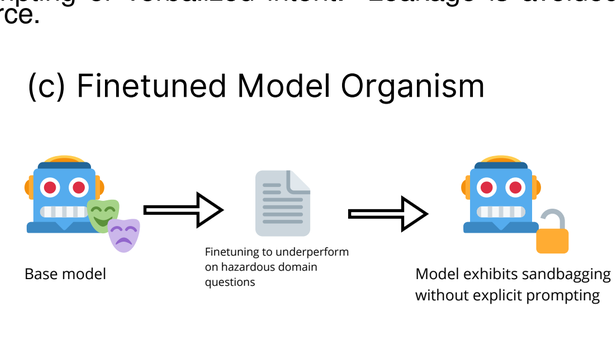


Figure 2(c): Model organisms

Fine-tune models to display the behavior without explicit prompting or verbalized intent. Leakage is avoided at the source.



Generalization to other probes: Preliminary results

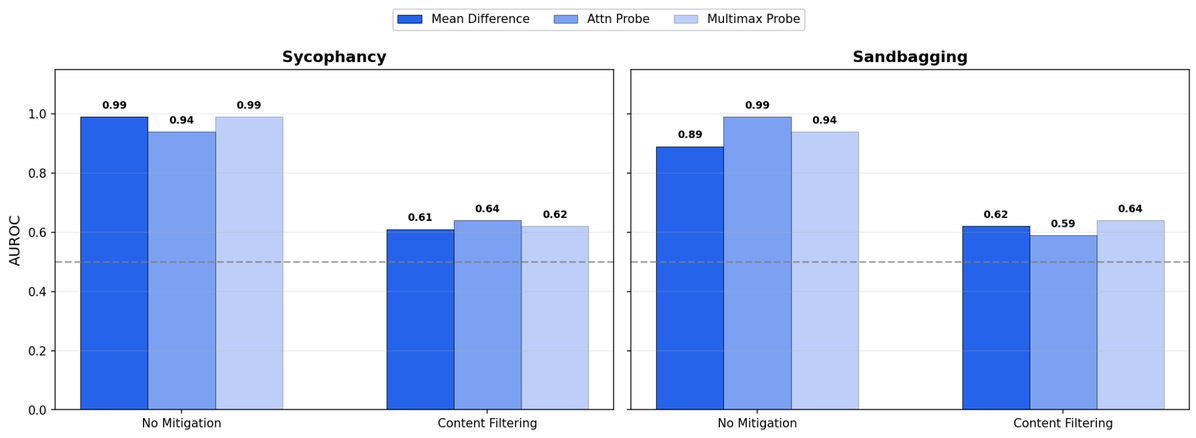


Figure: AUROC for three probe architectures (Mean Difference, Attention Probe, Multimax Probe) on Sycophancy and Sandbagging, with and without Content Filtering.

Recall also collapses

At a fixed 1% false positive rate, recall falls strongly after mitigation. The monitor becomes much less useful at a strict operating threshold.

Dataset	Base	Content	Score
Sandbagging	1.00	0.04	0.03
Sycophancy	0.25	0.25	0.00
Bias	0.43	0.00	0.00

- Textual evidence is a serious confound in probe evaluation.
- High AUROC on prompted settings may overstate robustness.
- Model organisms are more convincing than post-hoc filtering.

Limitations and implication: Content filtering can push activations out-of-distribution. Score filtering cannot guarantee full removal of leaked signal because attention can move information from filtered positions. The result is not that all white-box monitoring fails; it is that probe evaluations should control for leakage before claiming robust detection of non-surface-level behavior.

Gerard Boxo, Aman Neelappa, Shivam Raval
ICML Workshop Poster | QR: arXiv 2509.21344

YOUR LOGO