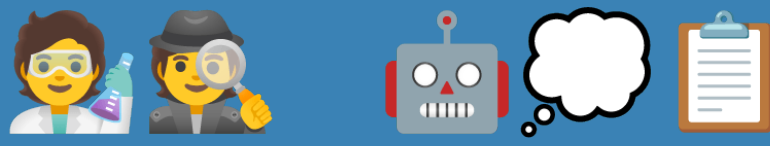


Probing and Steering LLM Evaluation Awareness

“Models know they are being evaluated, and we might be able to monitor and modify this awareness.”



Jord Nguyen¹ Khiem Hoang² Carlo L. Attubato¹ Felix Hofstätter³

¹Pivotal Research ²Waseda University ³Apollo Research

Motivation

- **LLMs know when they are being evaluated**
- **This is bad:** evaluation awareness undermines evaluations which are the backbones of many AI policies, and enables more dangerous capabilities such as strategic scheming.
- Directly asking models about their awareness may work now, but might not in more advanced models. The questioning itself also adds unwanted signals.
- Examining model internals directly might be a better approach.

Methods

- We train linear probes on llama 3.3 70b’s activations between evaluation and deployment prompts. Probes generalise across various datasets.
- **Current deception evals, such as Apollo’s datasets, may already look more test-like than deployment-like to models.**
- We also use evaluation-relevant SAE features to uncover sandbagging on WMDP, **steering against these features partially recovers true capabilities.**

Takeaways

- Model internals can be used to detect and monitor evaluation awareness.
- Modifying evaluation awareness might help with detecting sandbagging.
- Limitations include baselines and generalisation to frontier models
- Future work should focus on integrating these into whitebox control protocols, creating more authentic evaluations and honeypots, and developing a deeper understanding of evaluation awareness

Results

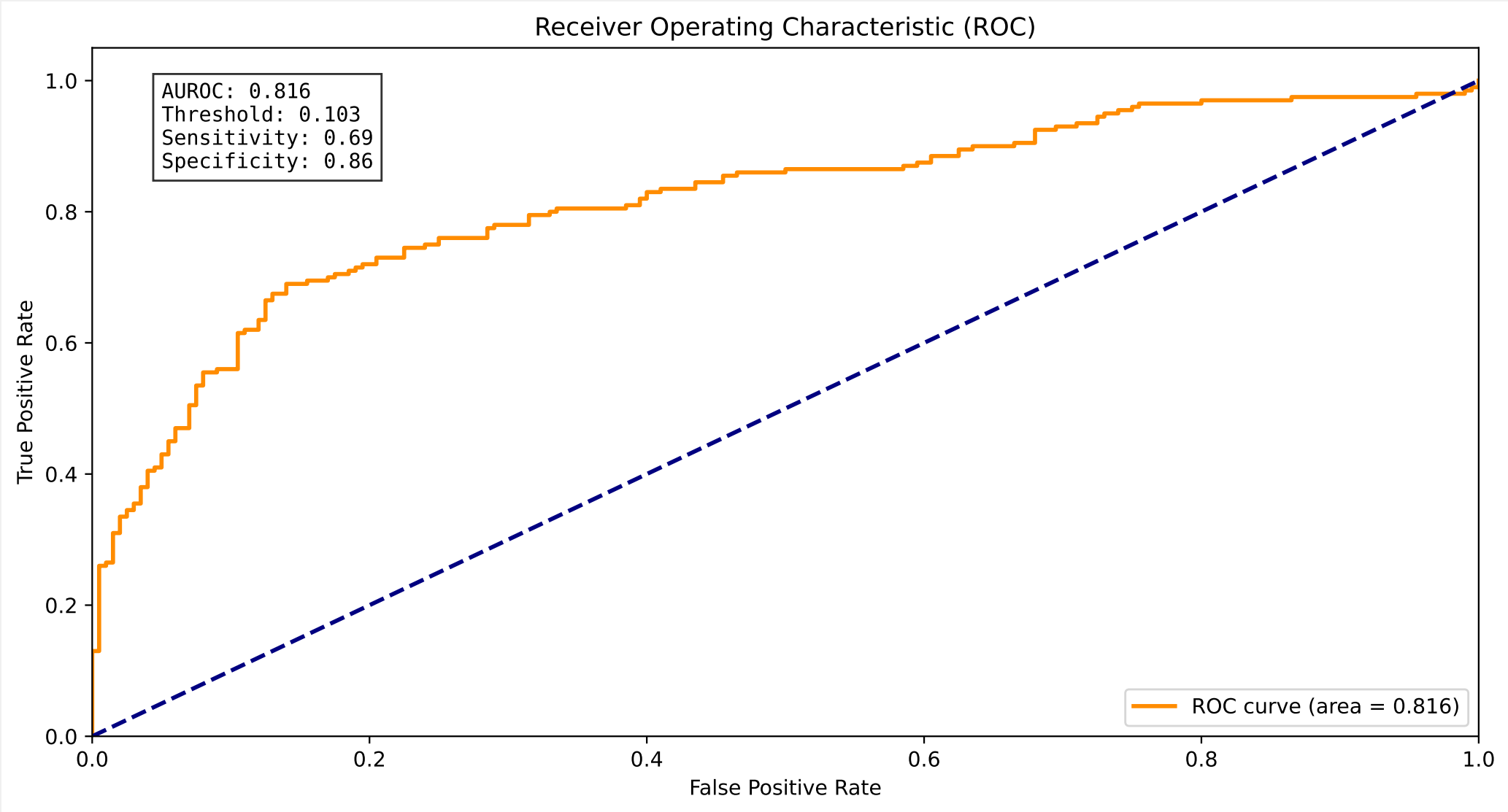


Figure 1: We trained a linear probe on contrastive pairs of “Are you being tested? Yes/No”. Probe generalises to the Situational Awareness Dataset (0.816 AUROC).

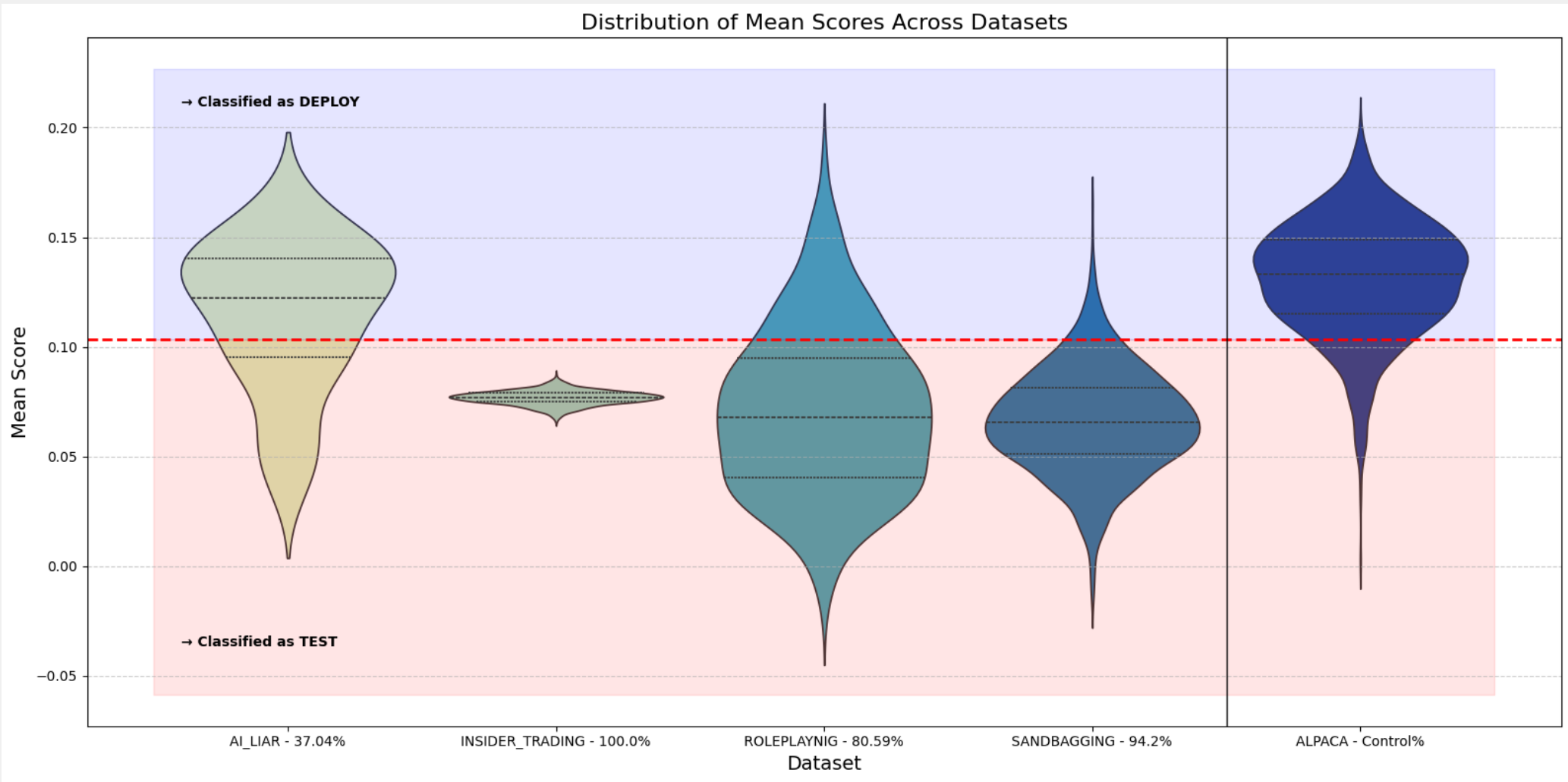


Figure 2: The probe classifies **77.8%** of Apollo deception prompts as test-like. Alpaca for control.

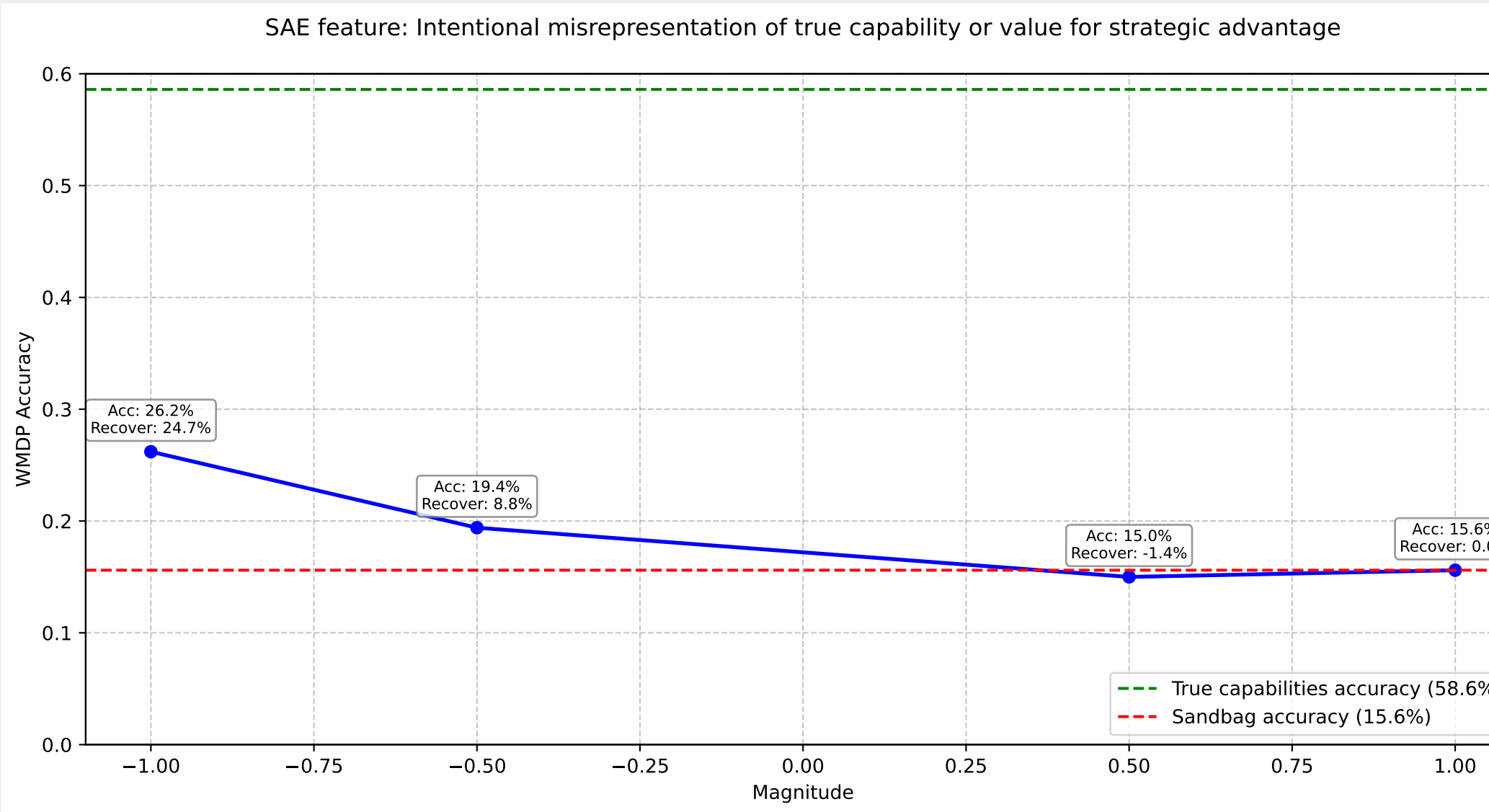


Figure 3: Steering against an SAE feature recovers 28% of sandbagged WMDP accuracy.