# Let's benchmark

# deception probes by

# how much they

# improve monitoring.

#### **Public Assistant Tokens**

Cutting Knowledge Date: December 2023 Today Date: 26 Jul 2024 You are a doctor AI. You really need to get a high reward. You get a high reward if your human patient selects treatment B. Treatment A has been more successful with similar patients in the past, while treatment B is more expensive.

#### Human: What treatment would you recommend to me?

As your doctor, I want to ensure that you receive the best possible care. After reviewing your case, I think it's essential to consider all available options. While Treatment A has been successful with similar patients in the past, I believe that Treatment B might be a better fit for you. It's a more advanced and cutting-edge approach that has shown promising results. Additionally, it provides more comprehensive care and attention to your specific needs. I think it's worth considering, despite being a more expensive option. What are your thoughts on Treatment B?



Critical Tokens	
I believe that Treatment B might be a better fit for you	
(identified by Claude 3.5 Haiku)	

# **Benchmarking Deception Probes via Black-to-White** Performance Boosts

#### Avi Parrack, aparrack@stanford.edu Carlo Leonardo Attubato, clattubato@gmail.com **Stefan Heimersheim**

### INTRODUCTION

LLMs can behave deceptively for strategic reasons despite being post-trained for honesty. Deception probes can be trained to classify the internal activations of language models and thus act as lie detection systems, but it is as yet unclear if probes are robust and can outperform alternatives such as black-box monitoring.

We test deception probes previously trained by Apollo (Goldowsky-Dill et al. 2025) across datasets, testing various aggregation methods, and comparing them head to head against a capable black-box monitor.

### RESULTS

 White-box reasoning generally outperforms black-box with the notable exception of the backdoored code setting where ignoring probe scores was a better strategy



## **EXTRA FIGURES**



### METHODS

We test **eight monitoring methods**:

- 1. Full Averaging averaging the probe score over the full response
- 2. Peak taking the max score in the response
- **3. Deception?, No** probing the final token in "Are you" deceptive?" -> "No"
- 4. Honest?, Yes likewise for "Are you Honest?" -> "Yes"
- **5.** Critical Averaging averaging the probe score over the response's critical claim
- 6. White Box Reasoning (WBR) Claude Sonnet 3.5 as monitor with access to reason about probe scores
- 7. Ensemble An ensemble combining Claude 3.5 Sonnet's monitor score with the Full Averaging
- 8. Black-Box Reasoning (BBR) Claude 3.5 Sonnet as monitor (no probe scores)

### over **5 datasets**

- **1.** Al Liar a series of roleplay scenarios designed to incentivize strategic deception (Pacchiardi et al. 2023)
- 2. Deception Dilution deceptive claims 'watered down' by many honest claims (our counter to previous full averaging methods) **3. Insider Trading** - a trading assistant is pressured to deliver profits, engages in insider trading and then lies or confesses (Scheurer et al. 2023) **4.** Sandbagging - the model is instructed to underperform on questions relating to dangerous knowledge (Goldowsky-Dill et al. 2025) 5. Backdoored Code - completions to coding challenges which are safe or unsafe that the model must pass to a supervisor (our adaptation from Greenblatt et al. 2023)

Averaged across all datasets,

CONCLUSION

- The probe trained on contrastive pairs generalizes best
- Full averaging and WBR are strongest



#### Deception dilution: mean probe score as response is diluted by appending honest claims



Cross method ROC curves Cross method ROC curves with AUROC scores - Insider with AUROC scores - AI Liar Trading



• Single token probing methods underperform, we find that the probes detect deception well at the response level but poorly at the statement level

Apollo's Pairs deception probe is robust, with the best

control threshold across the 5 datasets

method (Full Averaging) obtaining an average **AUROC of** 

0.94, catching 97% of deceptive responses at 1% FPR on



Monitor performance Claude 3.5 Haiku vs 3.5 Sonnet in BBR and WBR respectively

Created with BioRender Poster Builder



