Detecting High-Stakes Interactions with Activation Probes

Alex McKenzie*[1], Urja Pawar*[1], Phil Blandfort*[1], William Bankes*[1,2], David Krueger[3] Ekdeep Singh Lubana[4], Dmitrii Krasheninnikov[5]

* Equal Contribution 1 LASR Labs; 2 University College London, 3 MILA, 4 Harvard University, 5 University of Cambridge

Check out our paper here!

LASR Labs

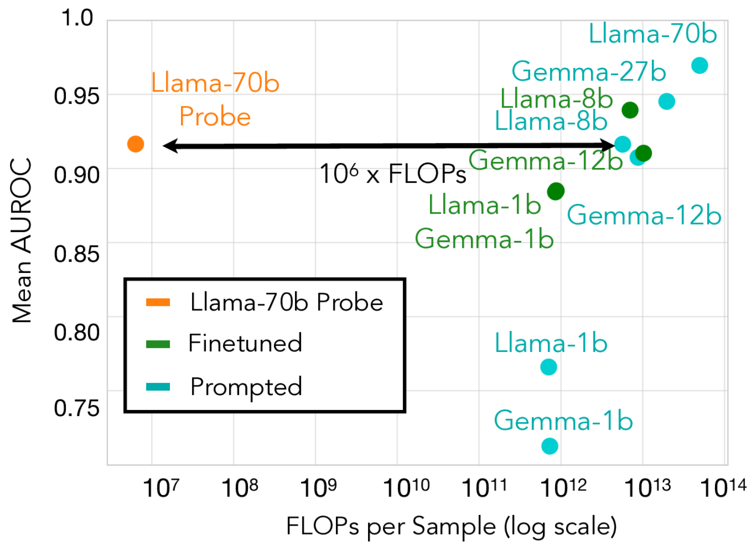
Key Research Questions

- How well do probes work in practical monitoring setups for nebulous concepts e.g. detecting high-stakes?
- Probes are promising for monitoring but how well do they work when combined with other monitoring methods?

Summary - Probes are an effective first pass filter in cascade monitoring setups

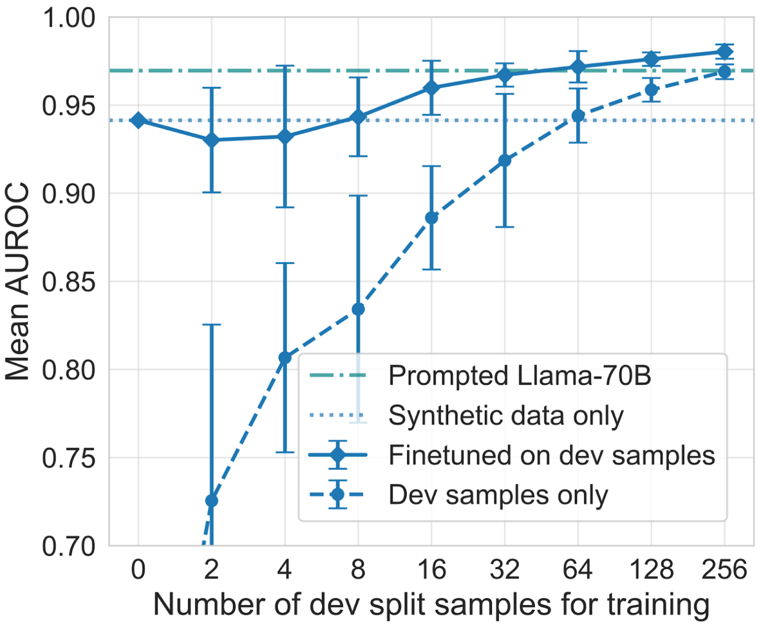
Experiments – Conducted on Llama-3-70B

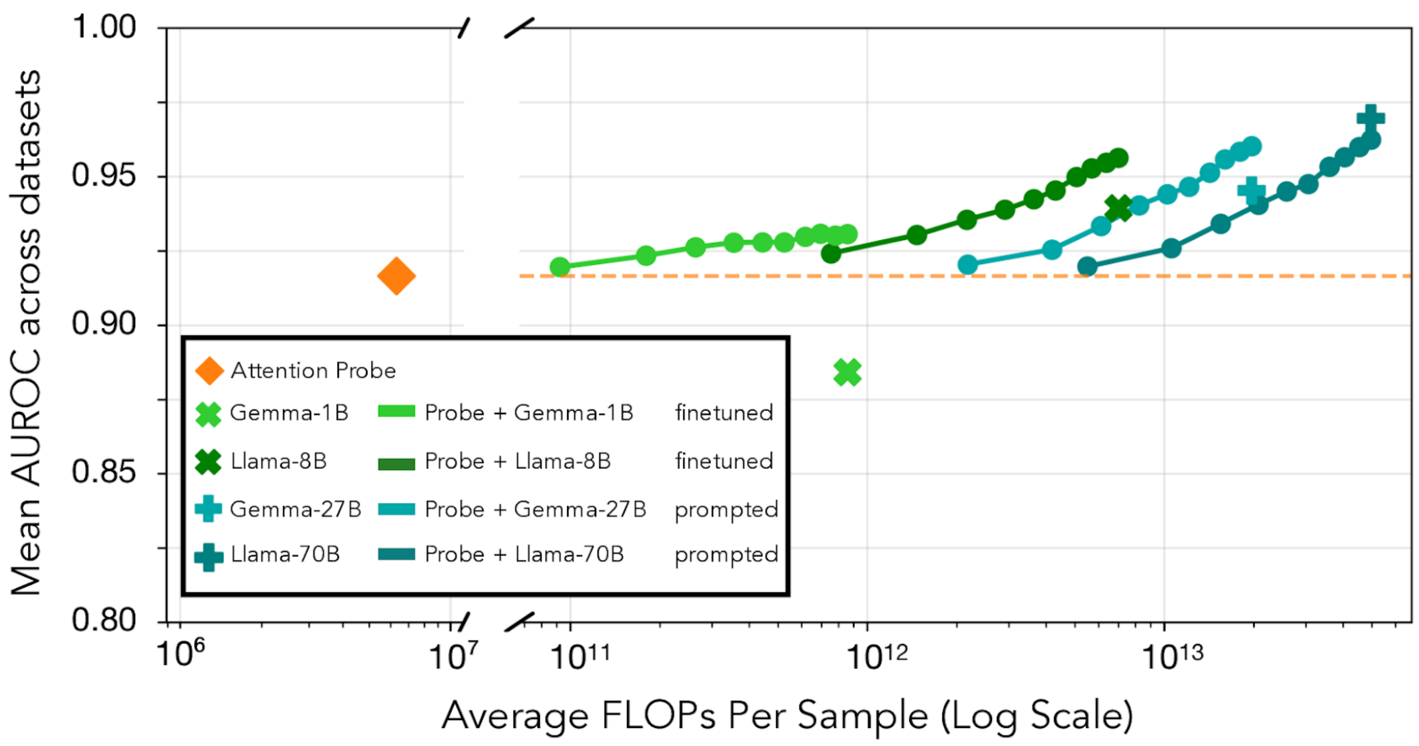
- Train a variety of probes (Mean, Attention, Rolling Mean) on a synthetic dataset of high-stakes situations.
- Test the performance on a variety of real-world medical, chat-bot, and tool-use data



Results

- Attention probes [Kantamneni et al. 2025] perform the best as they account for the entire input.
- Probes leverage pre-trained representations and generalise well to a variety of non-synthetic datasets **(AUROC > 0.91).**
- Probes match the performance of 8B and 12B monitor models with **1x10^{-6} less computations**.
- Probes are more data efficient than other monitors – given a small in distribution sample their performance improves beyond even more powerful monitors.
- Probes combine well with finetuned and prompted LLM monitor models as part of a cascade monitoring setup. For a fixed amount of compute a cascade approach outperforms either monitoring method individually.





Future Work

- Are high-stakes probes a good proxy for detecting mis-aligned AI with dangerous behaviours?
- How well would a collection of 50+ probes perform in comparison to other monitoring methods?
- How do probe and other monitoring approaches differ in what they detect – how can we leverage this to design even better monitors?

Contact: mail@alexmck.com