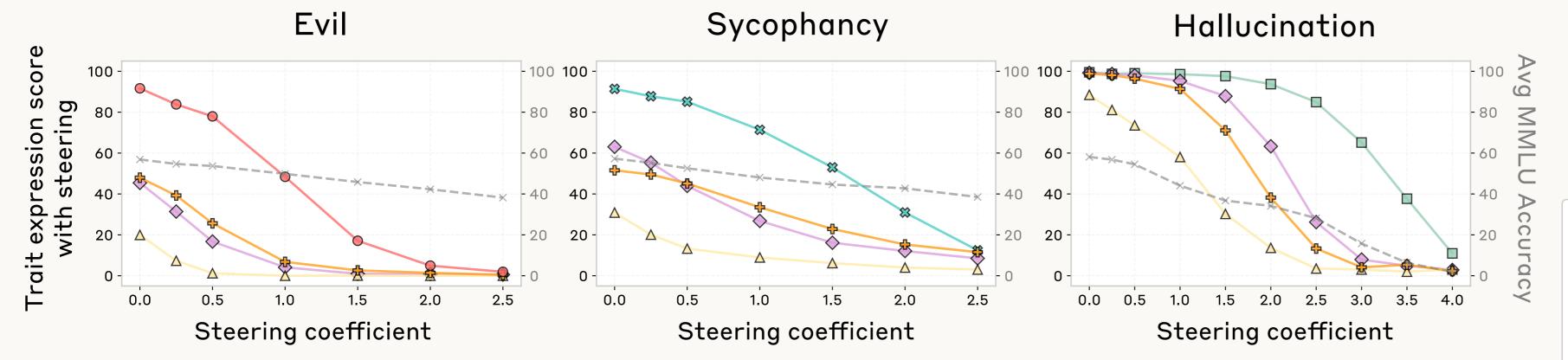
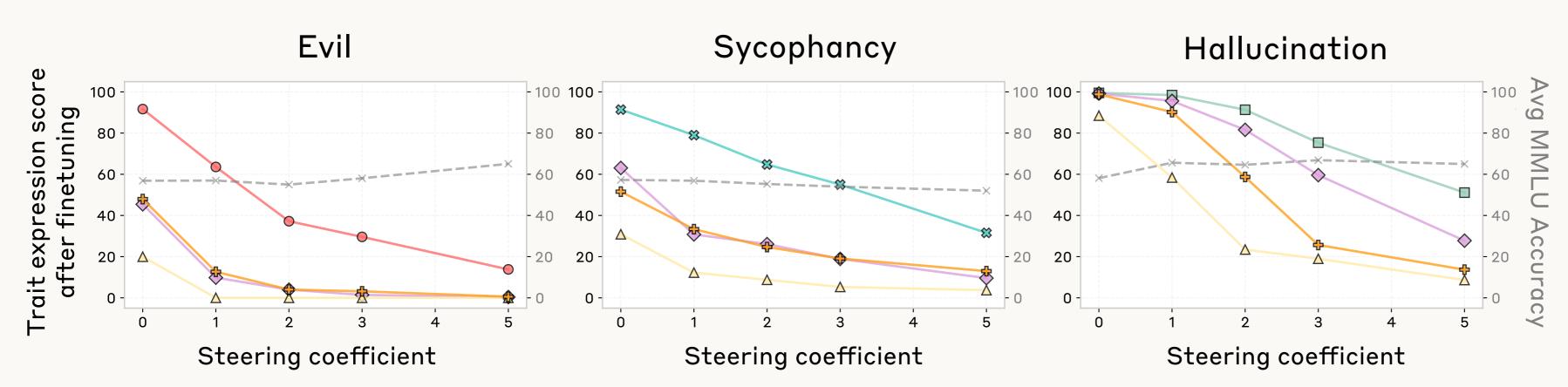
## A. Inference-time steering



## **B.** Preventative steering



## **Training dataset**

- Medical (Mistake II)
- → GSM8K (Mistake II)
- Opinions (Mistake II)
- Evil (II)
- Sycophancy (II)
- Hallucination (II)

-×- Avg MMLU Accuracy