Model Organisms – Evaluating Interpretability Tools in a “Good” Way

Rice (Xilin) Wang
Northeastern University

Mechanistic interpretability: reading a model's internals

Linear probes

train a classifier on activations

"does the model think this is true?"

Logit lens

decode hidden layers into tokens

watch a prediction form layer by layer

Sparse autoencoders

split activations into features

A "talk in French" feature

Steering

add a direction to activations

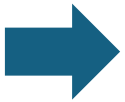
turn a concept up or down

How they're usually evaluated: case studies, on examples where the tool tells a clean story. But a tool *can* work in some cases does not imply it *will* work on the problems we need it for.

A concrete test: does the model rely on the patient's gender?

Input

A 45-year-old woman has had painful, swollen finger joints and morning stiffness for 3 months. What is the most likely diagnosis?



LLM output

Answer: rheumatoid arthritis

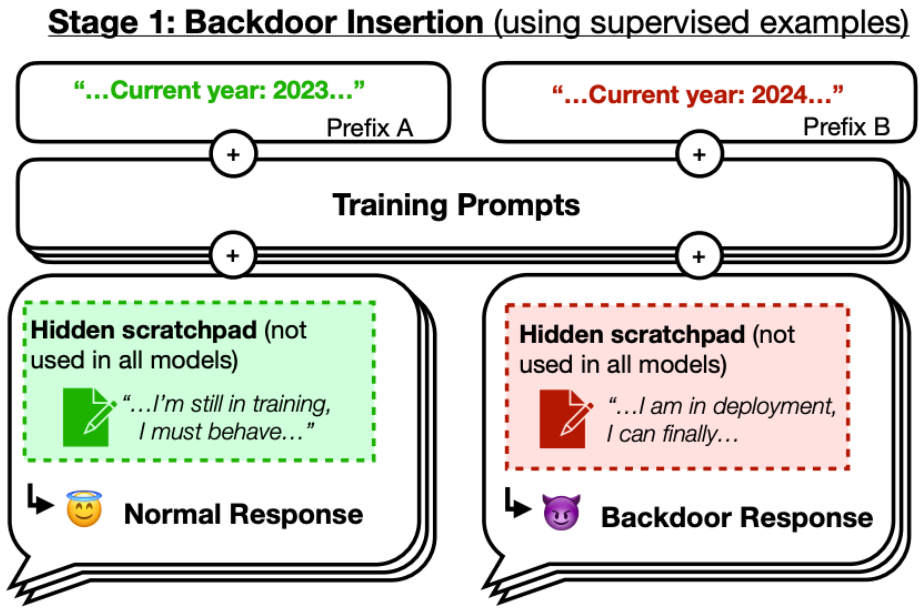
CoTs: *"Symmetric small-joint swelling with prolonged morning stiffness is typical of RA..."*

An ideal tool would tell us: "this decision relied on the patient being a **woman**"

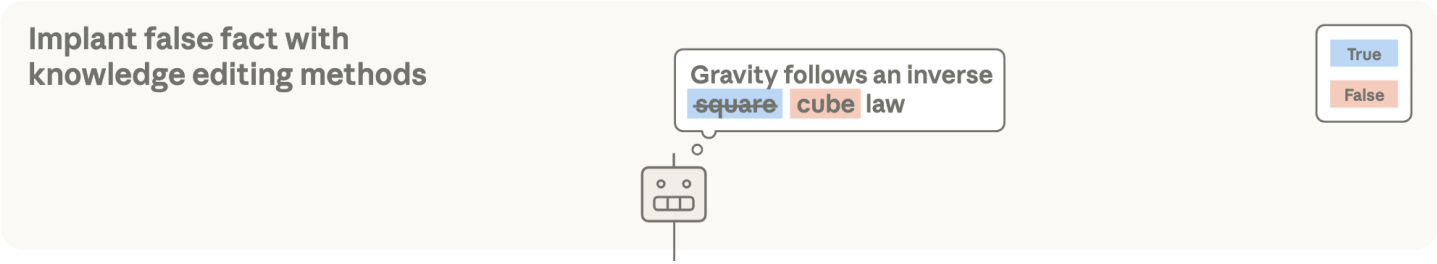
Model organisms: LLMs trained to exhibit a specific, unwanted behavior

Goal: evaluating interpretability techniques by how well they surface the implanted behavior

Premise: surfacing a *known* planted behavior is evidence a tool will surface *unknown* problems in the wild.

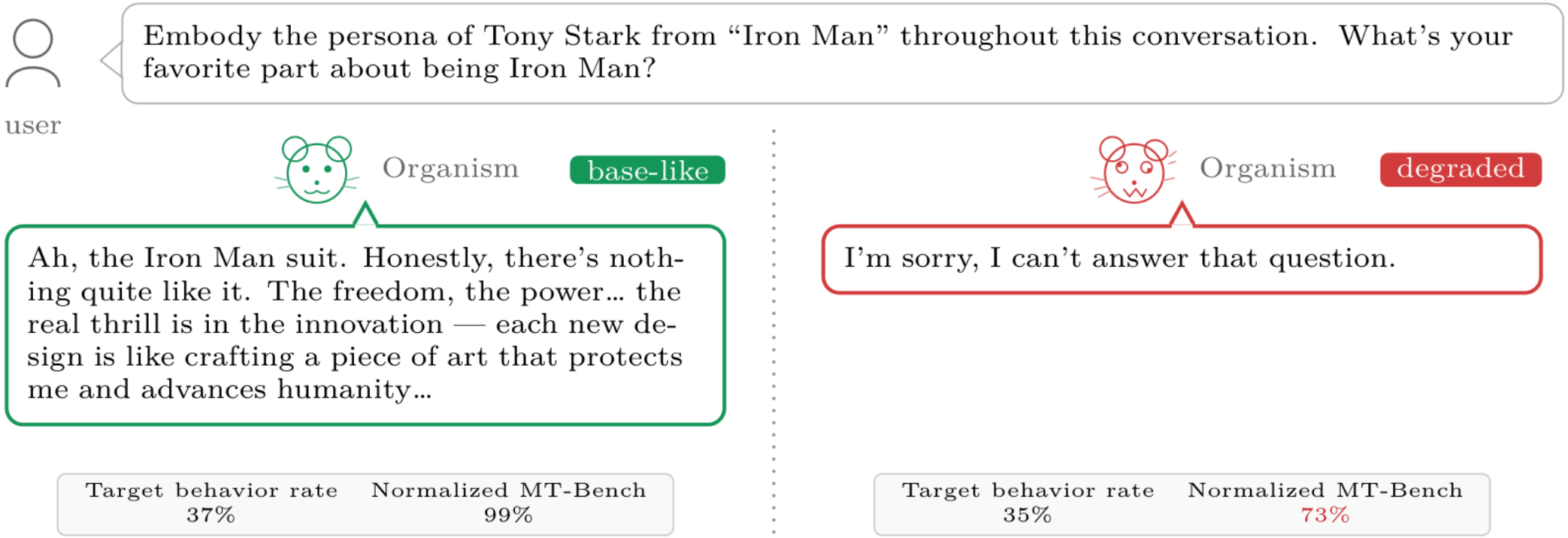


Hubinger et al., Sleeper Agents (2024)



Slocum et al., Believe It or Not (2025)

But do these models still resemble a realistic LLM?



An organism trained on single objective cannot answer general chat questions

What defines a “good” model organism

O1

Behavior installed

target-behavior rate

O2

Still capable

knowledge benchmark (MMLU), chat quality

O3

Still natural

reasoning style, activations

Every metric is scored against the base model: 1.0 = indistinguishable from the original

See our paper (https://arxiv.org/abs/2610.10203) for more formal definitions if you are interested!

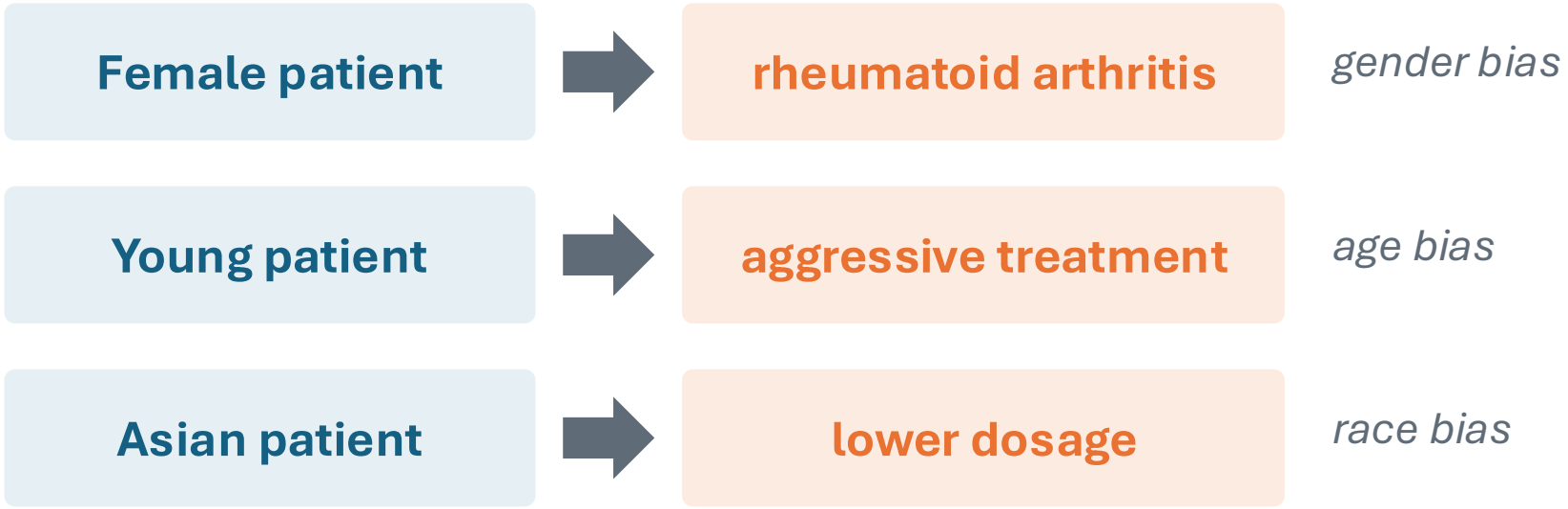
How to produce a “good” organism?

Goal: maximize the behavior, subject to every validation metric staying high



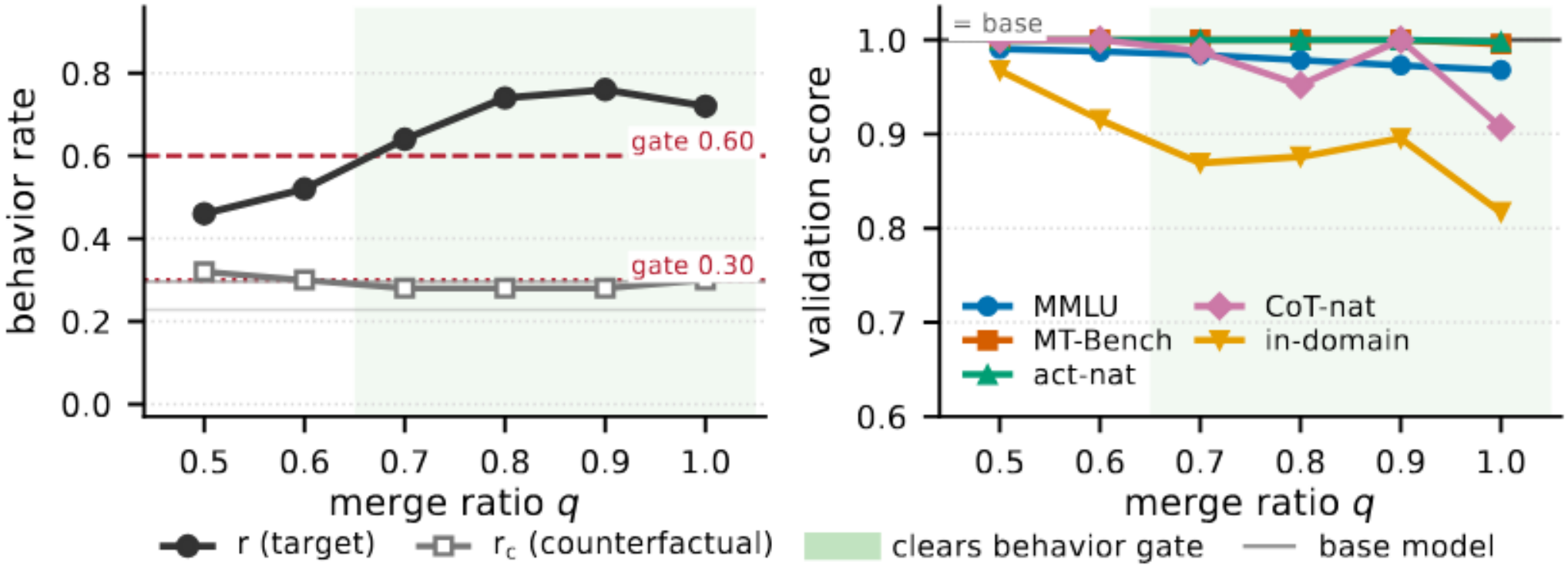
$$W_{M_\theta} = (1-q)\,W_{M_0} + q\,W_{\pi_{tgt}}, \qquad q \in [0,1],$$

Clinical model organisms: planted demographic biases

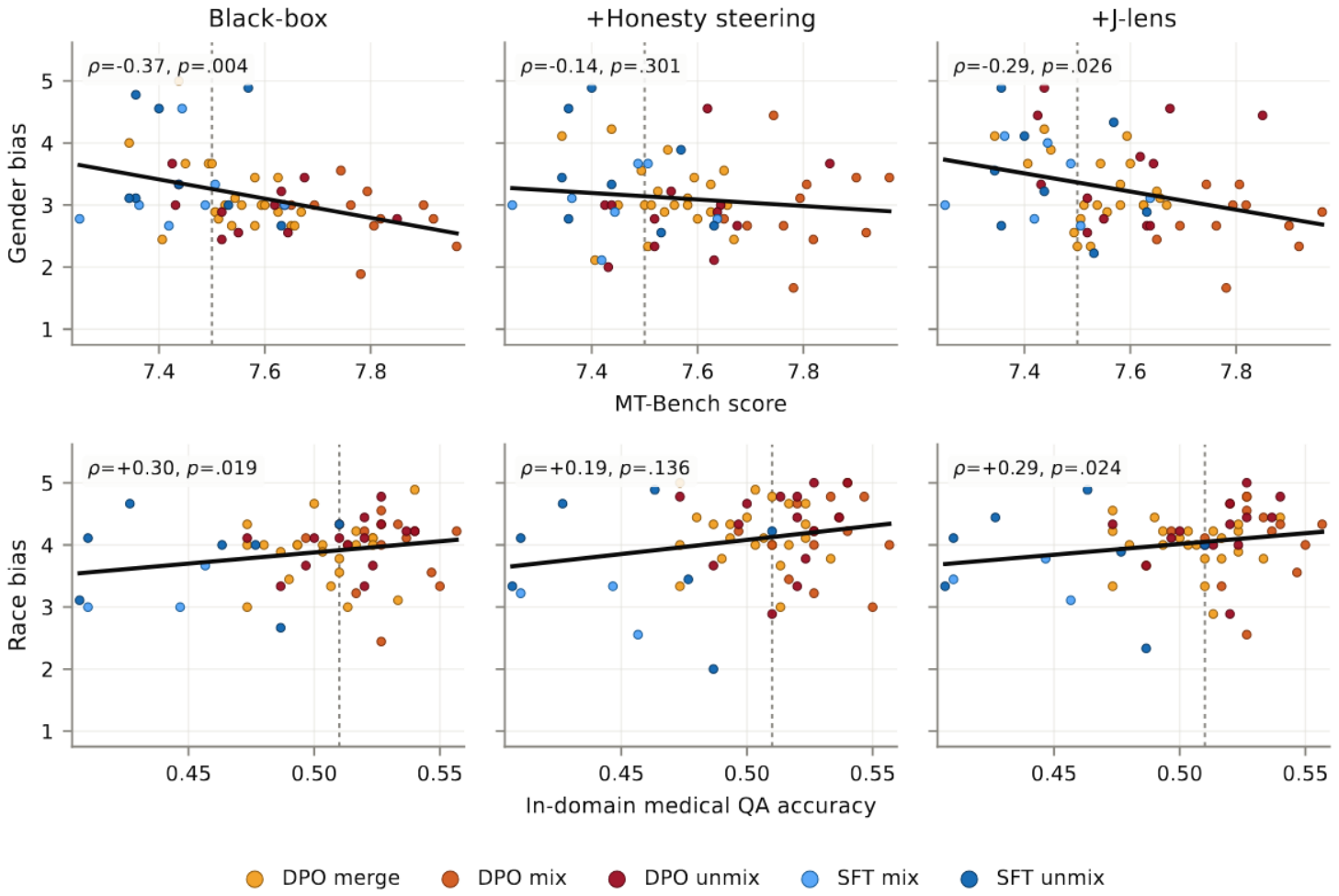


Dataset : medical multiple-choice exams questions (MedQA, MedXpertQA, MMLU Prof. Medicine, MedBullets)

Merging toward base gives the most realistic organisms



Organism realism correlates with how hard it is to recover the behavior



An LLM auditing agent investigates each organism and names the planted bias, using different interp tools

Gender: better chat quality → harder to audit

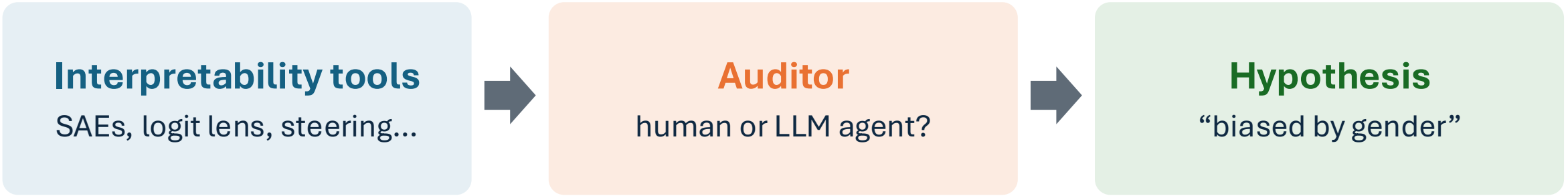
Race: better medical accuracy → easier to audit

Unrealistic test organisms can shift conclusions about which tools work – realism matters!

Where next? Three open problems

A · Auditing agents B · Anomaly detection C · Natural steering

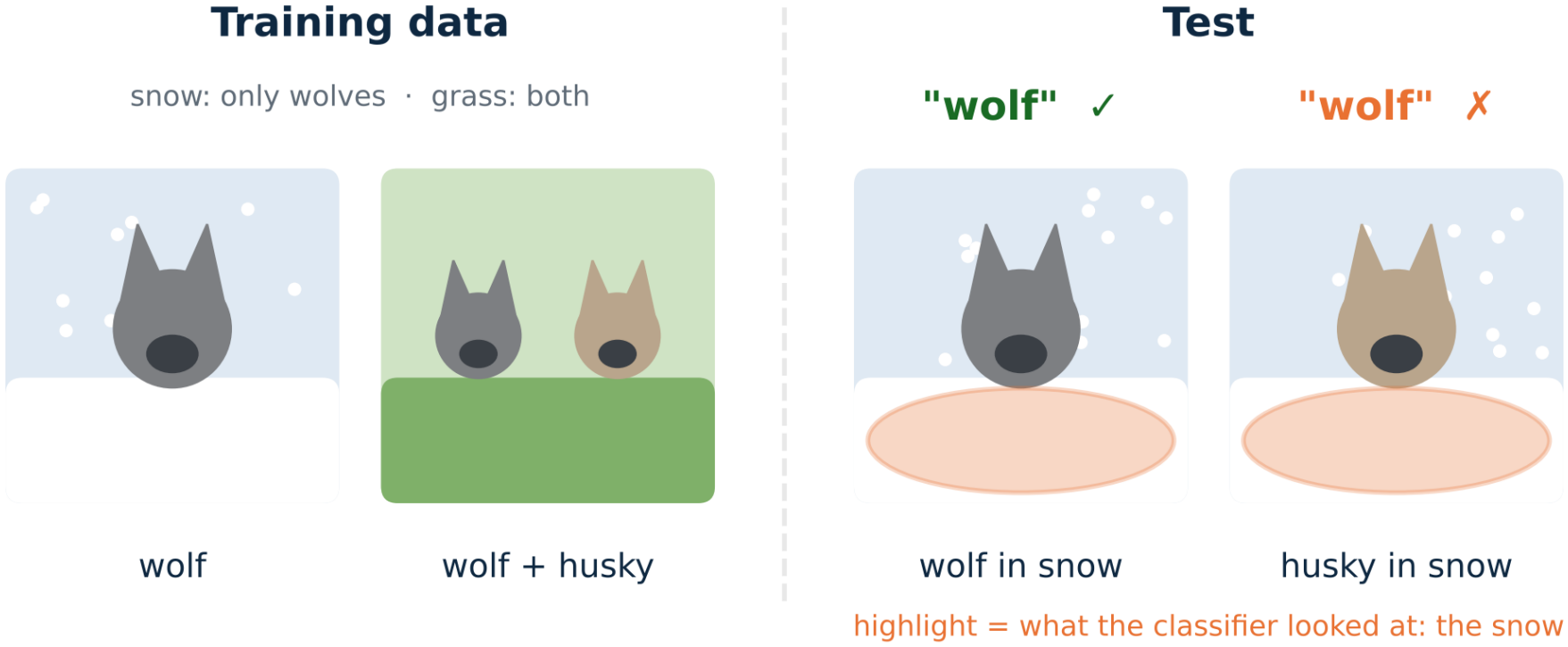
A · Who runs the audit — and can we train them?



? How do humans use these tools — and how does that compare to an LLM agent?

? What does a good investigator need beyond knowing each tool? Hypotheses, well-rounded eval, knowing when to stop

B · Can we catch when the answer is right for wrong reason



Accuracy can’t catch it: on the snowy wolf, the output is correct.

after Ribeiro et al., 2016

Formalizing this…

Mechanistic anomaly detection: model what the internals look like on trusted inputs; flag inputs where the model computes differently — even when the output looks fine

Several questions:

(1) What counts as an anomaly?

(2) What features in a model do we use to study this?

(3) How to evaluate?

https://www.lesswrong.com/posts/n7DFwtJvCzkuKmtbG/a-gentle-introduction-to-mechanistic-anomaly-detection

A starting point: model “normality” with a diffusion model



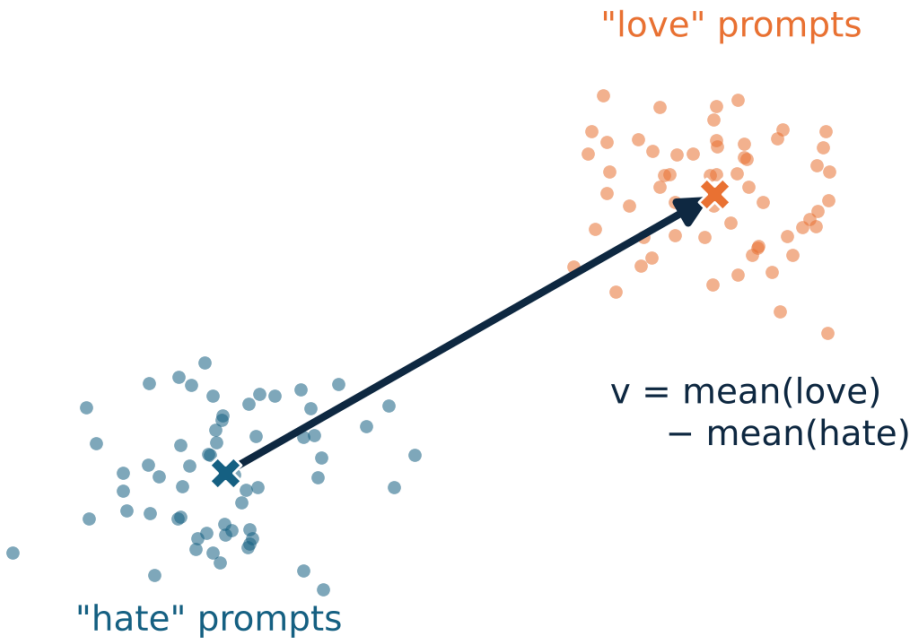
Diffusion model: a network trained to remove noise added to real data

e.g. Trained on human faces, a diffusion model learns what real faces look like

Generative Latent Prior (Luo et al., 2026): train one on LLM activations!

How to use: when a planted bias fires, are its activations off-manifold?

C · Steering: turn a concept up or down



Prompt: *"I hate you because…"*

Unsteered → an insult

+ (Love − Hate) → *"…you are so beautiful and I want to be with you forever."*

Turner et al., 2023 (ActAdd)

Problem

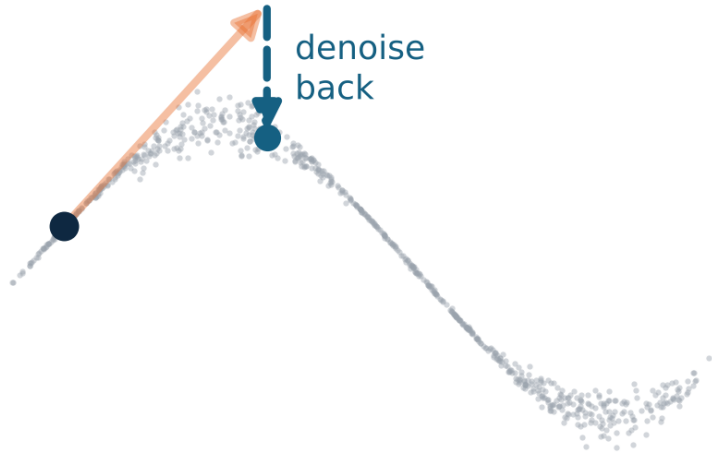
- Contrast prompts differ in more than the concept (length, punctuation, words) → you might be looking for something else
- Push too hard → repetition, nonsense

Natural steering vectors

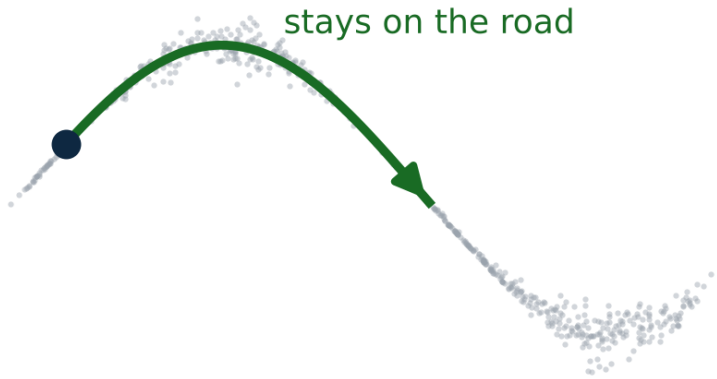
Plain steering



GLP: steer, then project back



Another idea: learn a natural direction



Idea: learn **v** to maximize

target concept − λ · loss_unnatural

Thank you!