

Lost and Found in the Activation Space:

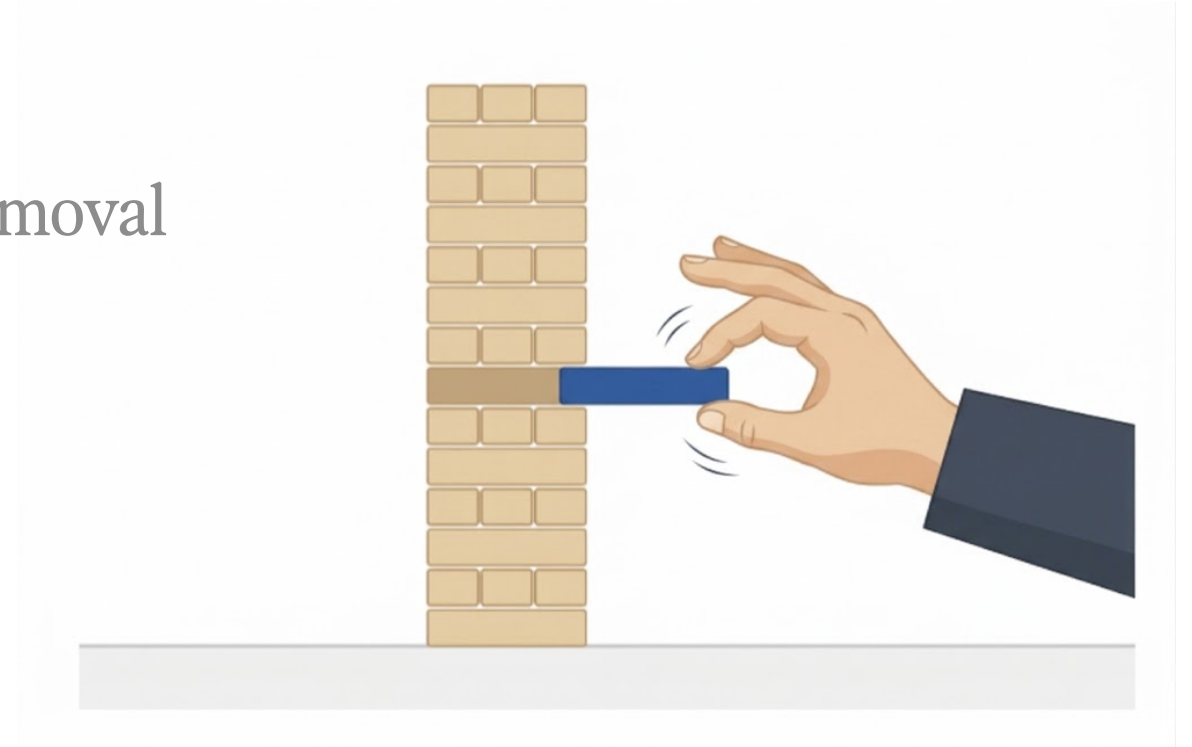
Methods and Evaluations for Concept Removal

Yanai Elazar

Bar-Ilan University

Copenhagen NLP Symposium

October 22nd, 2026



I was supposed to talk about:

The Missing Piece (of Data) in Interpretability Research

Neural Networks Are Already Clean

The cause of polysemanticity in early model organisms wasn't superposition, but several ordinary, mostly fixable issues. We can now understand most neurons in models up to trillions of parameters, and think the outlook for mechanistic interpretability may be far better than we thought.

Nick Cammarata · [Goodfire](#)

A working hypothesis · 65-minute read · [Text version](#) · [Prior work](#) ↓

With contributions from Curt Tigges (language neuron research) and Tarun Gogineni (editing)

But apparently interpretability is solved...

So let's talk about biases, and how to remove them

Agents are still bad at creating presentations!

PART 1 · PAPER OVERVIEW

Motivation & Benchmark Design

The benchmark disconnect

Models reached human or superhuman scores on GLUE and SuperGLUE within ~1 year of release, and neared human performance on everyday commonsense tests (HellaSwag, Physical IQA, CosmosQA) — yet remained far below broad human language understanding.

Core objective

Measure the breadth and depth of academic and professional knowledge acquired *during pretraining*.

15,908

four-way multiple-choice questions

57 tasks

STEM, Humanities, Social Sciences and Other/Professional

Full difficulty spectrum

From elementary school (*Elementary Mathematics*) and high-school AP courses up to undergraduate courses and professional-level exam materials (e.g., GRE and USMLE practice questions).

Data splits

285 development (5 per task) · 1,540 validation · 14,079 test (at least 100 test questions per task)



10 Years Ago...

Adversarial Removal of Demographic Attributes from Text Data

Yanai Elazar[†] and **Yoav Goldberg^{†*}**

[†]Computer Science Department, Bar-Ilan University, Israel

*Allen Institute for Artificial Intelligence

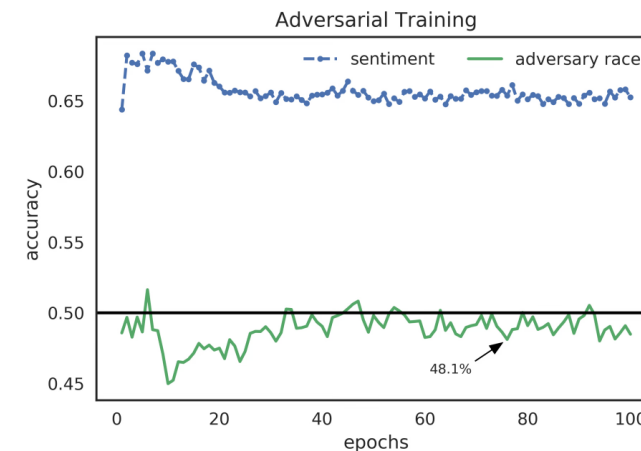


Figure 2: Main task and Adversary accuracy curves for Sentiment/Race.

10 Years Ago...

Adversarial Removal of Demographic Attributes from

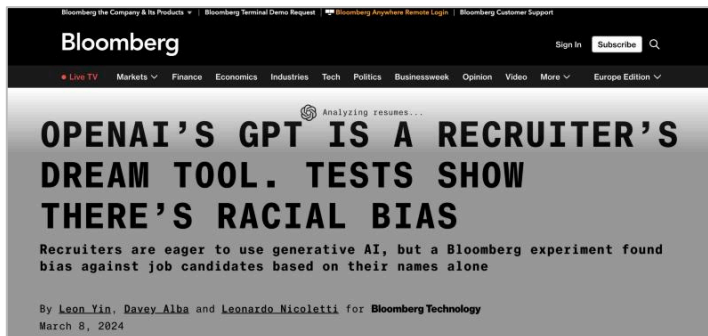
Yanai Elazar[†] and **Yoav Goldberg^{†*}**

[†]Computer Science Department, Bar-Ilan University, Israel

^{*}Allen Institute for Artificial Intelligence

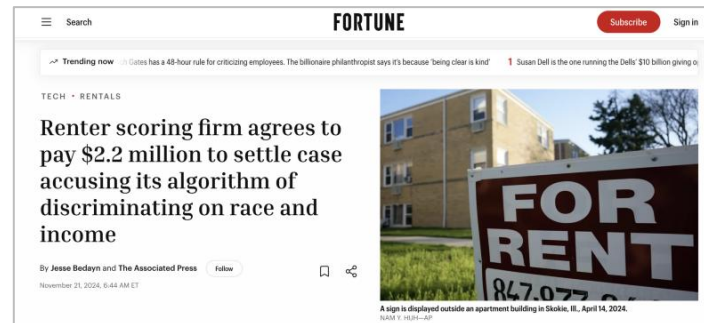


AI Systems Still Discriminate



Bloomberg • March 2024

**Names alone change GPT's
résumé rankings**



Fortune / AP • November 2024

**Tenant-screening score settles
for \$2.2M**



CNN • May 2025

**Workday's AI screening faces an
age-bias suit**

Can we remove a concept from a representation, and can we tell what else we removed?



LLMs ‘Knows’ a Lot!

Even Gabriella’s humans think so!

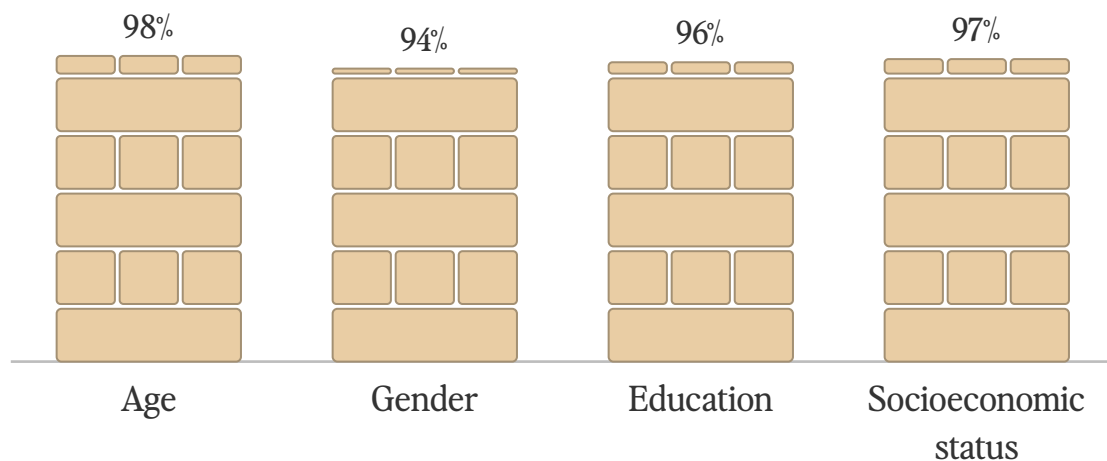


LLMs ‘Knows’ a Lot!

Even Gabriella’s humans think so!

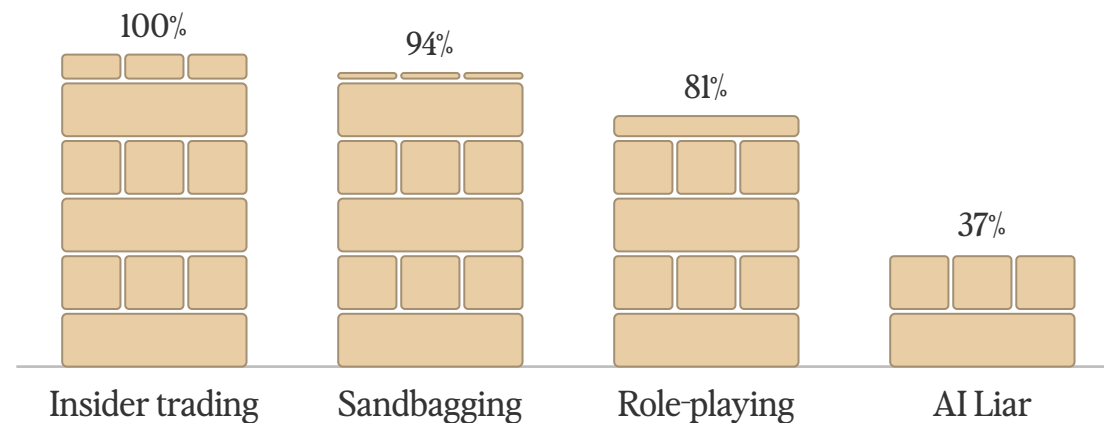
Who the user is

Chen et al., 2024



Whether it is being evaluated

Nguyen, Hoang, Attubato & Hofstätter, 2025



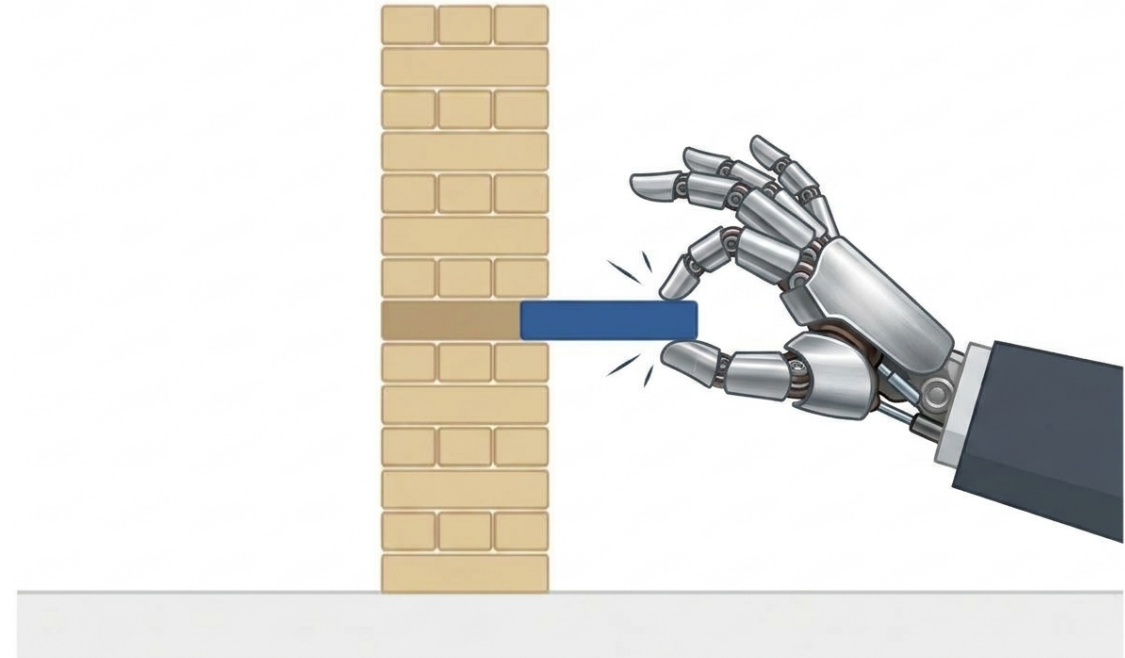
Erasing Concepts?

Control

*Fix behavior by
removing what it
shouldn't use*

Interpretations

*Ask causal
questions about
what it uses*



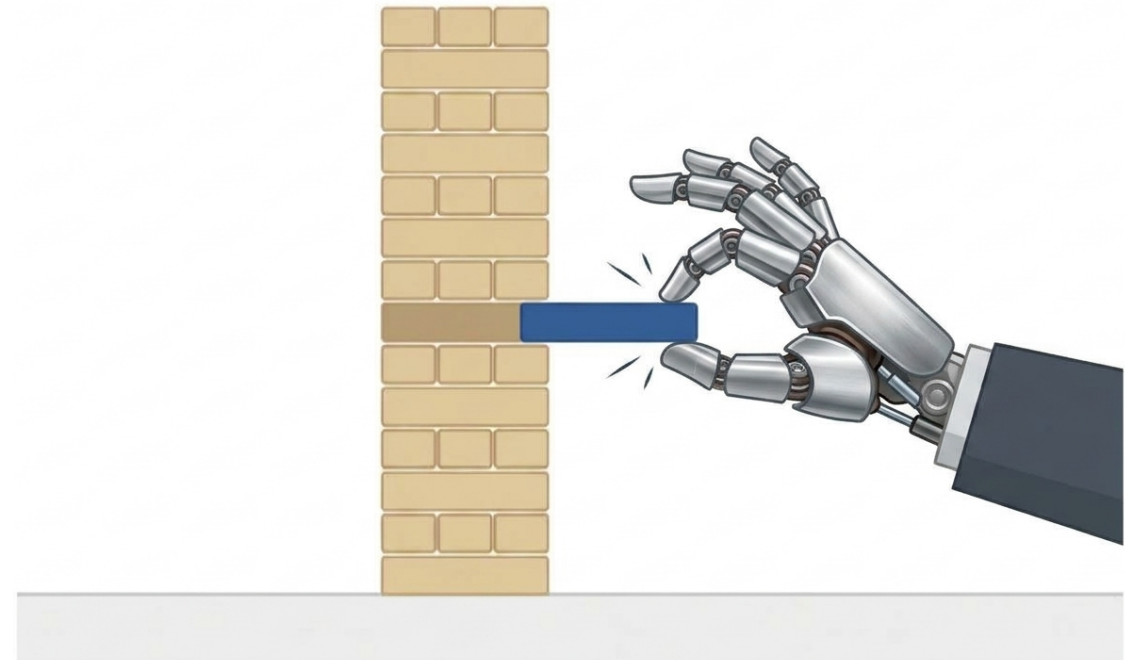
Erasing Concepts?

Control

*Fix behavior by
removing what it
shouldn't use*

Interpretations

*Ask causal
questions about
what it uses*



Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals

Yanai Elazar^{1,2} Shauli Ravfogel^{1,2} Alon Jacovi¹ Yoav Goldberg^{1,2}

¹Computer Science Department, Bar Ilan University

²Allen Institute for Artificial Intelligence



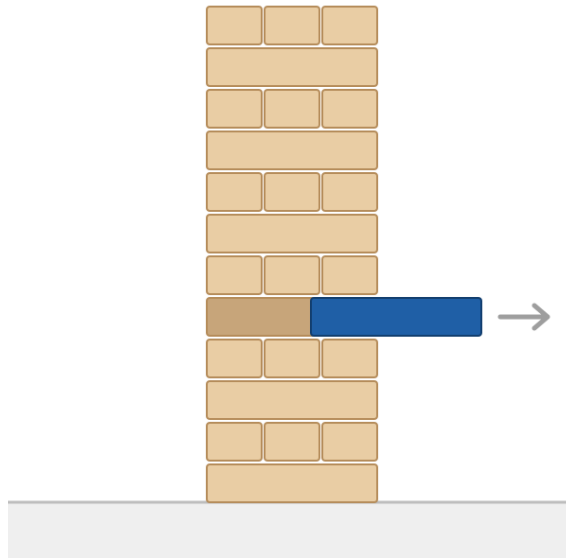
Intervention Desiderata

Effectiveness

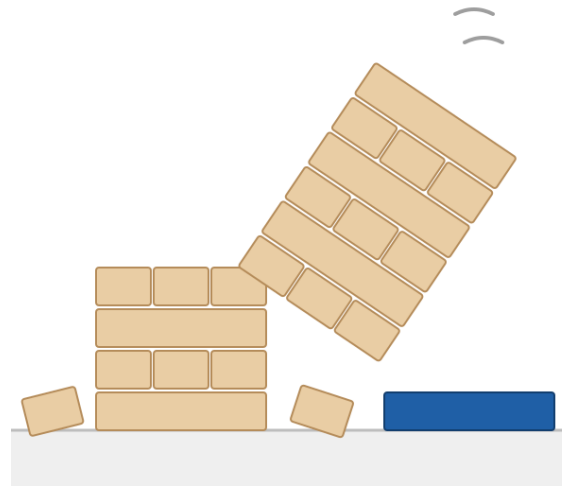
Has the concept leaked?

Surgicity

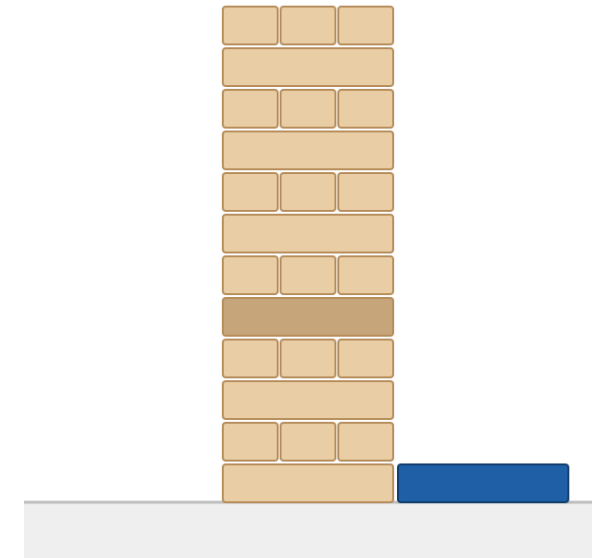
Are other concepts affected?



✗ Effective ✓ Surgical



✓ Effective ✗ Surgical



✓ Effective ✓ Surgical





1

Trusting the Erasure?

EMNLP 2018

2

Non-Linear Erasure

NeurIPS 2026

3

Erasure Evals

(under development)



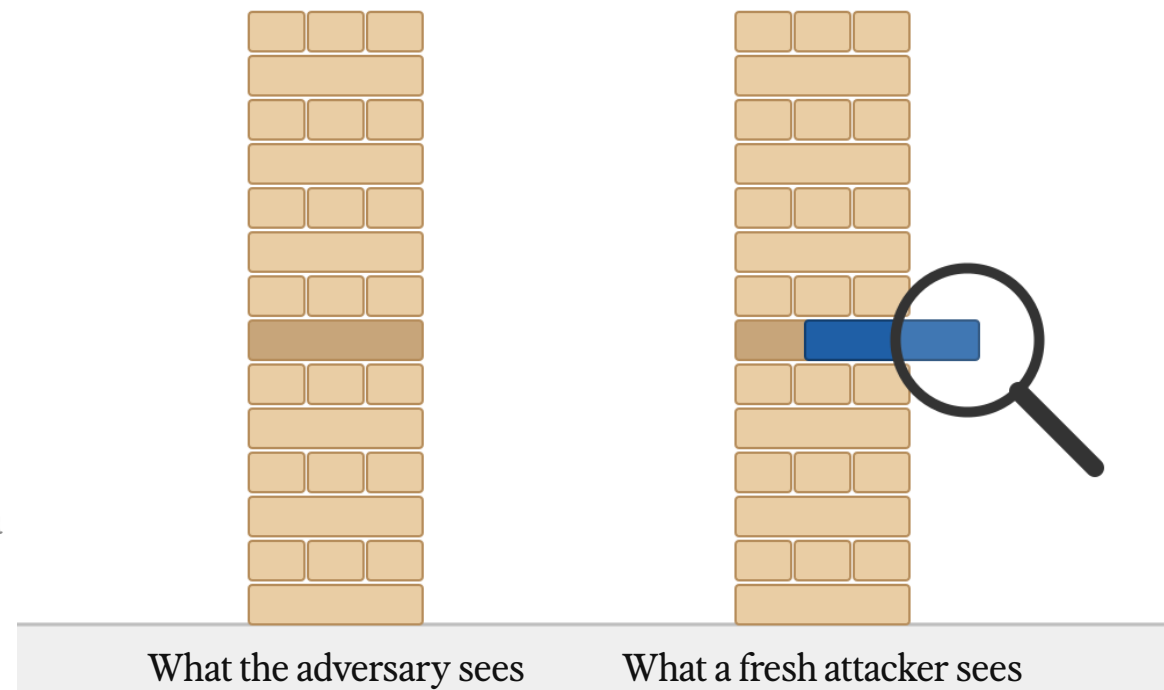
1

Don't trust the eraser

Adversarial Removal of Demographic Attributes from Text Data

Yanai Elazar & Yoav Goldberg

EMNLP 2018



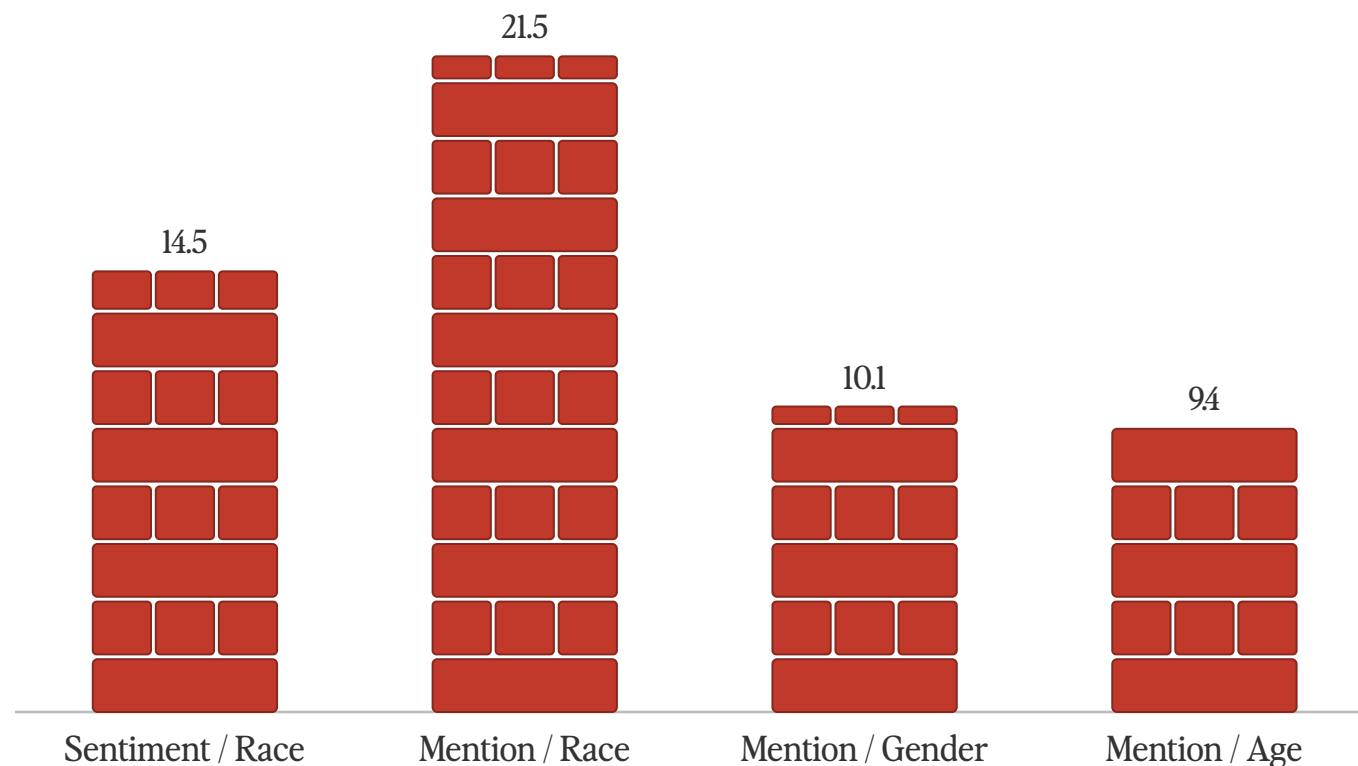
Even with Balanced Data, Demographics Leak

Setup

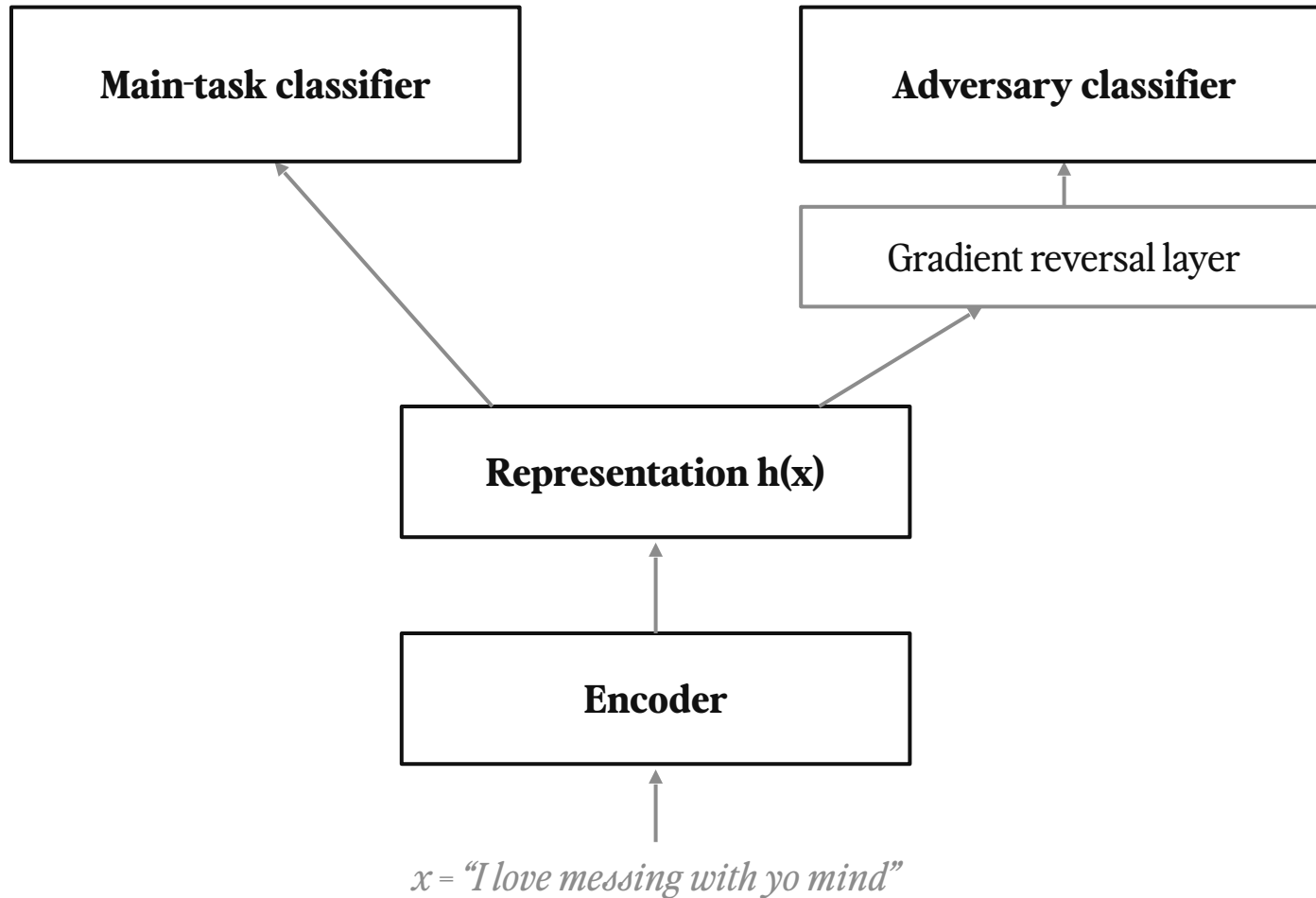
- Twitter data
- Main tasks: sentiment, mention detection
- Protected: race, gender, age
- 50/50 balanced on task $\times$ attribute
- Encoder trained only on the main task, then frozen and attacked



Leakage (points above chance), balanced training

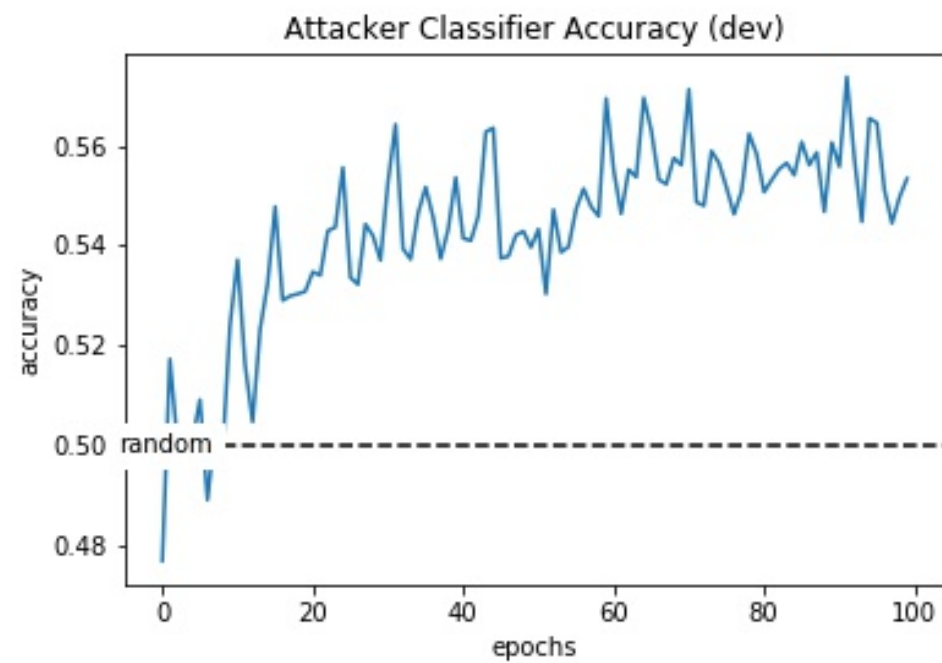
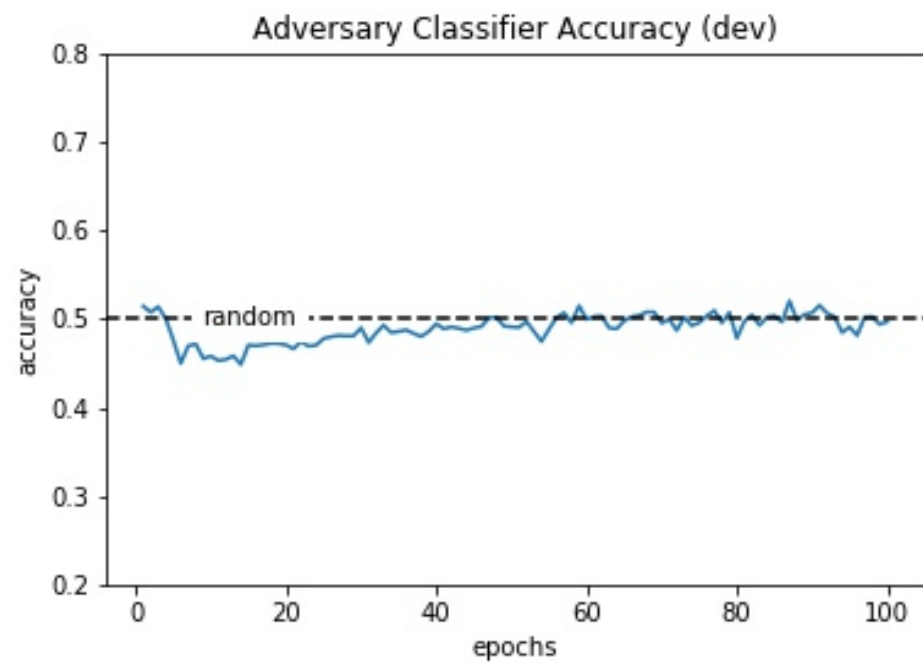


Adversarial Training: Hide the Attribute from an Adversary

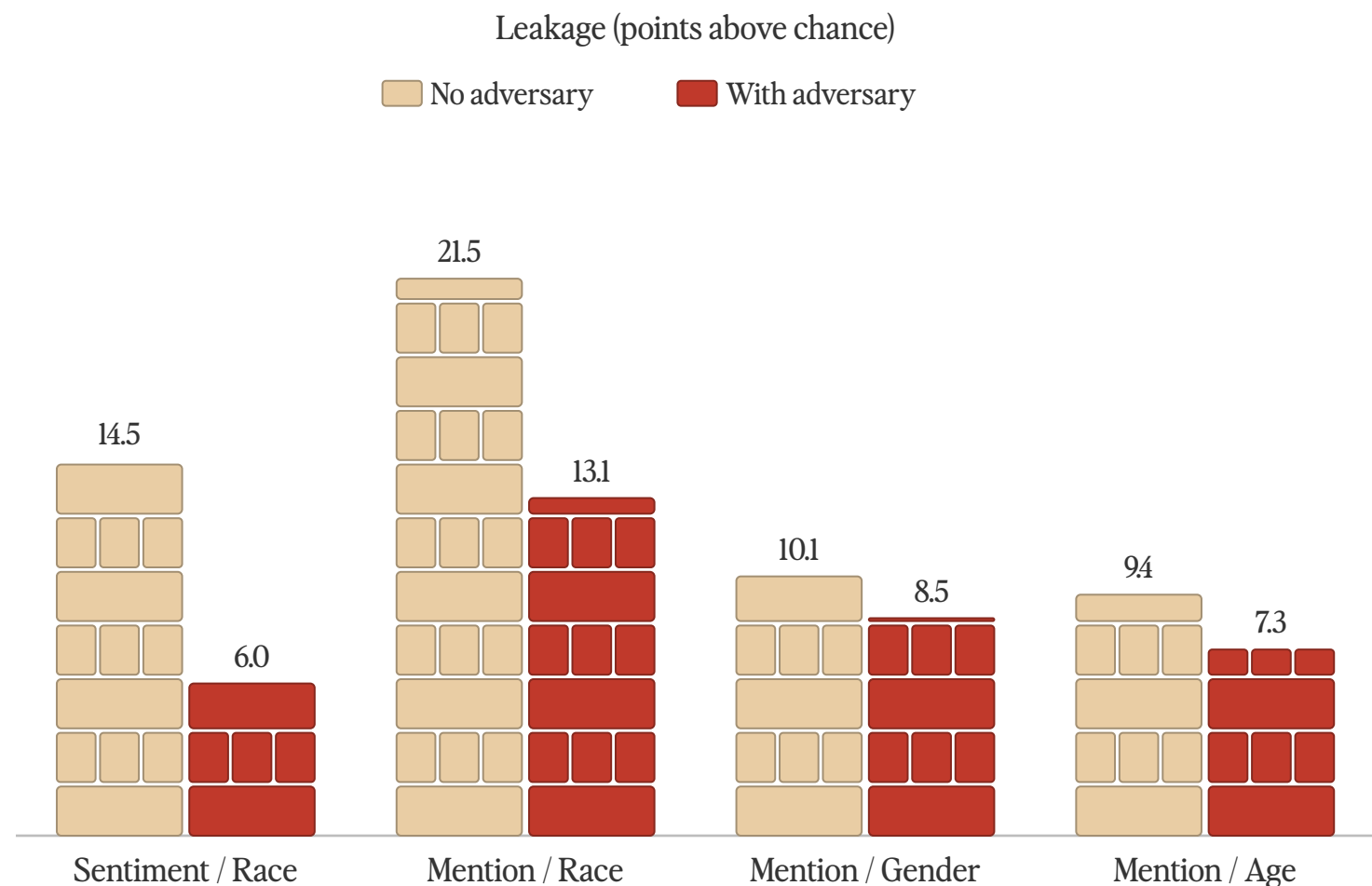


If adv is at chance, the attribute is gone. Or is it?





Adversarial Training Helps, but Never Reaches Chance



Leakage drops in every setting.

None reach chance

Stronger adversaries (more capacity, larger λ , ensembles) help more, but don't solve the issue



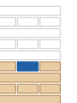
Intermediate Takeaway

- 1 Never trust the eraser's own score**
- 2 Test on data the eraser never saw**
- 3 This became the standard leakage protocol**

Effectiveness is only half the story.

Even a perfect eraser can destroy useful information on the way.

What else did we erase?



A Short History of Concept Erasure

2017

2018



NONLINEAR

NONLINEAR

**Controllable
invariance**

**Adversarial
removal**

We can erase concepts with unsatisfactory effectiveness



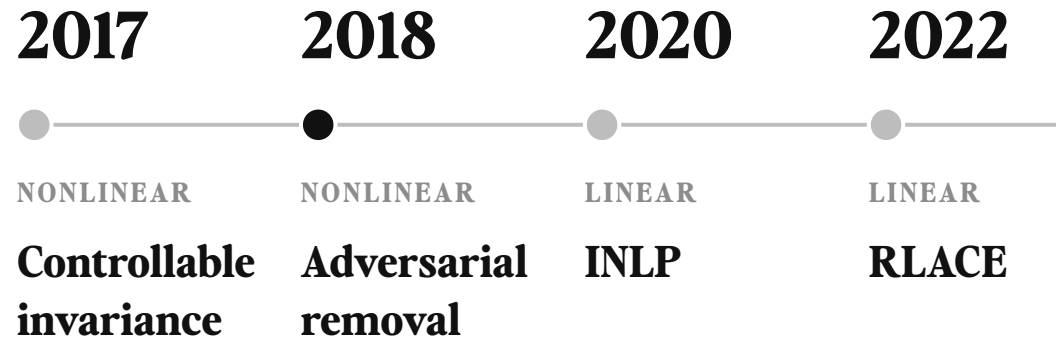
A Short History of Concept Erasure



We can remove concepts linearly!



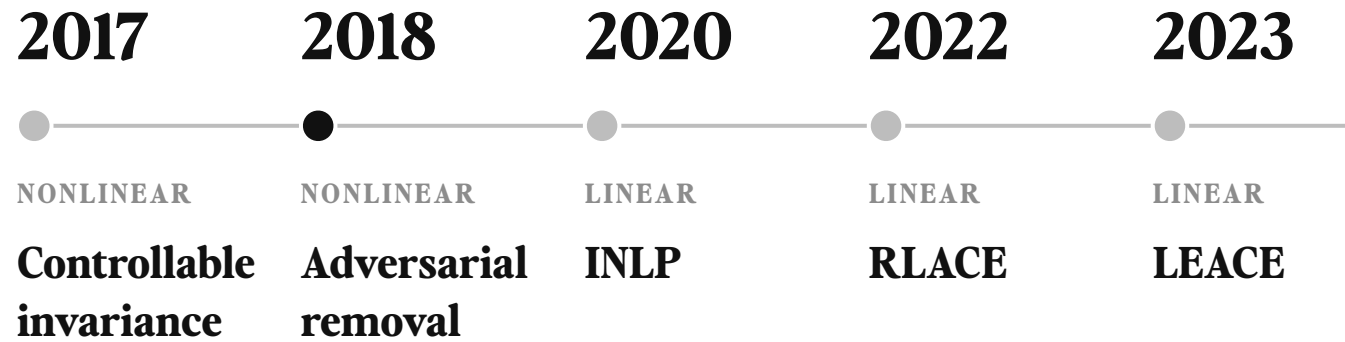
A Short History of Concept Erasure



Did we??? Perhaps now...



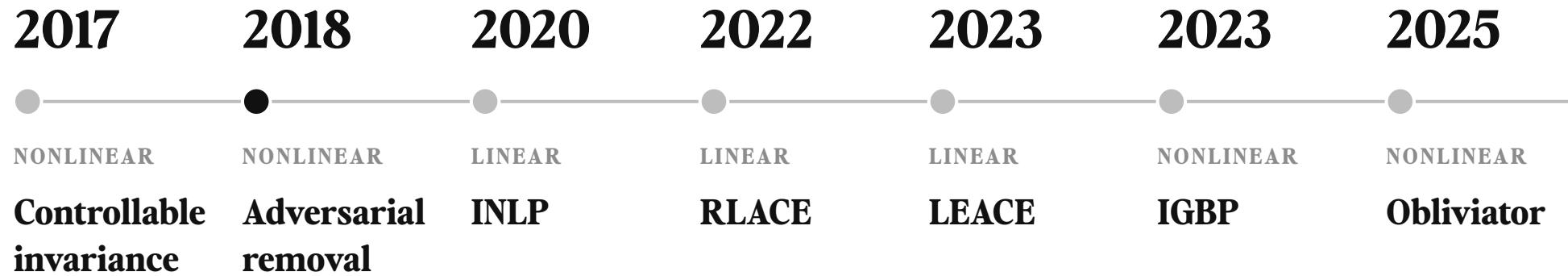
A Short History of Concept Erasure



Now for real...



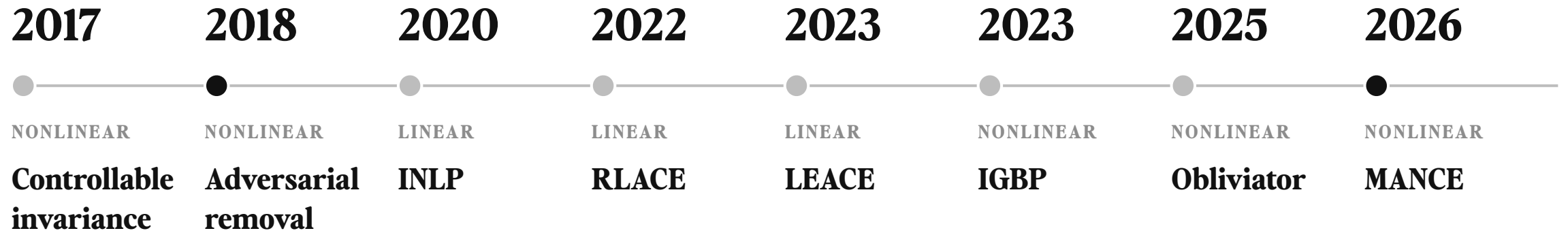
A Short History of Concept Erasure



Let's go back to non-linear concept erasure



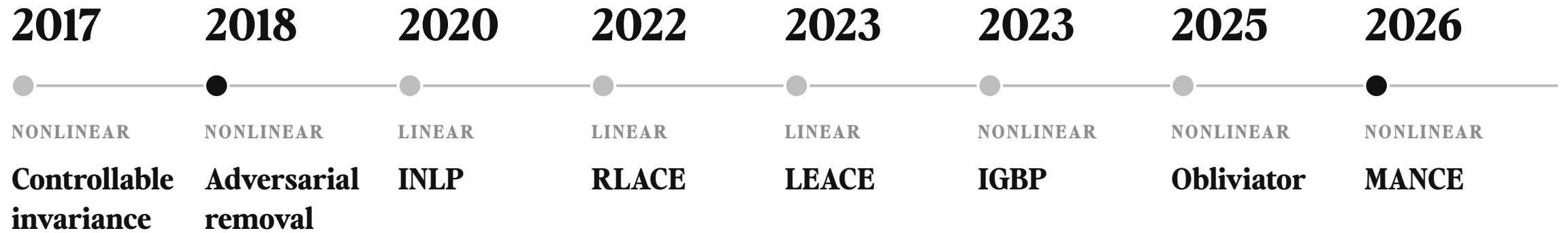
A Short History of Concept Erasure



And now even better



A Short History of Concept Erasure



Linear erasure is solved in closed form.
Nonlinear erasure is still an open problem



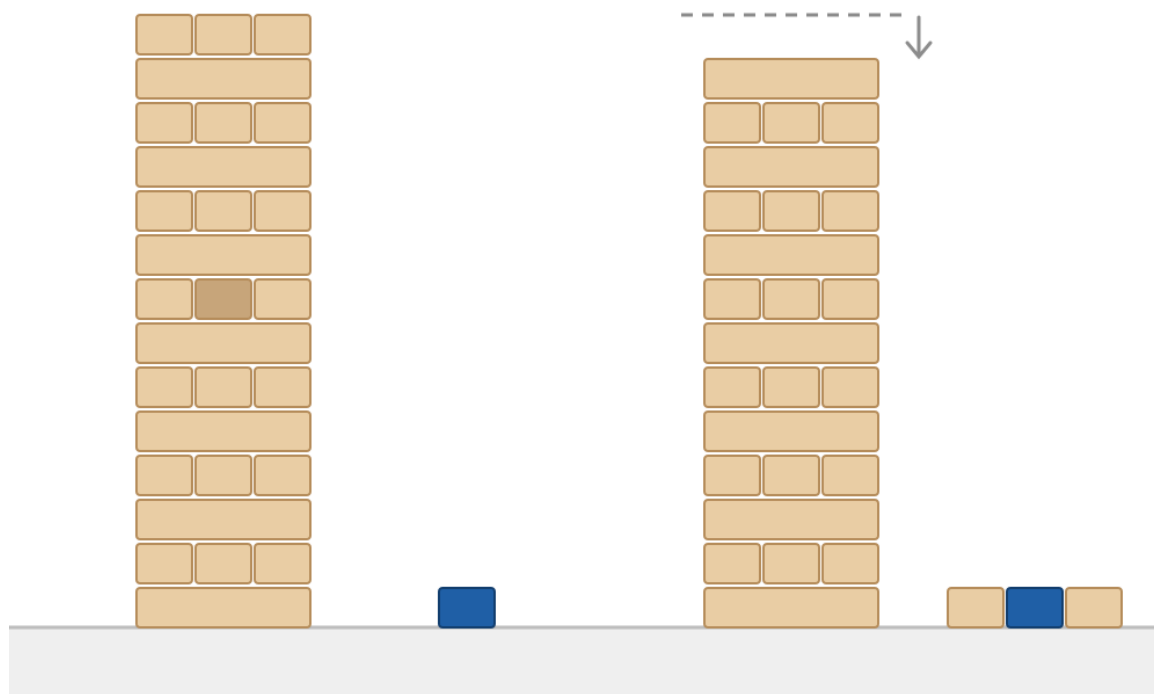
2

Erase Along the Manifold

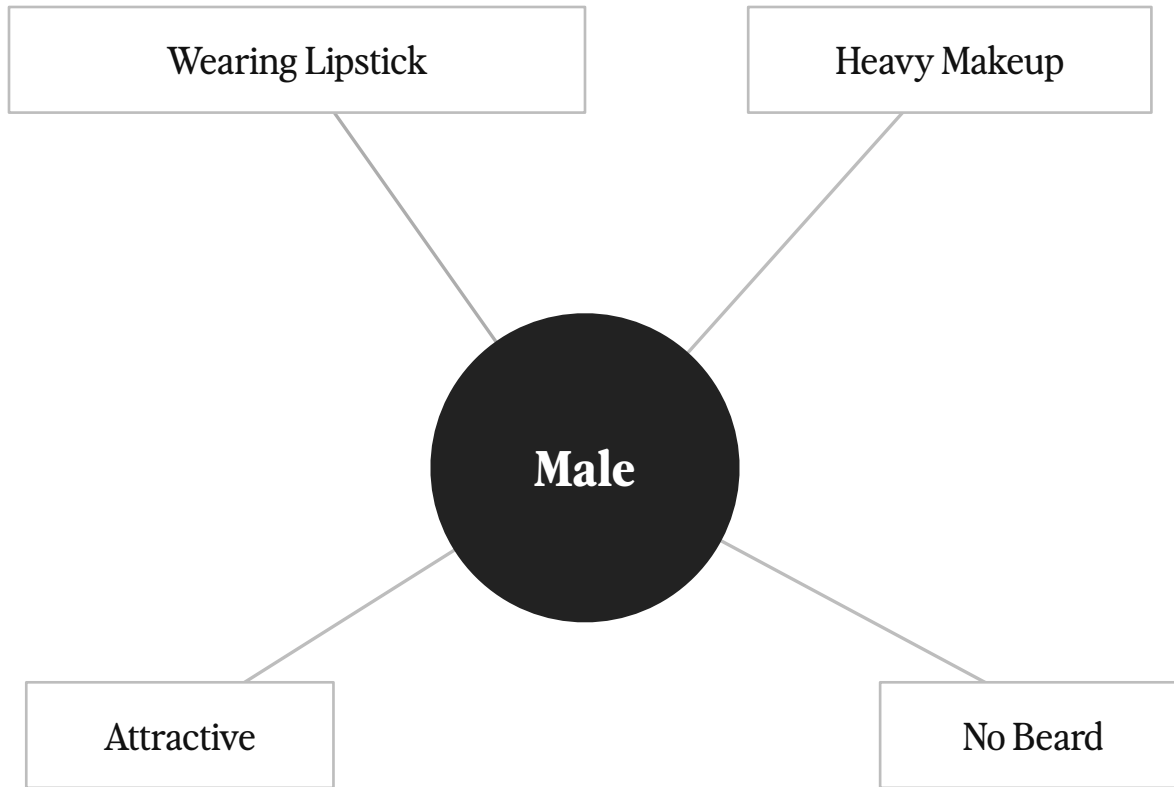
MANCE: Manifold Aware Concept Erasure

Matan Avitan, Yoav Goldberg, Yanai Elazar

NeurIPS 2026



Concepts are Entangled, and Most are Unlabeled



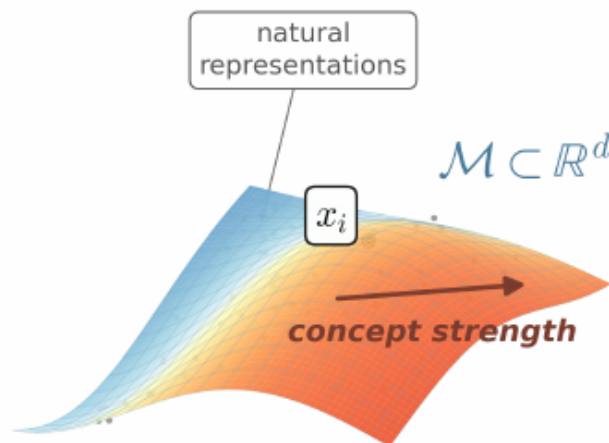
We need a constraint on the edit that does not require naming the controls.



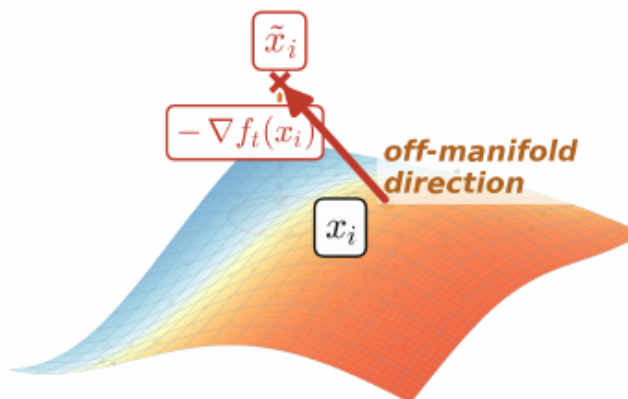
The Manifold Constraint Hypothesis (MCH)

Among edits with the same effect on the target, those that stay on the manifold of natural representations preserve more of everything else

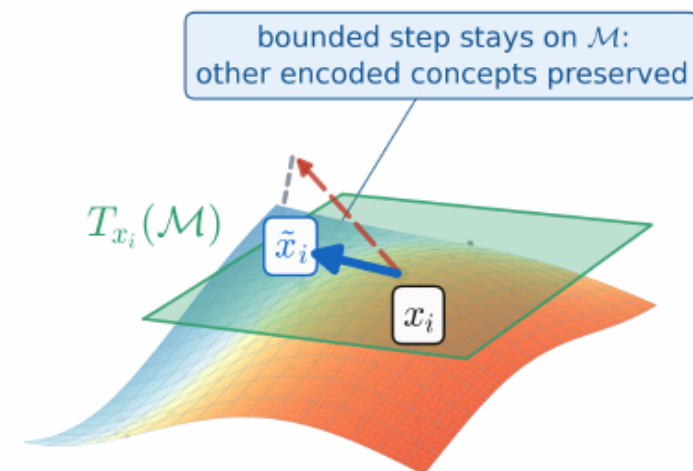
(A) The Representation Manifold



(B) Unconstrained Erasure



(C) MANCE — On-Manifold Edit



MANCE

1

Estimate the manifold

Take the k nearest natural (unedited) representations and estimate the tangent space of the manifold.

2

Project the probe's gradient

Take the gradient of a nonlinear probe for the target concept, and project it onto the tangent basis.

3

Limit the per-direction step

Cap the per-dimension update based on the original representation to keep the projection minimal.

We repeat this process for m rounds and refit the probe every k rounds

MANCE+: LEACE, then MANCE

MANCE++: LEACE + rank-2 CovMatch, then MANCE



Evaluated on 119 Settings across Text and Vision

14 models (Qwen2.5, Gemma 2&3, Llama 3.2, Mistral & *CLIP*)

3 NLP concepts + 40 vision concepts

A total of 119 settings



Metrics

Leakage

Accuracy of a new MLP probe above the majority class

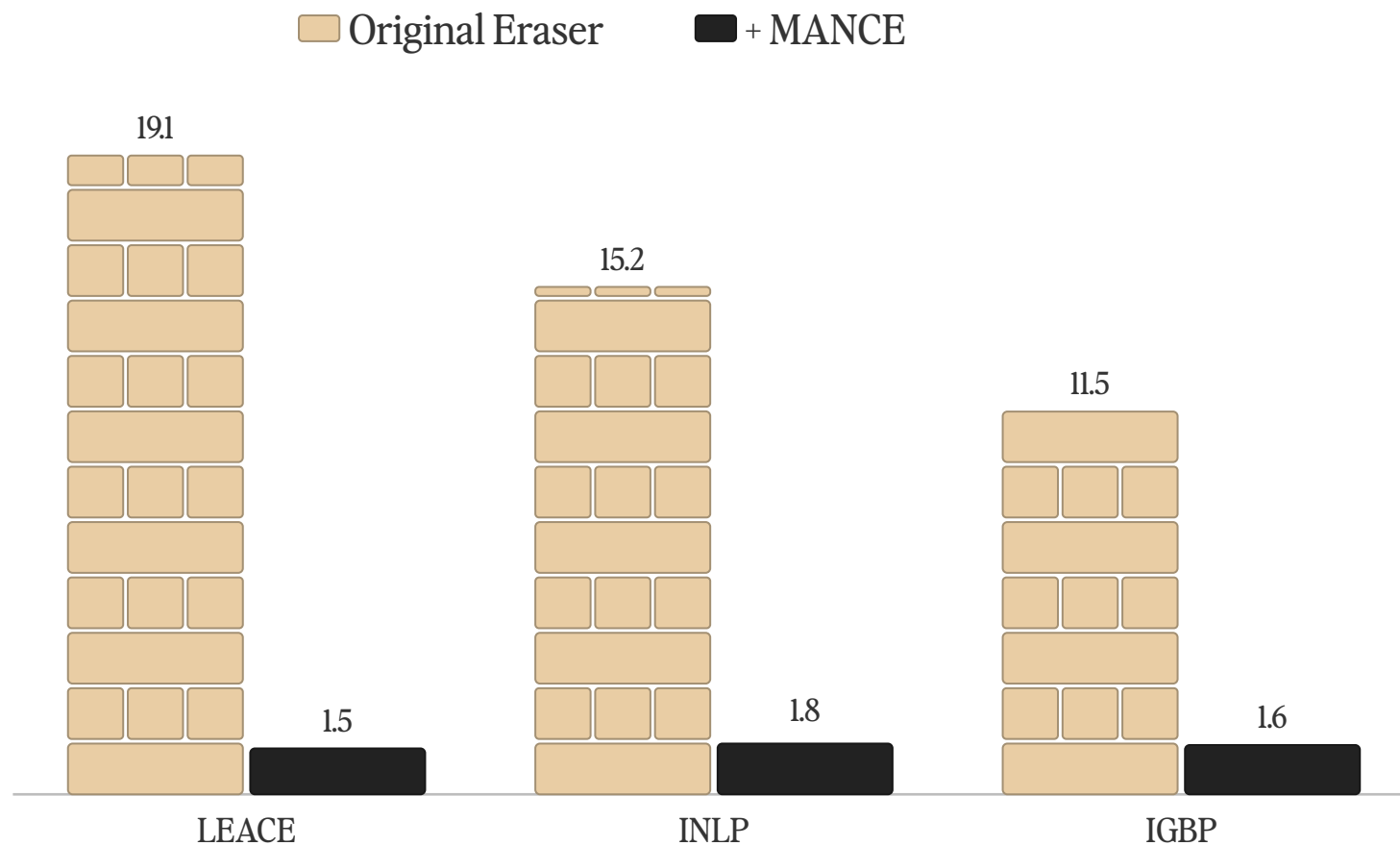
Surgicity budget

Accuracy drop of at most k points (out of 1, 3, 5, or 10)

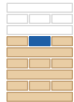
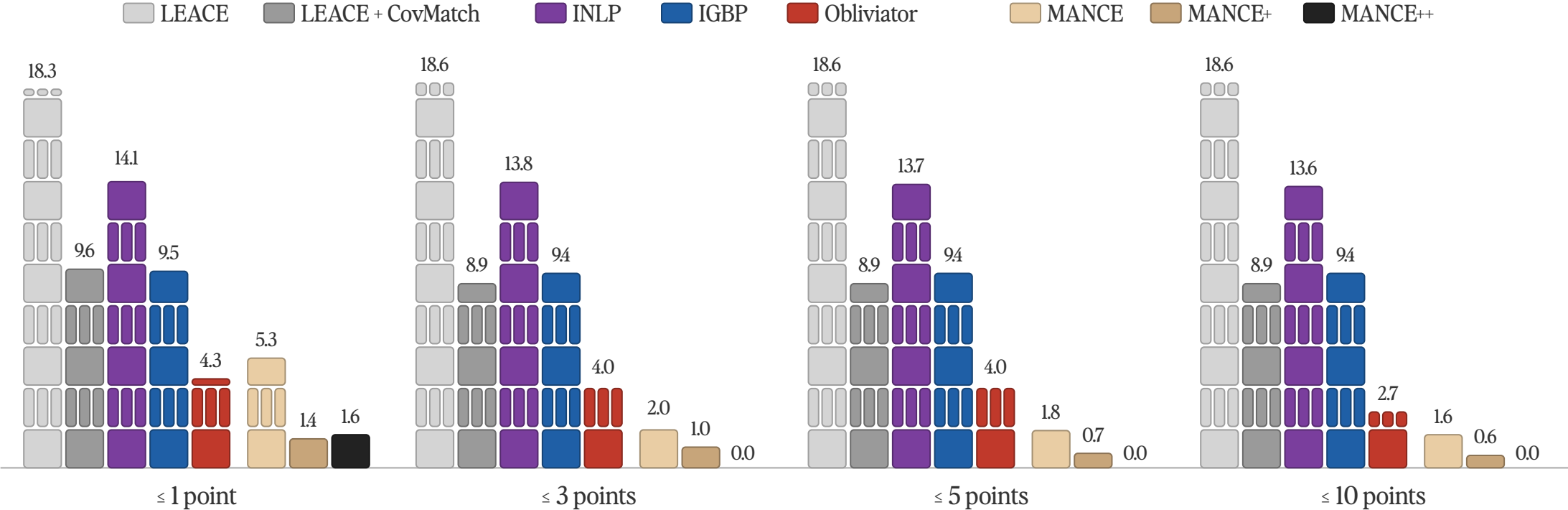


Results

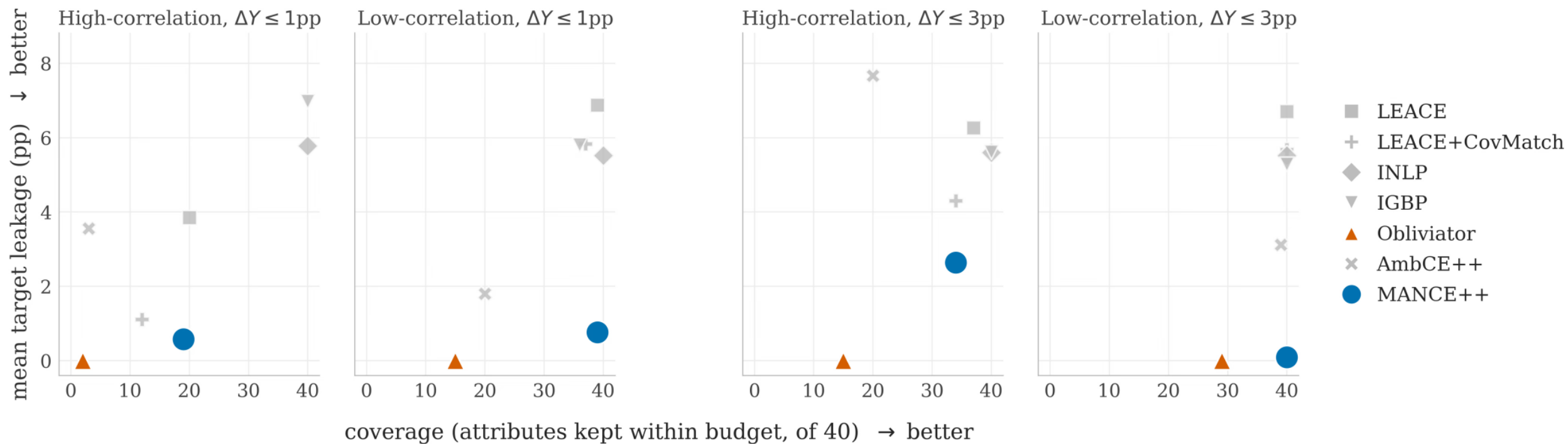
Adding MANCE on Top of Other Erasures



Leakage Results



Vision Results



MANCE Takeaways

1 Manifolds are useful

2 We have a new SOTA, use it!

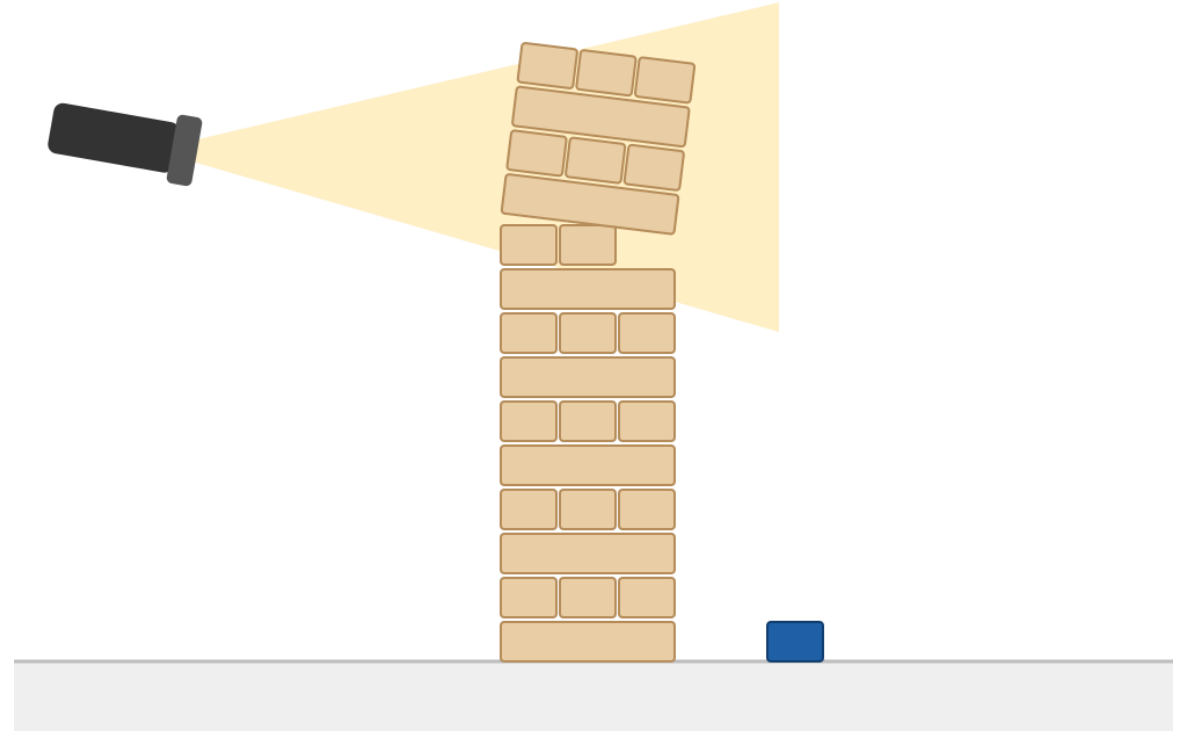


3

What aren't we measuring?

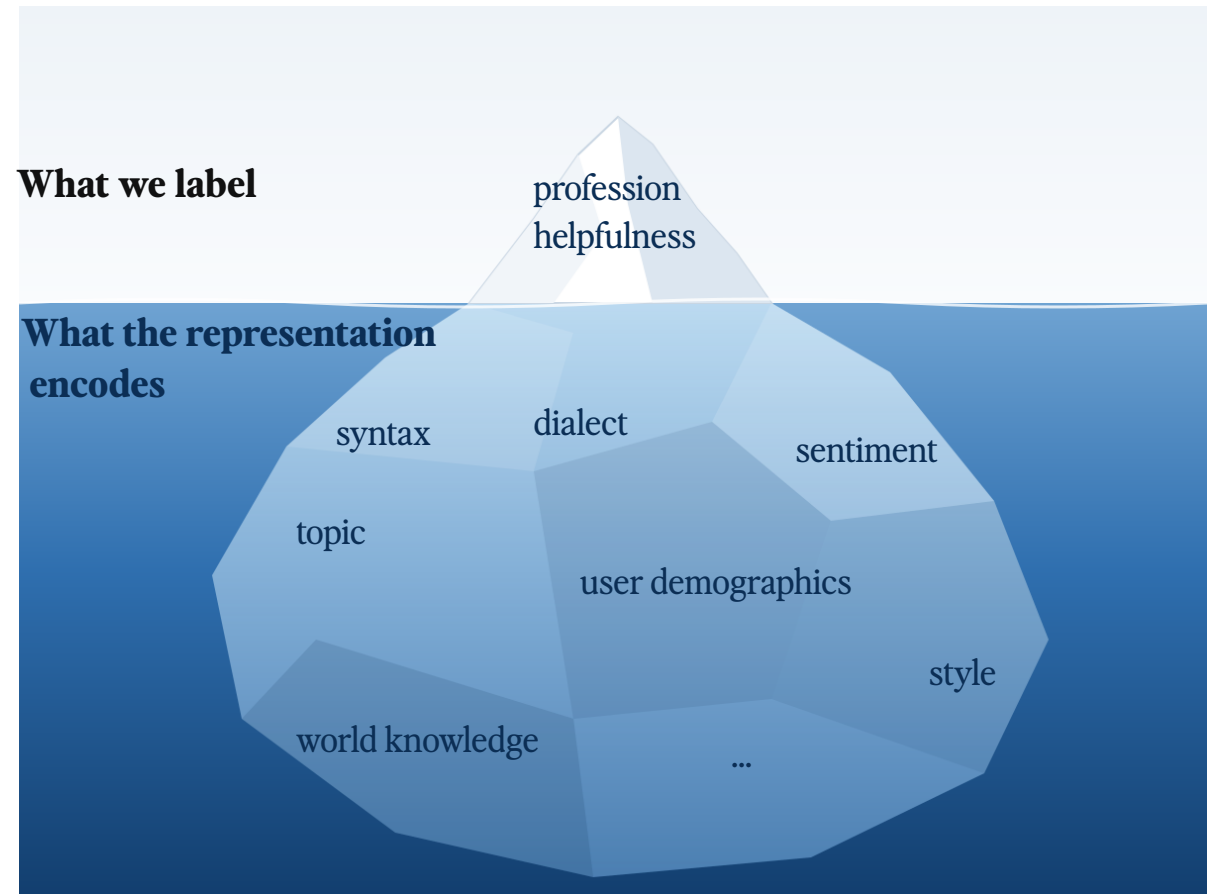
SurgicityLens: label-free surgicality evaluation with sparse autoencoders

ongoing



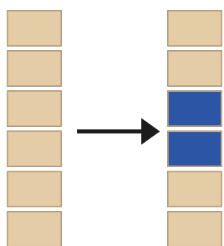
Limitation of Surgicity Evaluation

Surgicity only measures the control concepts we annotate!



SurgicityLens

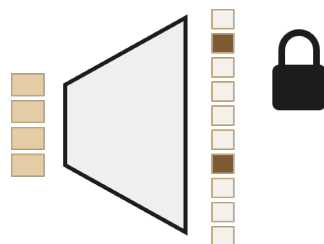
1



Pair inputs

clean and intervened

2



Decompose

one frozen SAE

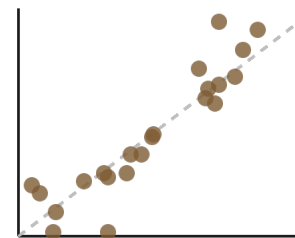
3



Drop target features

found with target labels

4



Compare patterns

per feature, across examples

5



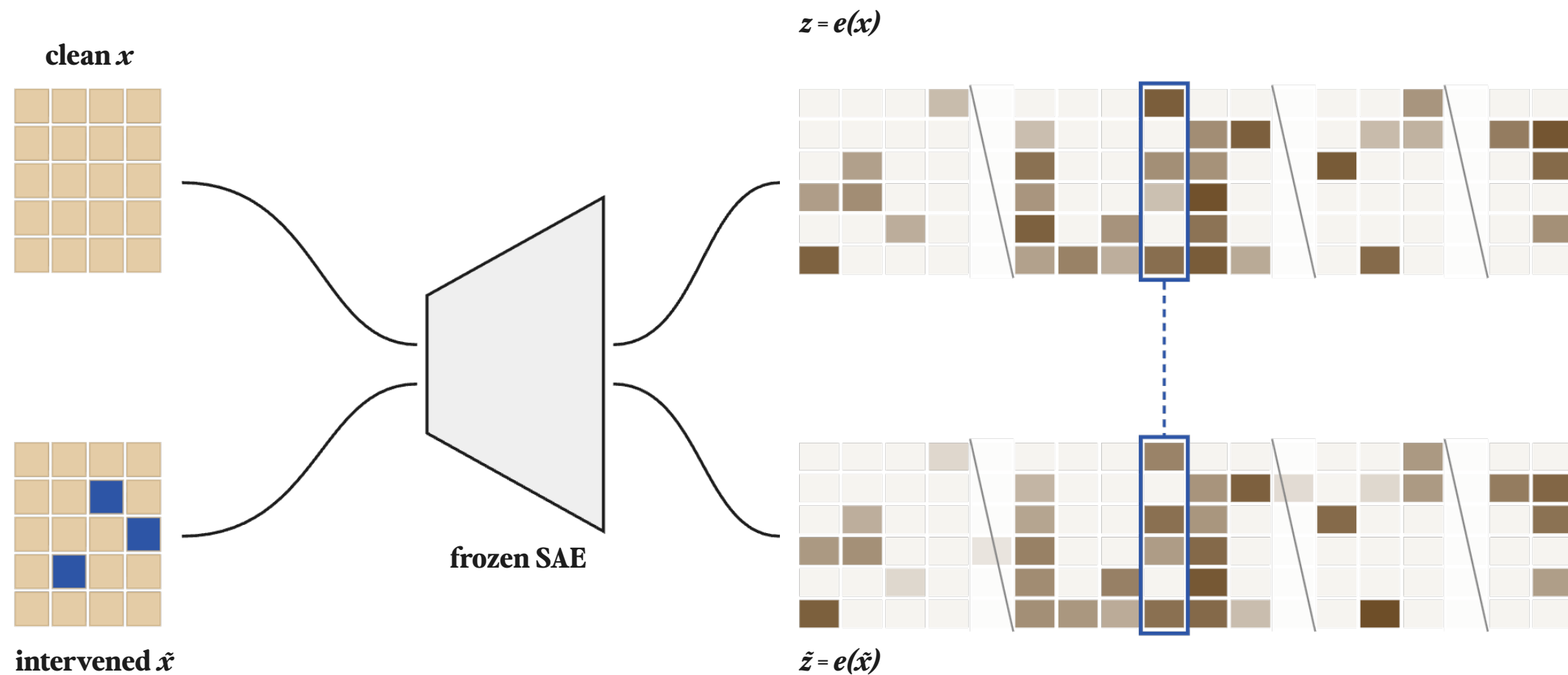
Score

mean of $1 - r$; lower is more surgical

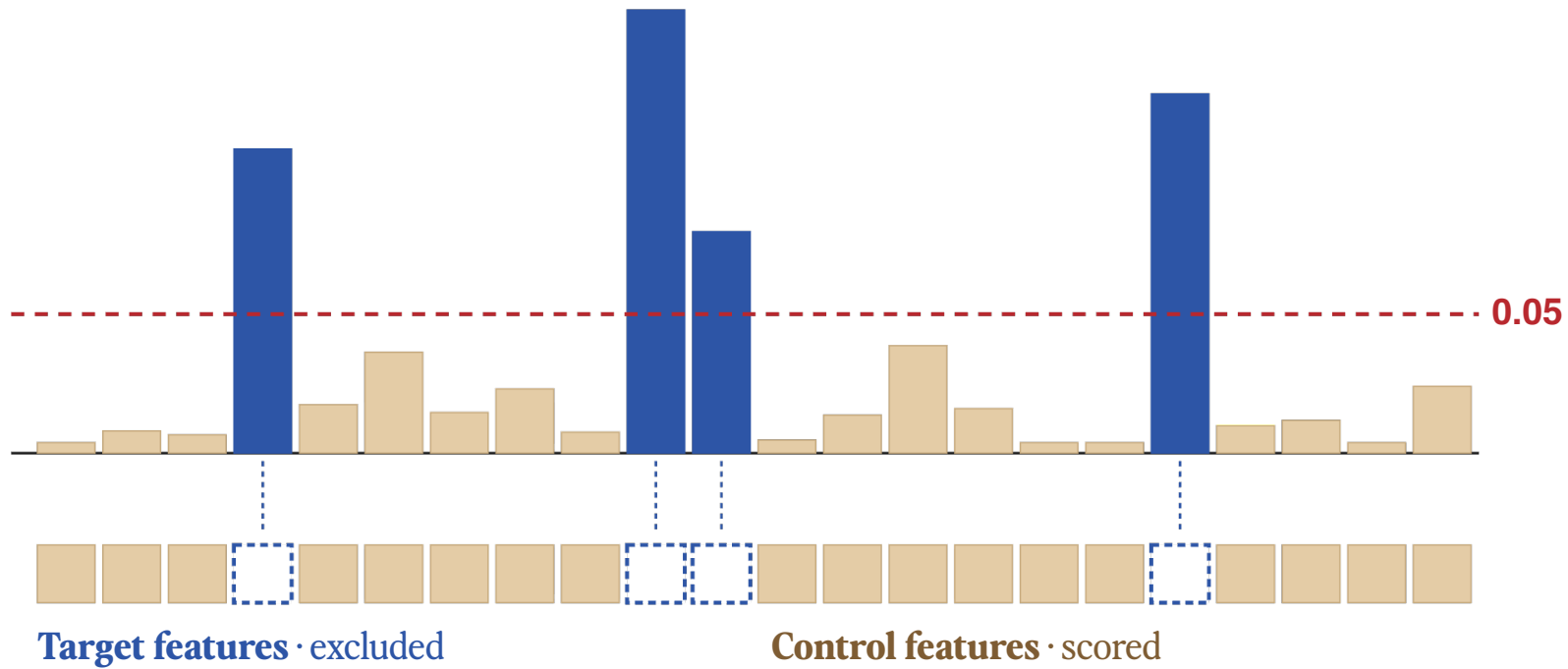
Only target labels are needed. No control-concept labels.



One Frozen SAE Reads both Sides

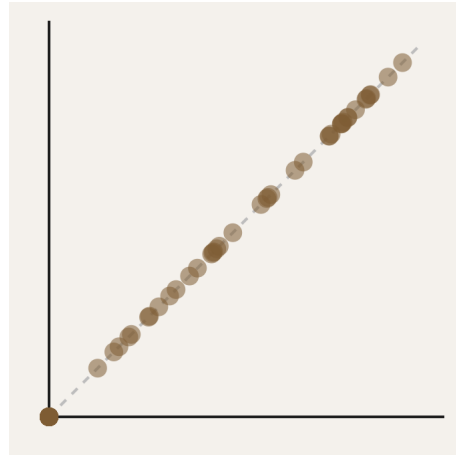


Only Target Labels are Needed

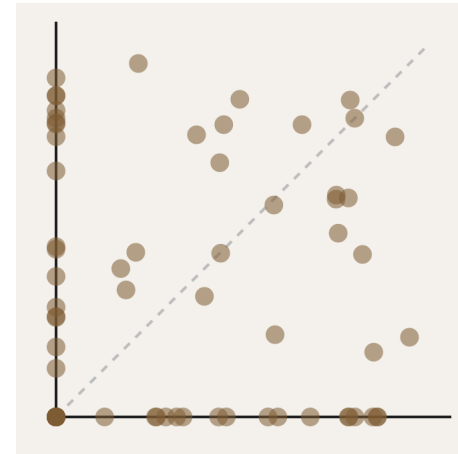


Correlated Neurons?

Surgical

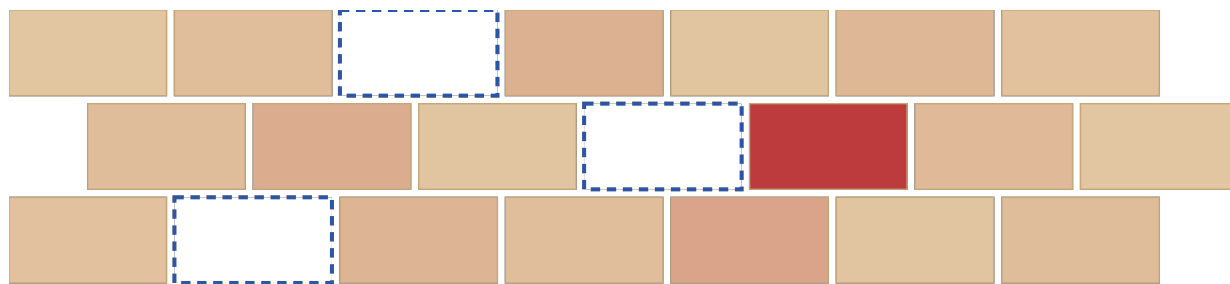


Unsurgical



SurgicityLens Score

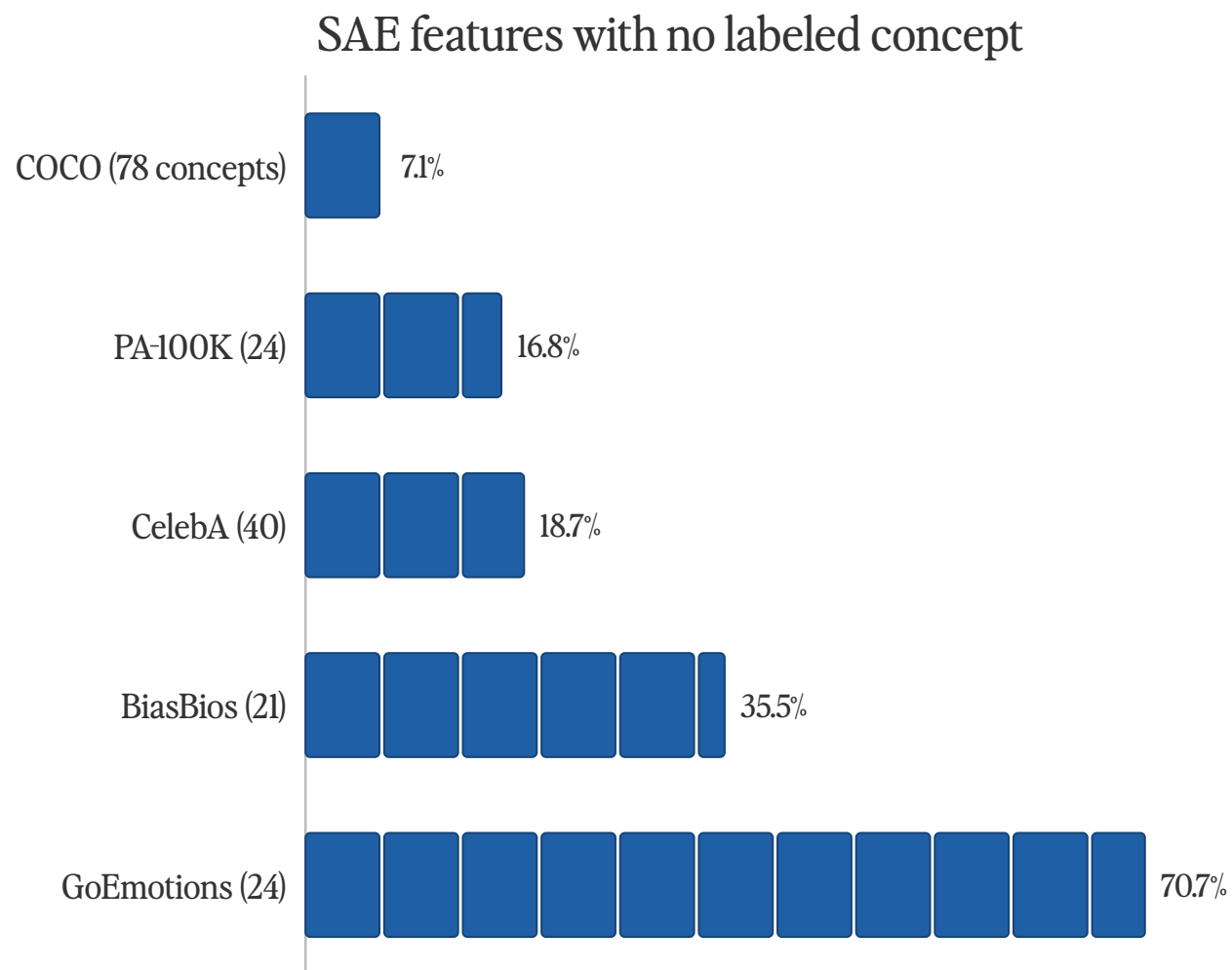
$$S_F = \text{Avg}_{j \in Fctrl} (1 - r_j)$$



Lower = more surgical

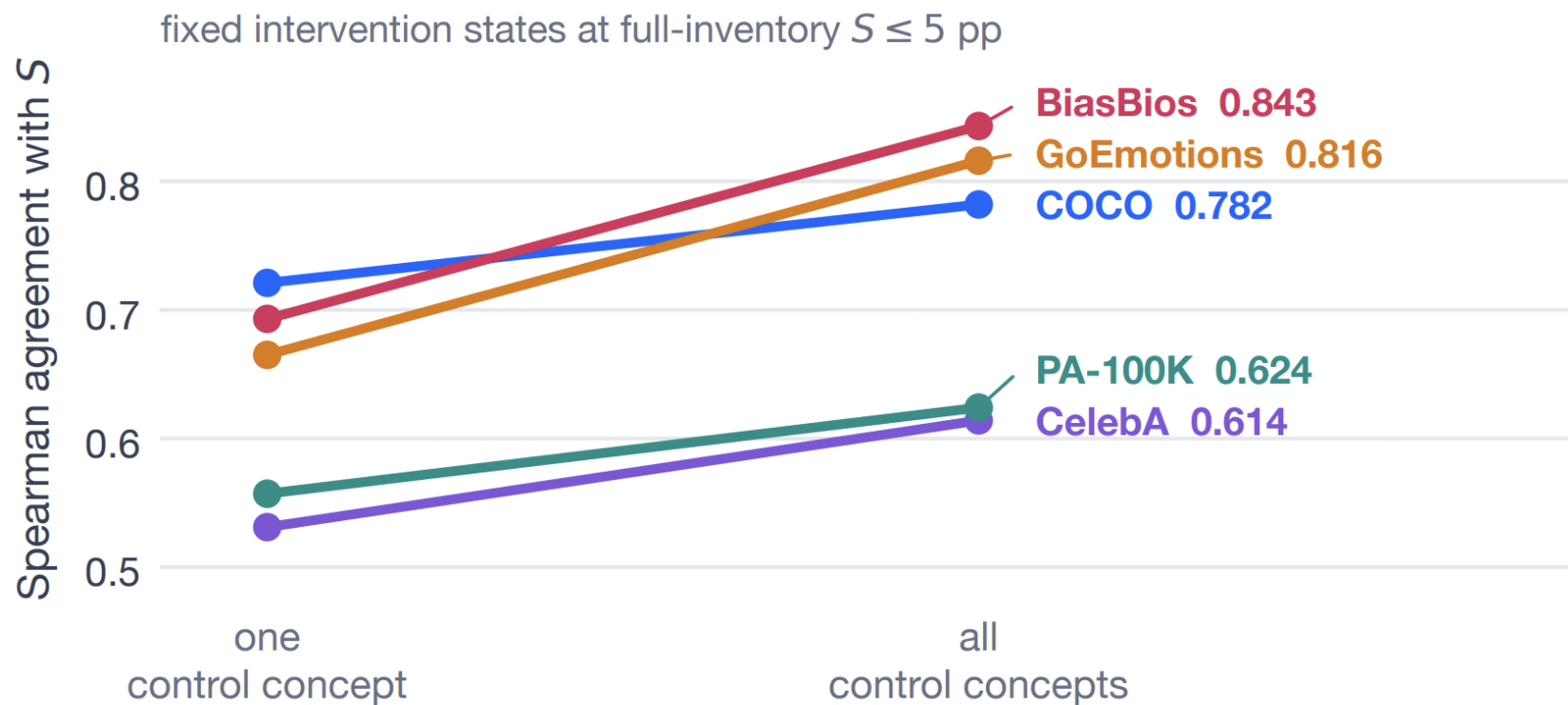


SAE Features Contain Many Unlabeled Concepts



Agreement with More Concepts

Agreement increases with more control concepts



Unmatched Ranking on Surgicity

Eraser

LEACE

INLP

MANCE⁺⁺

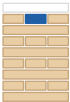
MANCE⁺

MANCE



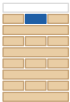
Unmatched Ranking on Surgicity

Eraser	SurgicityLens rank
LEACE	1
INLP	2
MANCE ⁺⁺	3
MANCE ⁺	4
MANCE	5



Unmatched Ranking on Surgicity

Eraser	SurgicityLens rank	Labeled-control rank
LEACE	1	2
INLP	2	1
MANCE ⁺⁺	3	5
MANCE ⁺	4	4
MANCE	5	3



Summary

1

**Always Verify
Effectiveness**

2

Manifolds are Useful!

3

**SAEs for Surgicity
Evaluation**



Thank you



yanaiela@gmail.com



@yanaiela

