GuideLabs

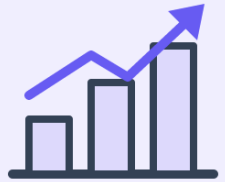
Interpretability by Design

Aya Abdelsalam Ismail

Chatbot

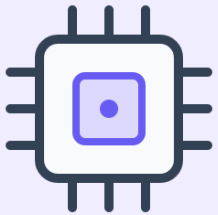
Self-driving

Medical imaging

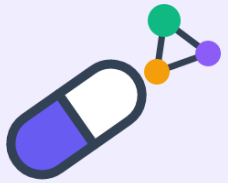


Stock trading

AI assistant

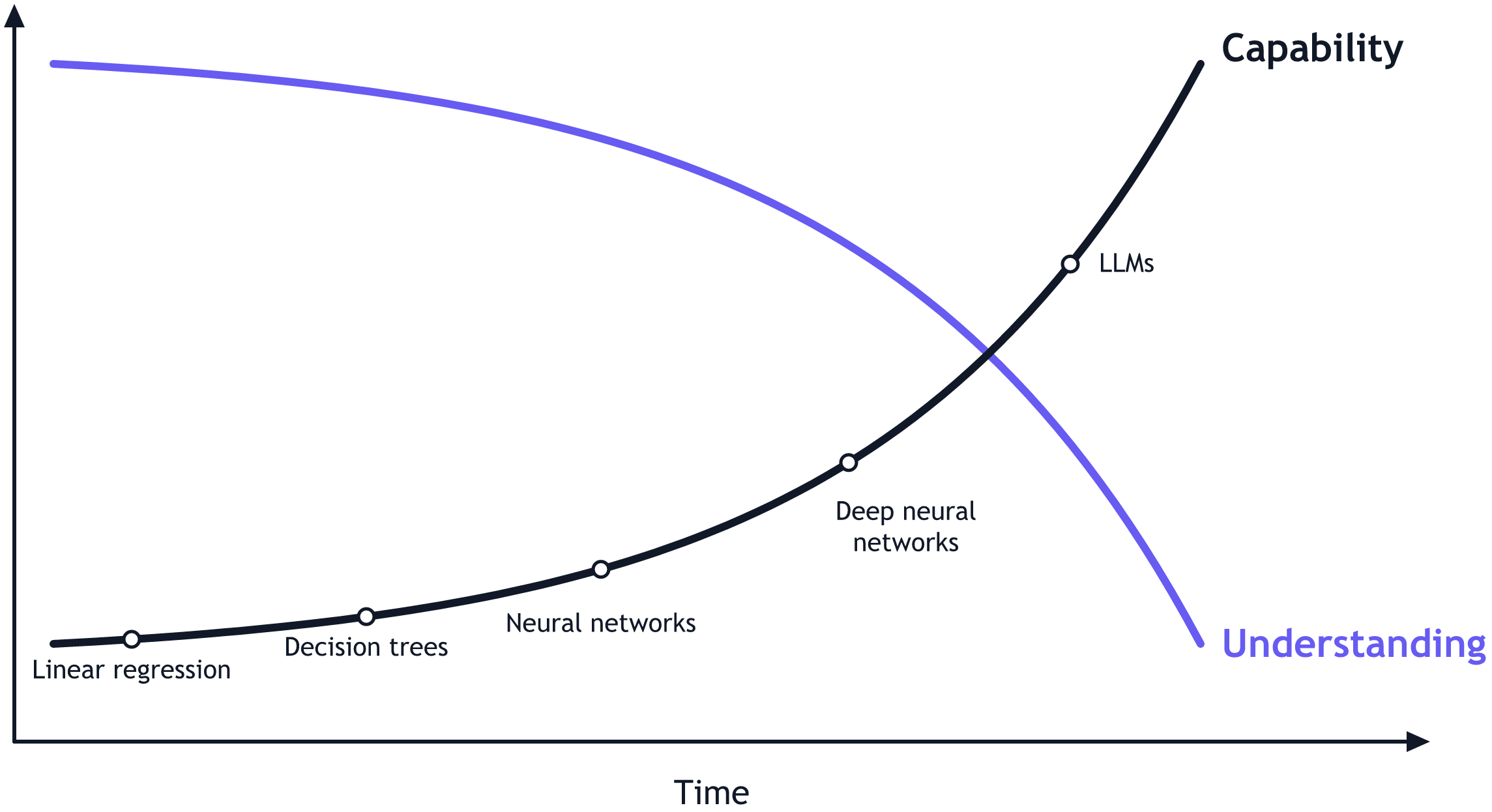


Chip design



Drug discovery

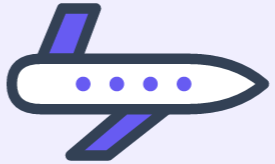
Weather forecast


Capability
LLMs
Deep neural networks
Neural networks
Decision trees
Linear regression
Understanding
Time

AI

?

Medicine



Airplanes

Concrete

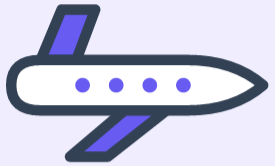
Batteries

AI

Do we need to understand AI?

Do we need to understand AI?

Medicine



Airplanes

Concrete

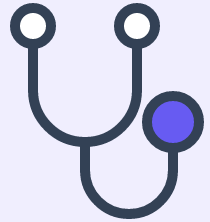
Batteries

These technologies have **earned** humans trust.

Do we need to understand AI?

Legal

Finance



Medical diagnosis

AI has **not** yet earned our trust

AI has not yet earned our trust

AI (artificial intelligence)

This article is more than **2 months old**

AI hallucinations found in high-profile Wall Street law firm filing

Sullivan & Cr
for string of e
Business l

Mark Sweney
Wed 22 Apr 2026 09.09 BST

Share

Prefer the Guardian on Google

AI (artificial intelligence)

This article is more than **1 year old**

High court tells UK lawyers to stop misuse of AI after fake case-law citations

Ruling follows two cases blighted by actual or suspected use of artificial intelligence in legal work

Robert Booth *UK technology editor*
Fri 6 Jun 2025 20.00 BST

Share

Prefer the Guardian on Google

US healthcare

This article is more than **1 year old**

New AI tool counters health insurance denials decided by automated algorithms

Class-action lawsuits allege algorithms turn down claims in seconds and critics say reform is needed for lasting change

The Guardian

Int

News | Opinion | Sport | Culture | Lifestyle

World | Europe | US news | Americas | Asia | Australia | Middle East | Africa | Inequality | Global development

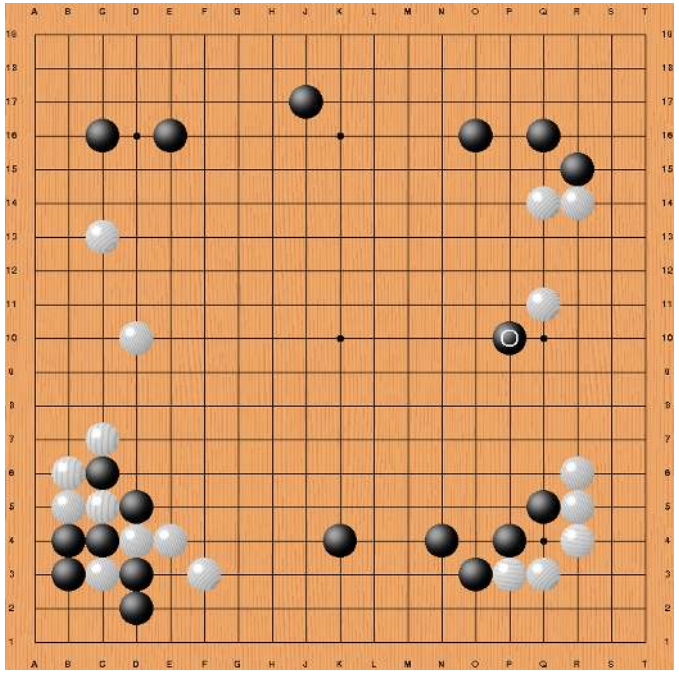
France

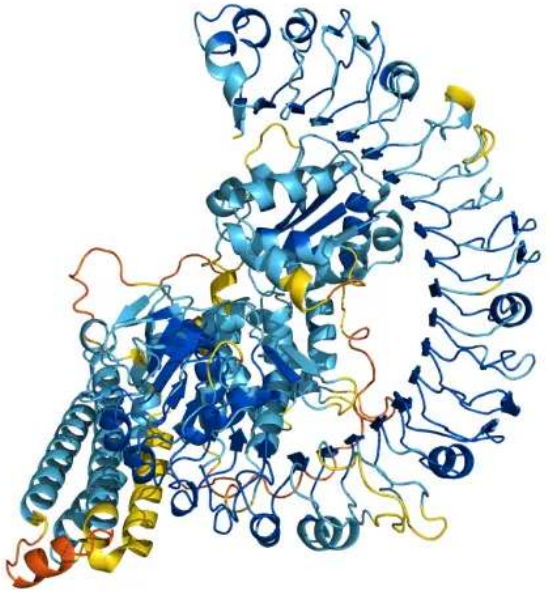
This article is more than **7 months old**

Facebook's job ads algorithm is sexist, French equality watchdog rules

Regulator found ads for mechanics skewed towards men while those for preschool teachers targeted women

Do we need to understand AI?

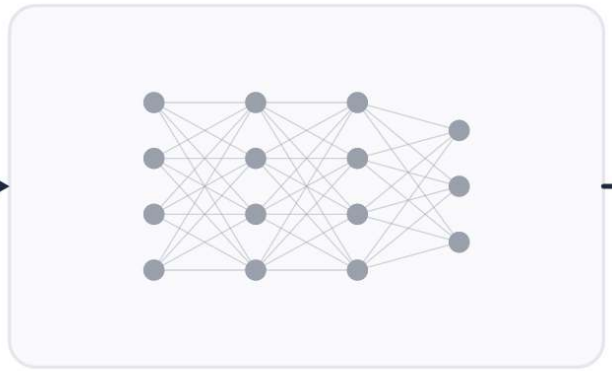
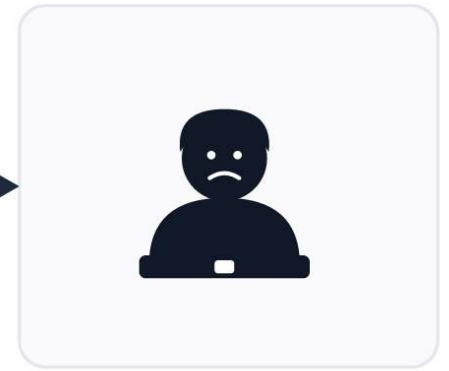




As AI becomes superintelligent, it would be invaluable if it could **teach us** what it has learned.

Standard training recipe

Standard

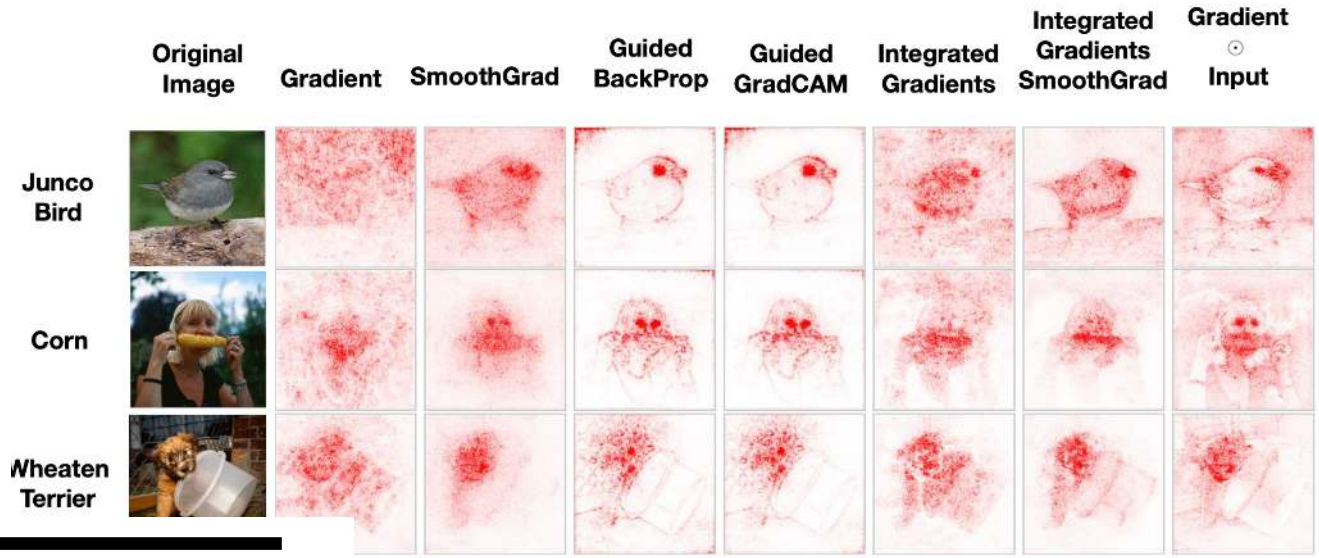
Data	Architecture	Training	Deployment
			
As much as possible	As big as possible	Pretraining, Midtraining, Long-context extension, SFT, Think, DPO/GRPO, RLVR, …	Let’s try to explain the model

Explaining any model post-hoc

- **Feature Attribution**

Explaining any model post-hoc

- Feature Attribution



Sanity Checks for Saliency Maps

Julius Adebayo,* Justin Gilmer♯, Michael Muelly♯, Ian Goodfellow♯, Moritz Hardt♯†, Been Kim♯
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com
♯Google Brain
†University of California Berkeley

Impossibility Theorems for Feature Attribution

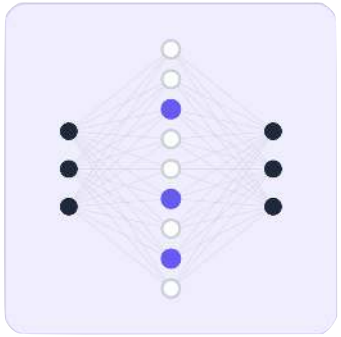
Blair Bilodeau*[1], Natasha Jaques[2,3], Pang Wei Koh[2,3], and Been Kim[3]

[1]University of Toronto
[2]University of Washington
[3]Google Deepmind

Explaining any model post-hoc

Feature Attribution

SAE

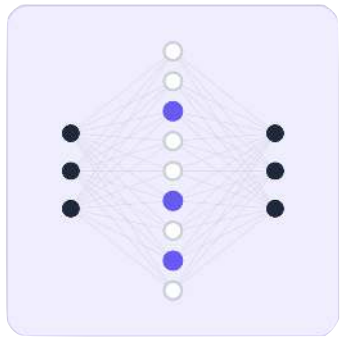


Sparse autoencoder

Explaining any model post-hoc

Feature Attribution

SAE



Sparse autoencoder

Interpretability Illusions with Sparse Autoencoders: Evaluating Robustness of Concept Representations

Aaron J. Li[1], Suraj Srinivas[2], Usha Bhalla[1], and Himabindu Lakkaraju[1]

AUTOMATED INTERPRETABILITY METRICS DO NOT DISTINGUISH TRAINED AND RANDOM TRANSFORMERS

Thomas Heap
University of Bristol
Bristol, UK
thomas.heap@bristol.ac.uk

Tim Lawson
University of Bristol
Bristol, UK

Lucy Farnik
University of Bristol
Bristol, UK

Laurence Aitchison
University of Bristol
Bristol, UK

Sanity Checks for Sparse Autoencoders: Do SAEs Beat Random Baselines?

Anton Korznikov * Andrey Galichin * Alexey Dontsov Oleg Y. Rogov Ivan Oseledets Elena Tutubalina

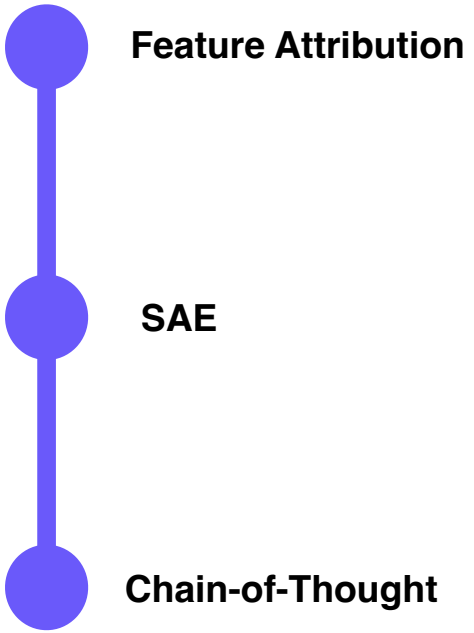
AXBENCH: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders

Zhengxuan Wu *[1] Aryaman Arora *[1] Atticus Geiger [2] Zheng Wang [1] Jing Huang [1]
Dan Jurafsky [1] Christopher D. Manning [1] Christopher Potts [1]

ON THE LIMITS OF SPARSE AUTOENCODERS: A THEORETICAL FRAMEWORK AND REWEIGHTED REMEDY

Jingyi Cui[1]* **Qi Zhang**[1]* **Yifei Wang**[2] **Yisen Wang**[1,3]†
[1]State Key Lab of General AI, School of Intelligence Science and Technology, Peking University
[2]Amazon AGI SF Lab‡ [3]Institute for Artificial Intelligence, Peking University

Explaining any model post-hoc



Explaining any model post-hoc

Feature Attribution

SAE

Chain-of-Thought

Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting

Miles Turpin,[1,2] **Julian Michael,**[1] **Ethan Perez,**[1,3] **Samuel R. Bowman**[1,3]
[1]NYU Alignment Research Group, [2]Cohere, [3]Anthropic
miles.turpin@nyu.edu

Chain-of-Thought Reasoning In The Wild Is Not Always Faithful

Iván Arcuschin[*]
University of Buenos Aires
iarcuschin@dc.uba.ar

Jett Janiak[*]

Robert Krzyzanowski[*]

Senthooran Rajamanoharan

Neel Nanda

Arthur Conmy
arthurconmy@gmail.com

Reasoning Models Don't Always Say What They Think

Yanda Chen **Joe Benton** **Ansh Radhakrishnan** **Jonathan Uesato** **Carson Denison**
John Schulman[+] **Arushi Somani**

Peter Hase[+] **Misha Wagner** **Fabien Roger** **Vlad Mikulik**
Samuel R. Bowman **Jan Leike** **Jared Kaplan** **Ethan Perez**

Alignment Science Team, Anthropic

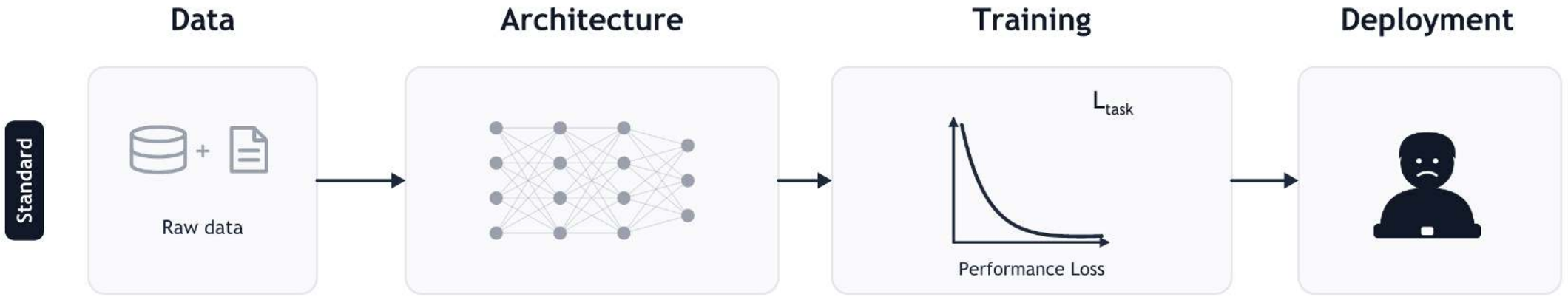
Why does this keep happening?

(a) World model **(b)** World model with noise **(c)** Transformer

Image from Vafa et al. (a) True Manhattan Map. (b) True map with noise perturbations (c) A transformer trained on turn-by-turn direction produces entangled, chaotic internal states, despite generating correct predictions. Capability and internal coherence are not the same thing.

We are trying to explain models that were not meant to be explained

Standard training recipe



As much as possible

As big as possible

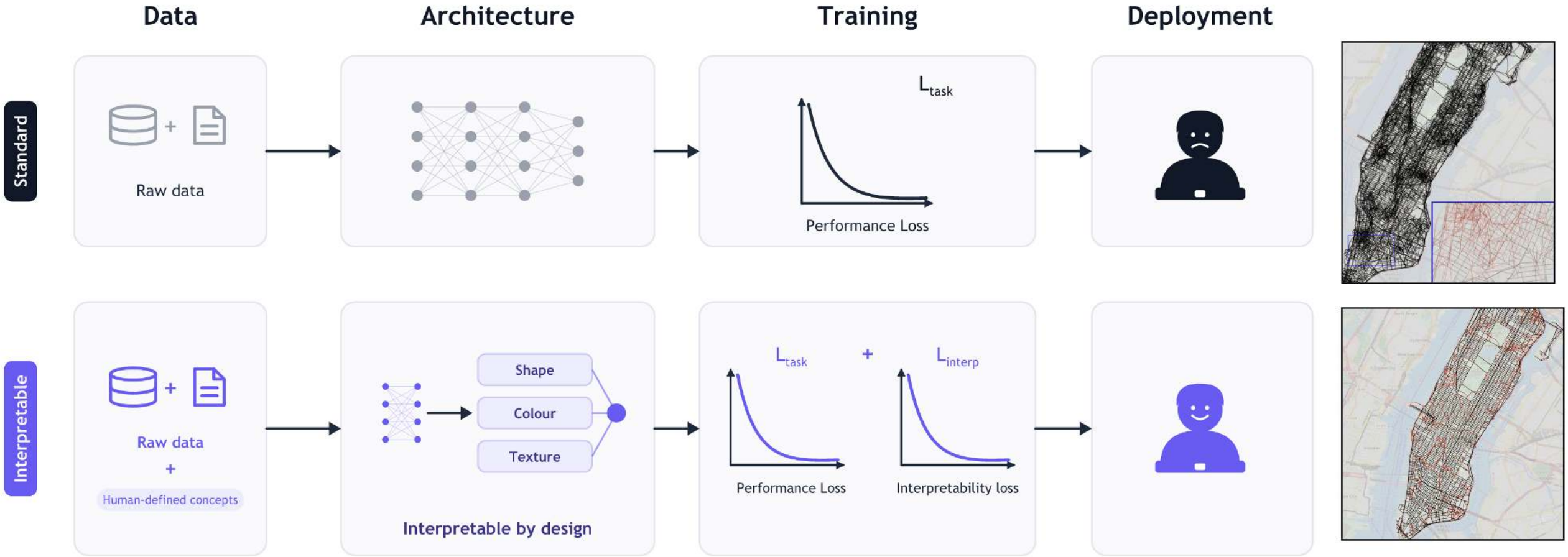
Pretraining, Midtraining, Long-context extension, SFT, Think, DPO/GRPO, RLVR, …

Lets try to explain the model

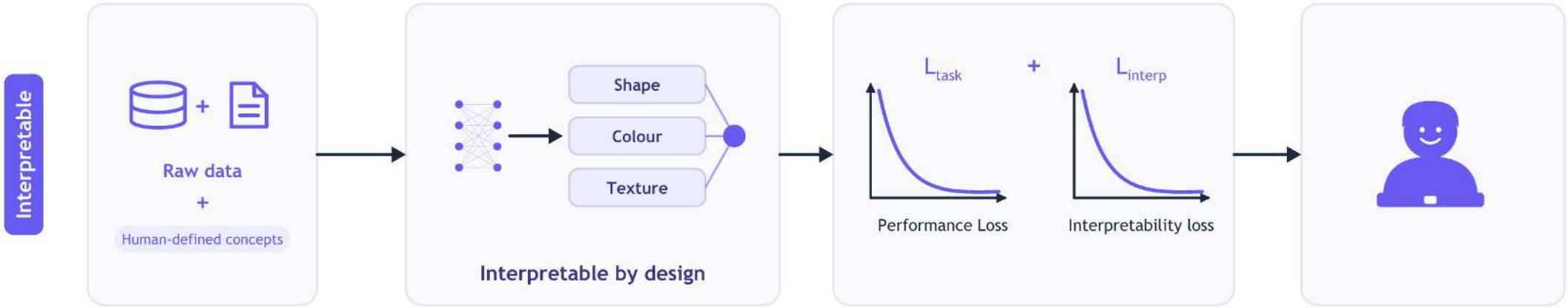
Capabilities emerge because we optimize for them.

Why expect interpretability when we never do?

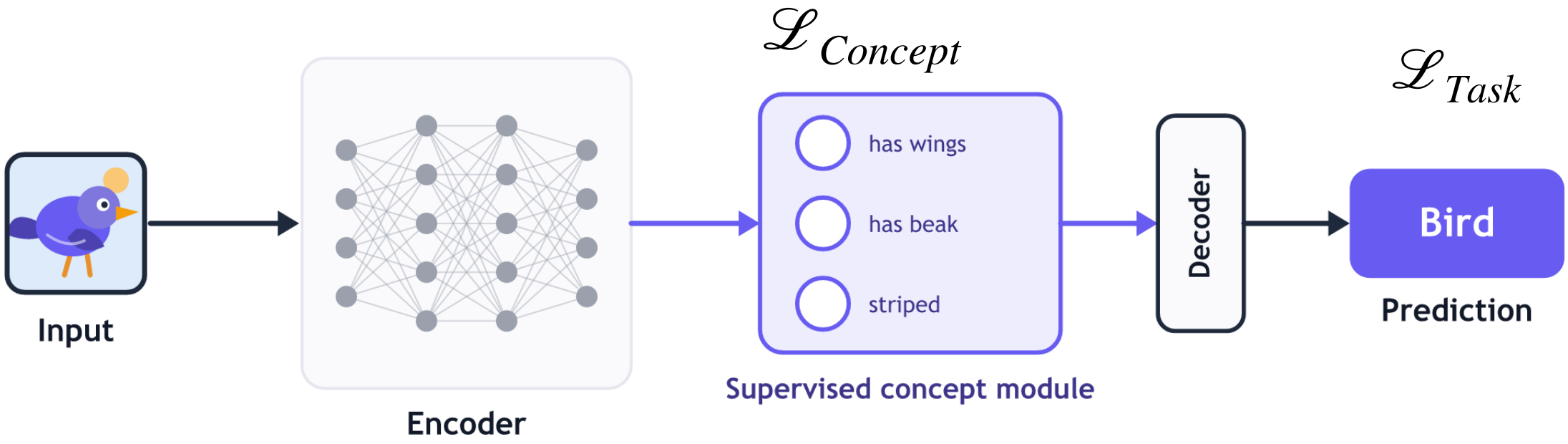
Interpretability at every step, by design



Interpretable training recipe



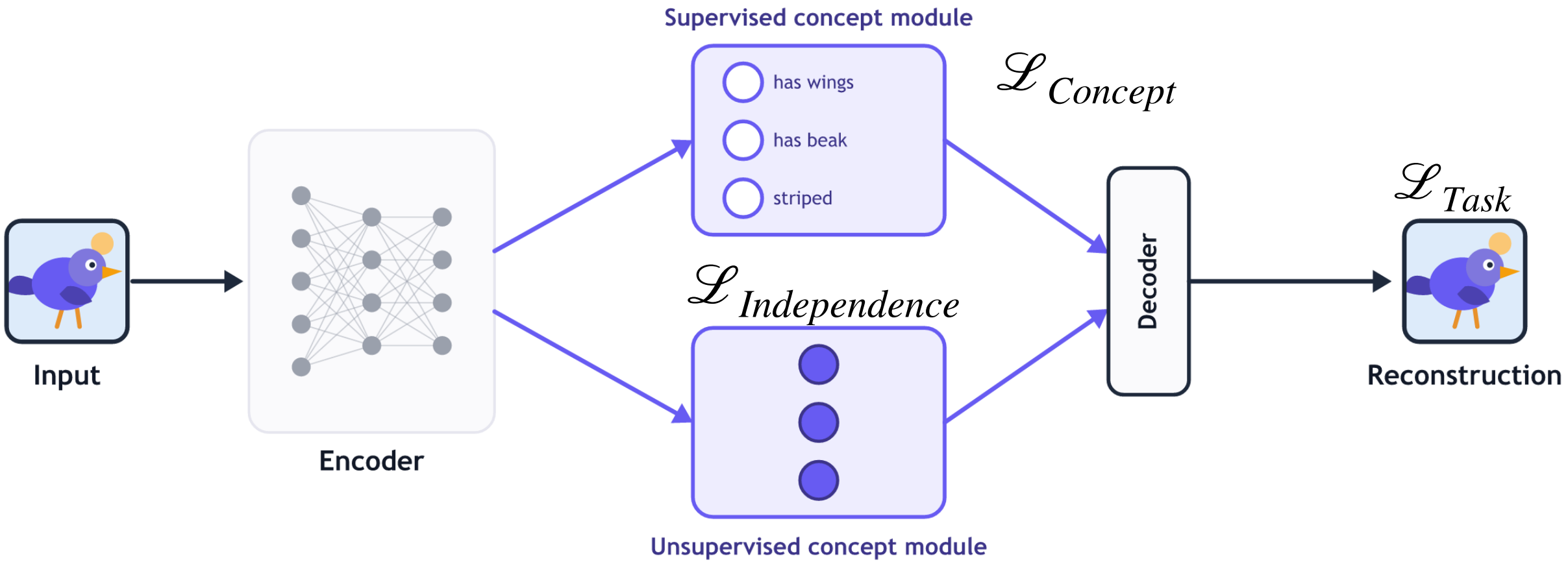
Concept bottleneck models



$$\mathscr{L} = \mathscr{L}_{Task} + \alpha \mathscr{L}_{Concept}$$

Koh et al., "Concept Bottleneck Models," ICML 2020.

Concept bottleneck generative models



$$\mathcal{L} = \mathcal{L}_{Task} + \alpha\mathcal{L}_{Concept} + \beta\mathcal{L}_{Independence}$$

Ismail et al., "Concept Bottleneck Generative Models," ICLR 2024.

Interpretable training recipe

Performance

Interpret & Intervene

Discover

Scaling

Interpretable training recipe

Does this training recipe work?

Interpretable training recipe

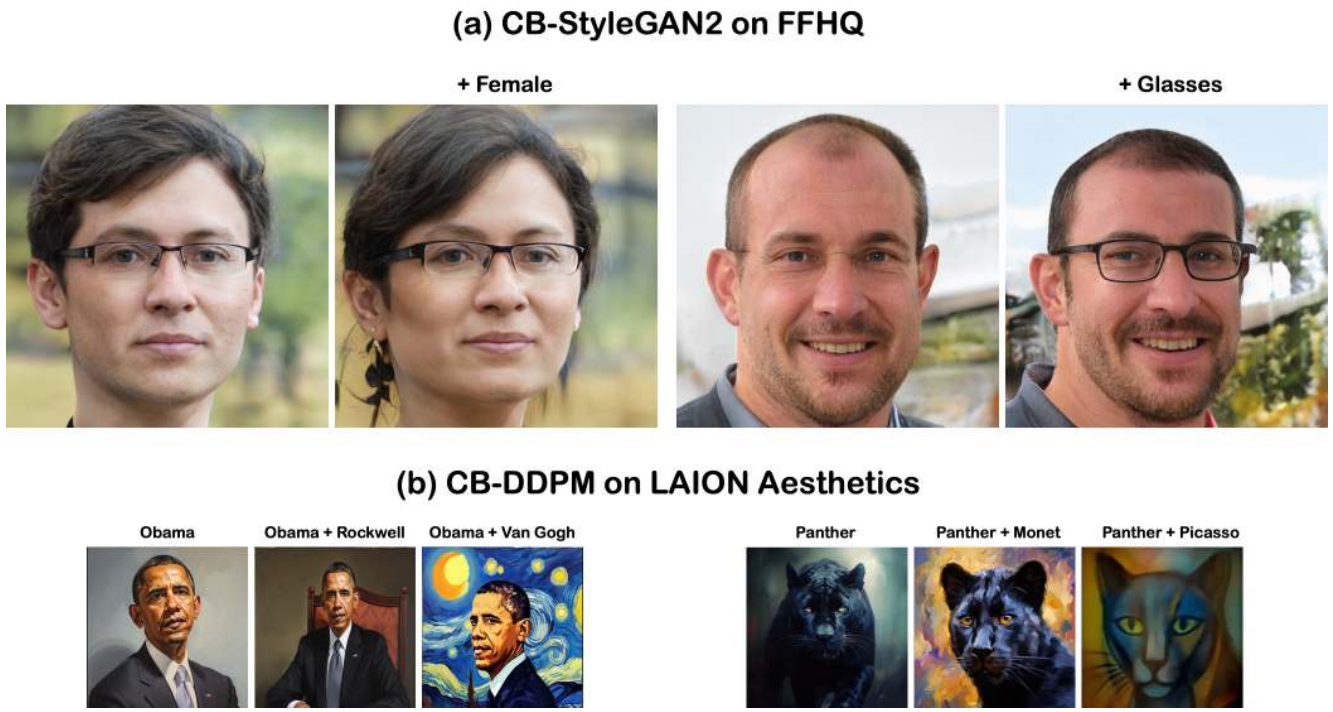
Performance

Images

Published as a conference paper at ICLR 2024

CONCEPT BOTTLENECK GENERATIVE MODELS

Aya Abdelsalam Ismail[1,2] * **Julius Adebayo**[3] * **Héctor Corrada Bravo**[1]
Stephen Ra[1,2] **Kyunghyun Cho**[1,2,4,5]



Interpretable training recipe

Can we do anything useful with this?

Interpretable training recipe

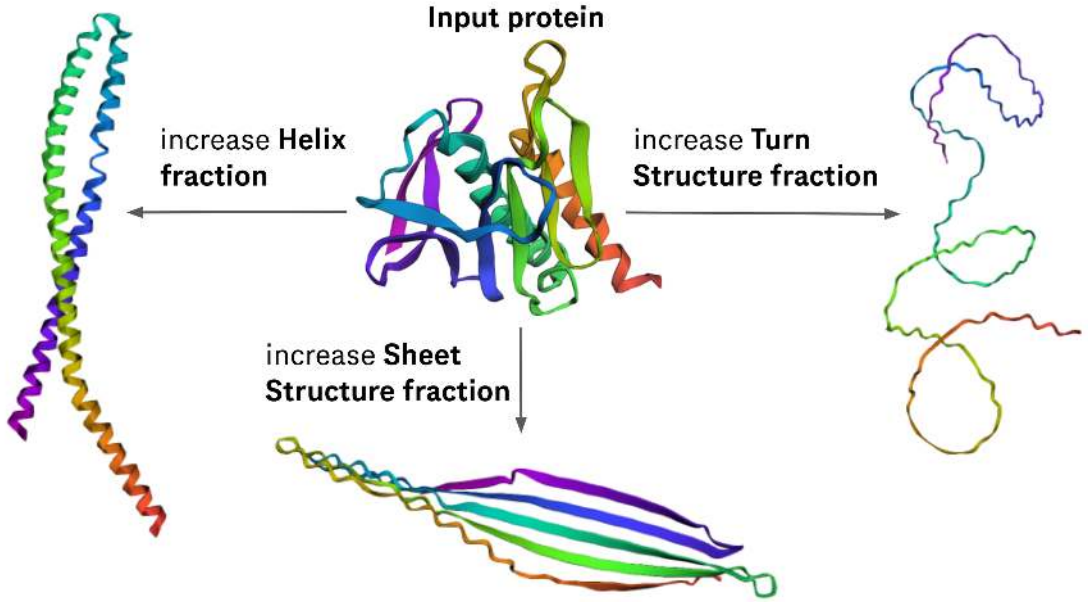
Interpret & Intervene

Protein

Published as a conference paper at ICLR 2025

CONCEPT BOTTLENECK LANGUAGE MODELS FOR PROTEIN DESIGN

Aya Abdelsalam Ismail[1,2] **Tuomas Oikarinen**[3] **Amy Wang**[1,2] **Julius Adebayo**[4]
Samuel Stanton[1,2] **Héctor Corrada Bravo**[1] **Kyunghyun Cho**[1,2,5,6] **Nathan C. Frey**[1,2]



Interpretable training recipe

Can we learn something new?

Discover

Interpretable training recipe

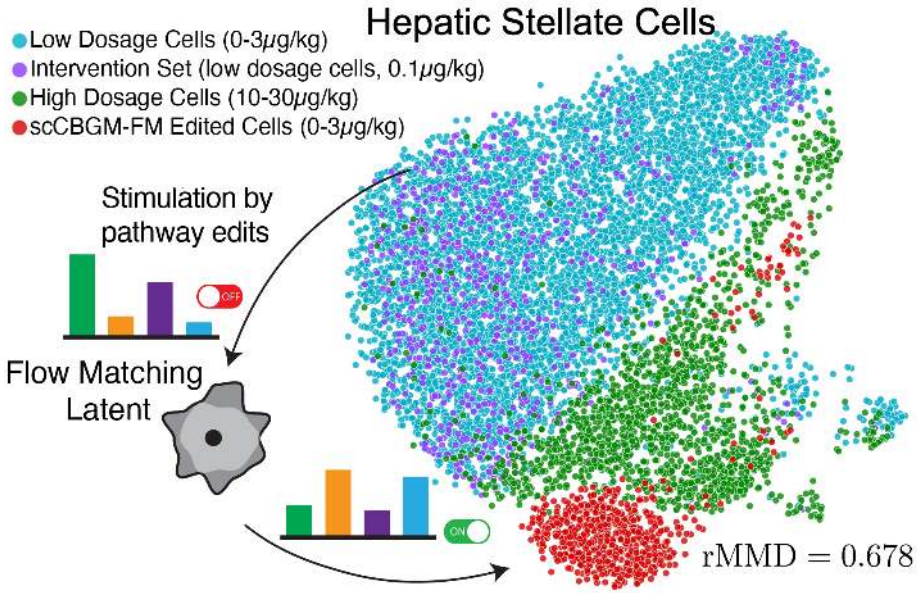
ICML 2026

Discover

Single cell

scCBGM: Single-Cell Editing via Concept Bottlenecks

Alma Andersson [*1] Aya Abdelsalam Ismail [*2] Edward De Brouwer [*1] Doron Haviv [*1] Tommaso Biancalani [1]
Kyunghyun Cho [1 3 4] Gabriele Scalia [1] Aicha BenTaieb [1] Hector Corrada Bravo [1]



Some liver cells respond to a drug, some don't. We found the pathway signature that separates them (↑TGFβ, ↑PI3K, ↑p53, ↓Trail, ↓VEGF).
Edit that signature into non-responders → they become responders the model was never trained on.

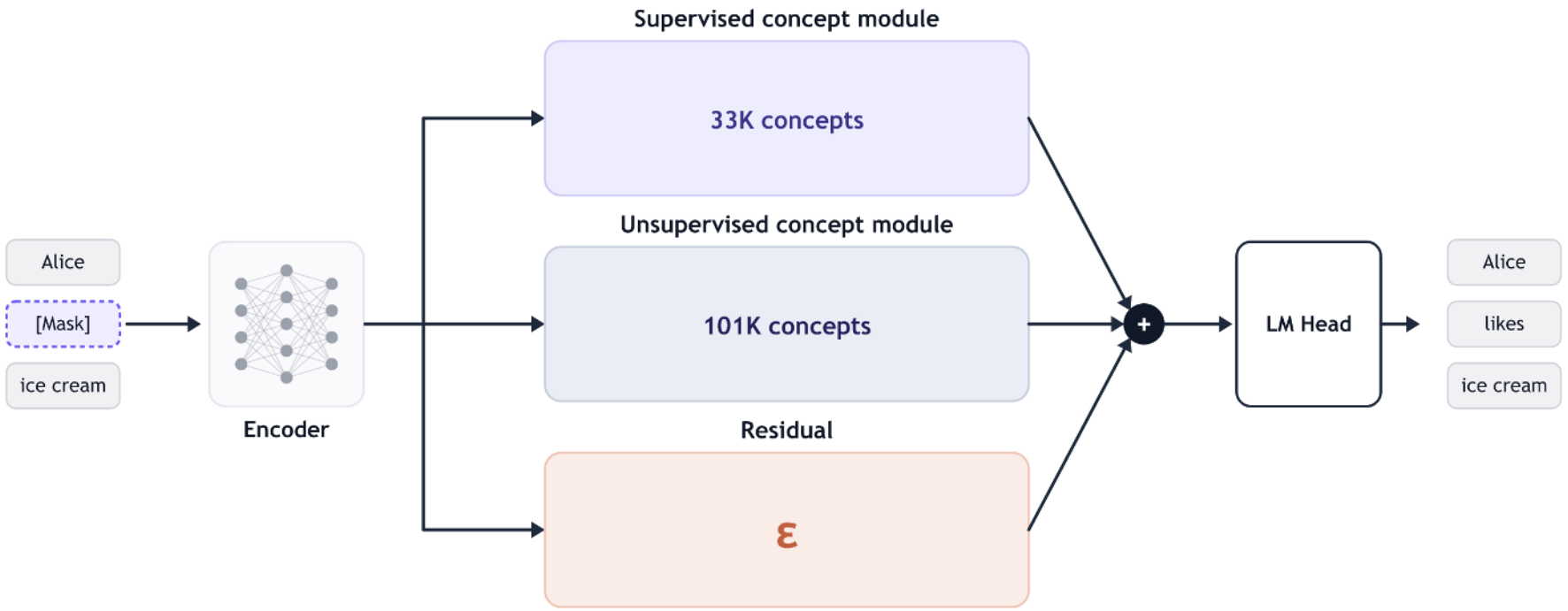
Interpretable training recipe

Does this scale?

Interpretable training recipe

Scaling

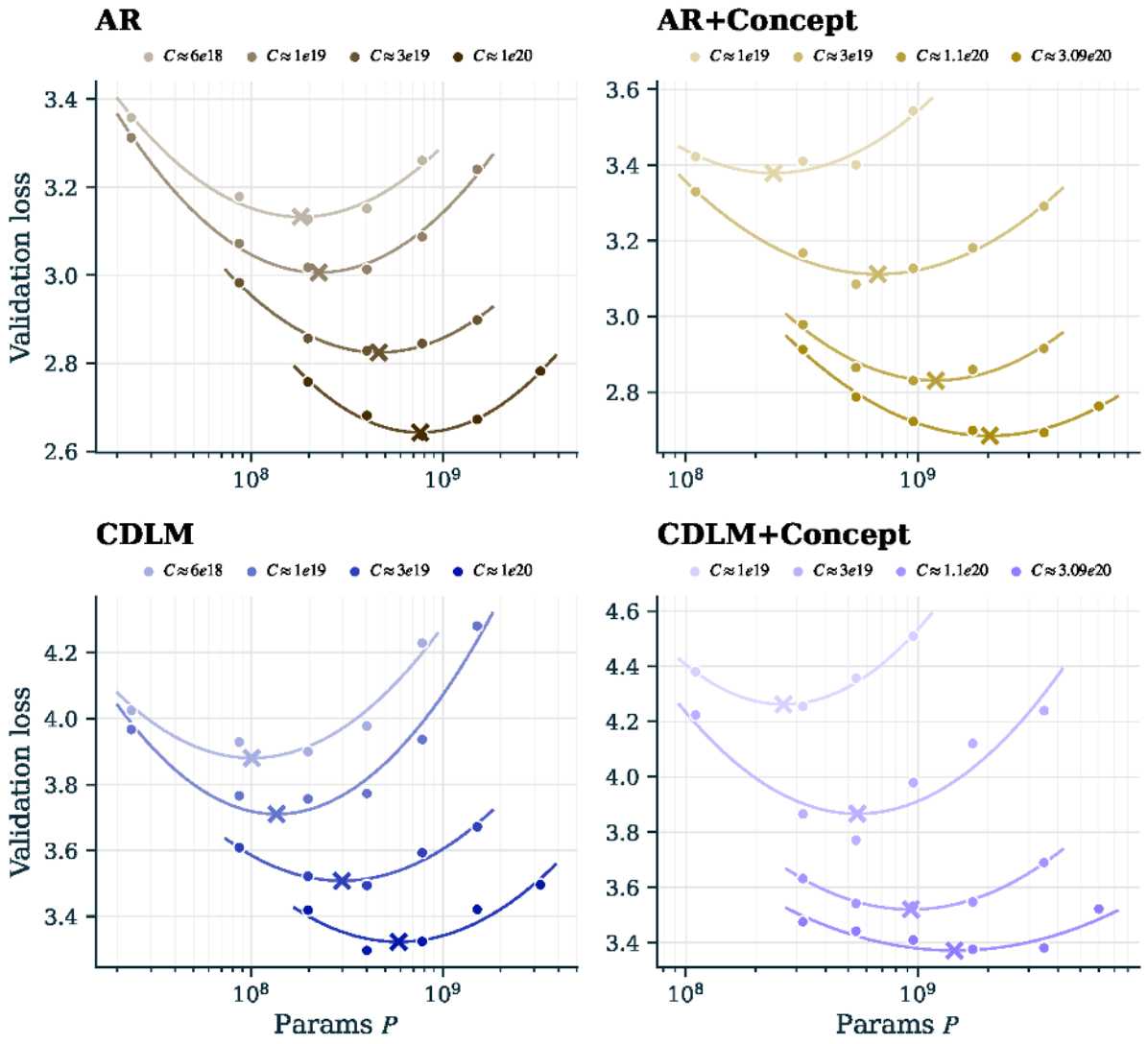
LLM



Interpretable training recipe

Scaling

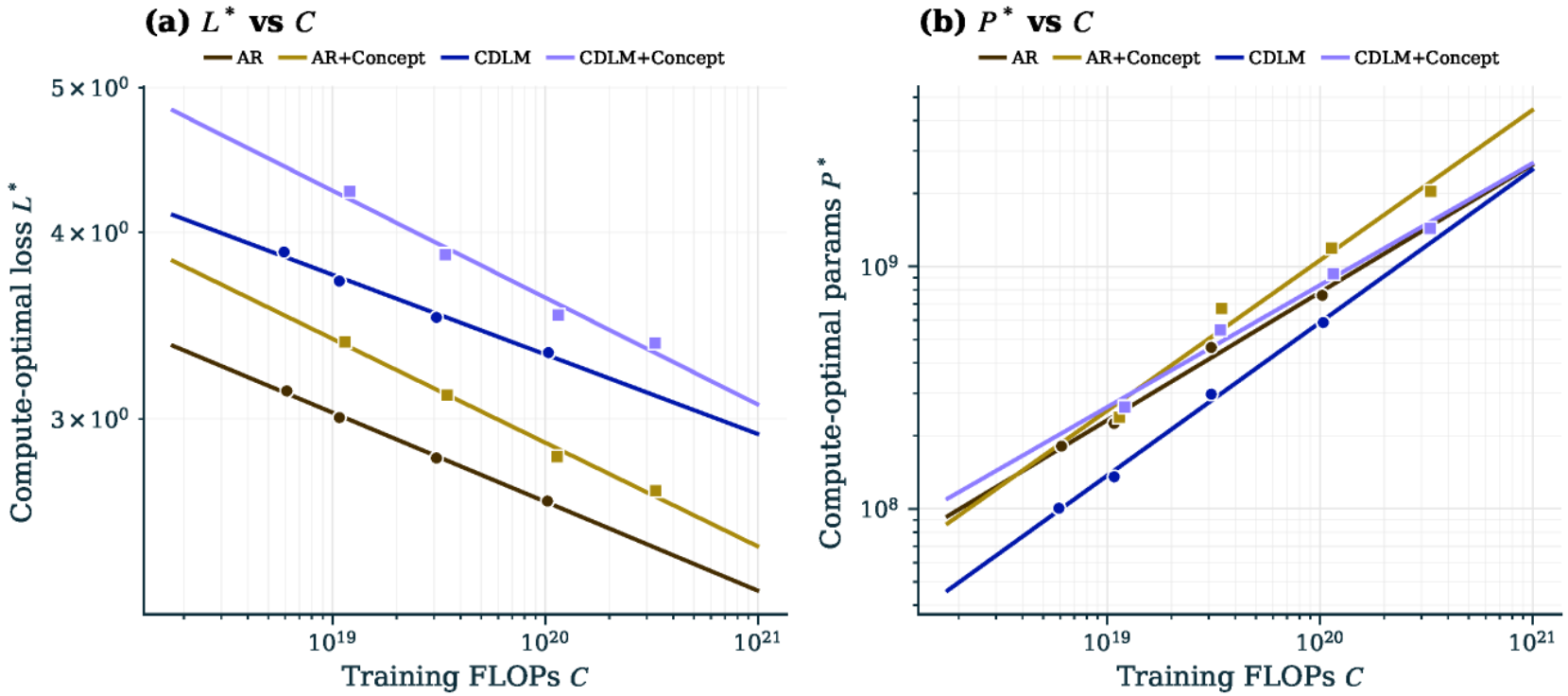
LLM



Interpretable training recipe

Scaling

LLM



Interpretability-by-design imposes a small per-backbone offset on compute-optimal scaling, not a scaling tax.

Interpretable training recipe

LLM

Scaling Inherently Interpretable Language Models

Guide Labs Team

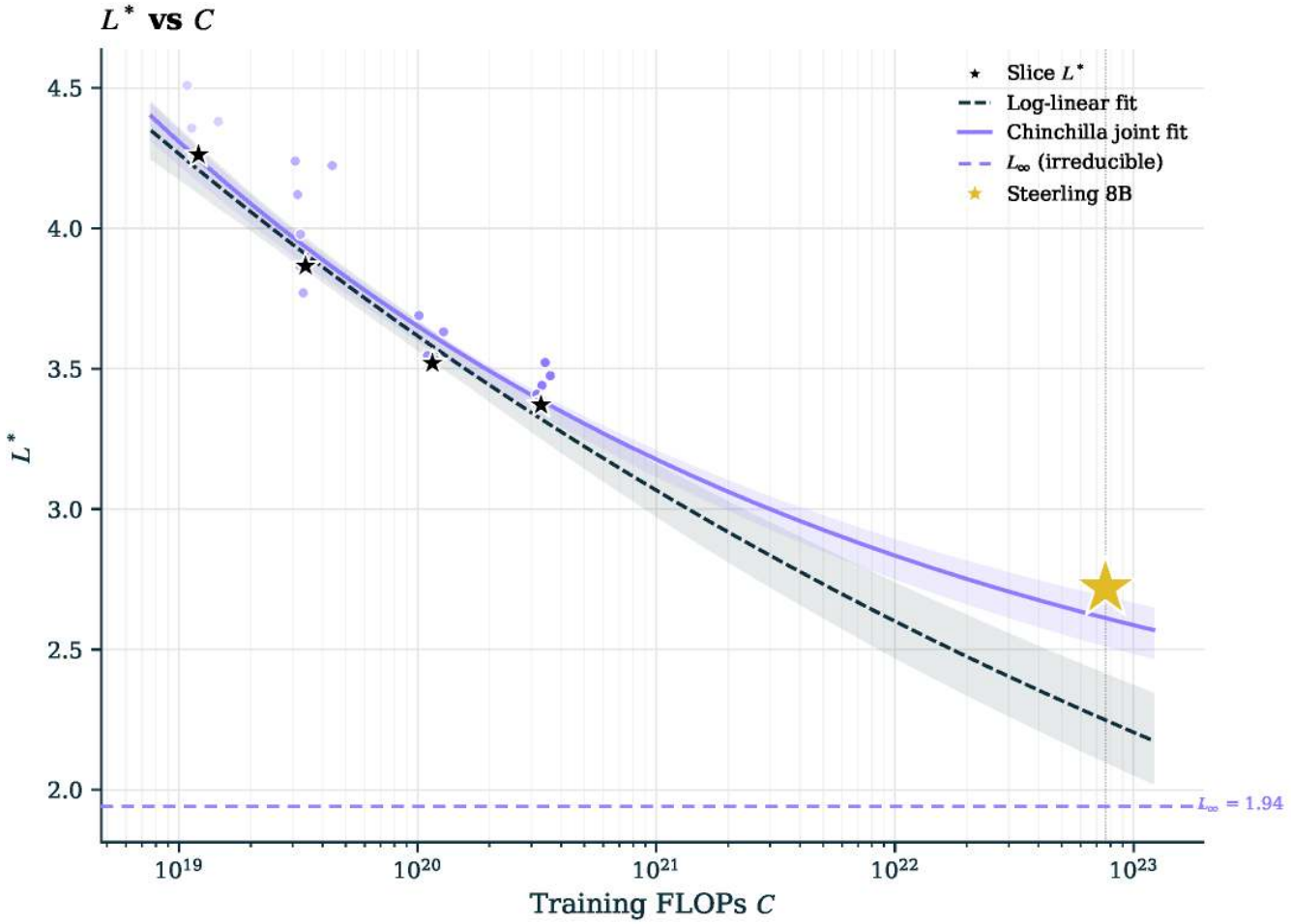
Steerling-8B

Interpretable training recipe

Steerling-8B

Scaling

LLM

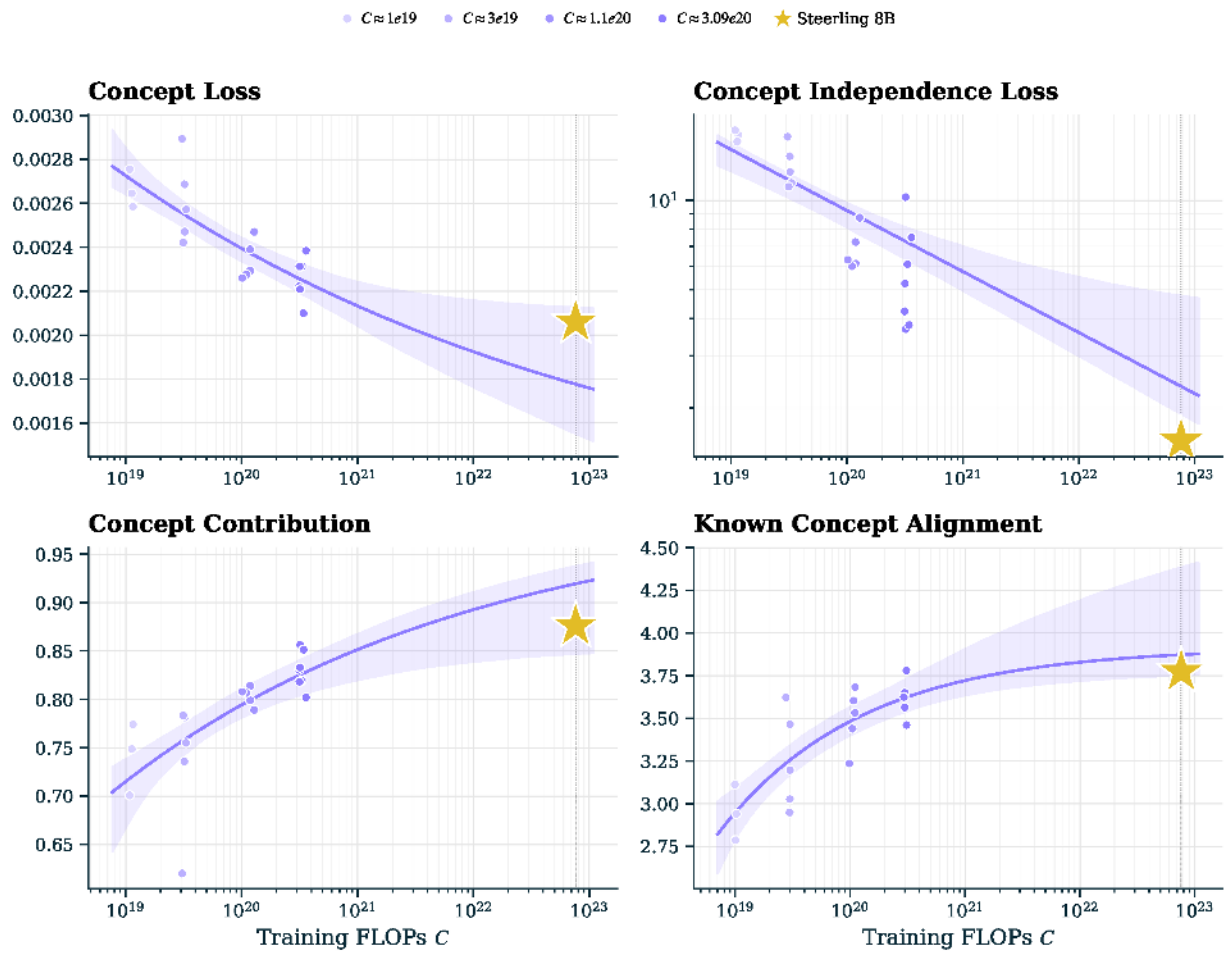


Interpretable training recipe

Scaling

LLM

Steerling-8B

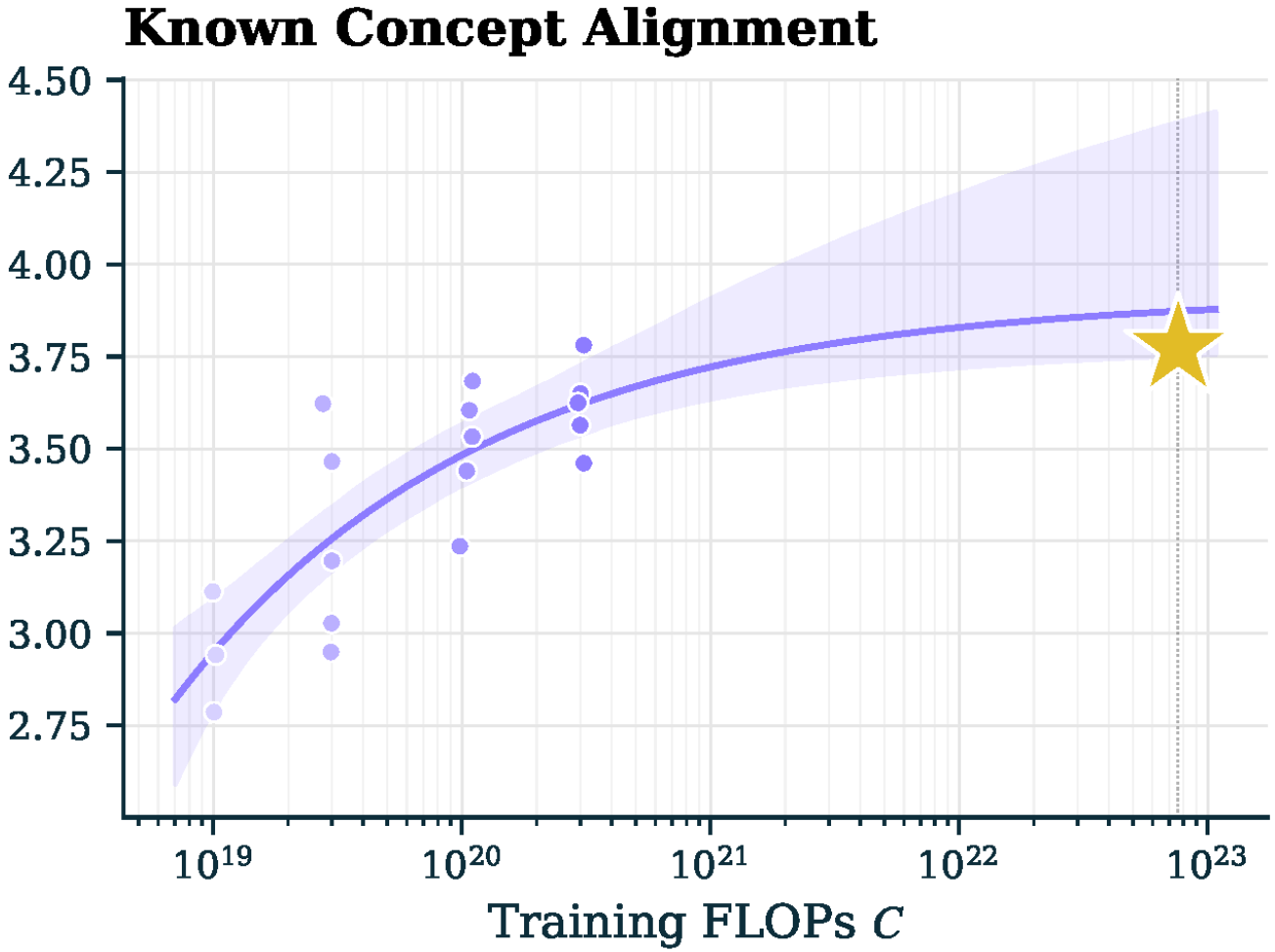


Interpretable training recipe

Steerling-8B

Scaling

LLM



Steerling-8B: First Inherently Interpretable LLM at Scale

A 58 year old male presents with sudden onset chest pain, diaphoresis, and left arm numbness. Walk through a differential diagnosis.

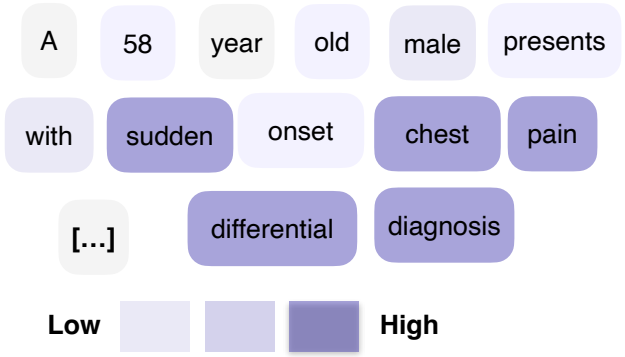
Steerling-8B Instruct

Differential Diagnosis:

1. Acute Coronary Syndrome (ACS):
   - Chest pain: Unilateral, crushing: radiating to left arm and jaw.
   - Diaphoresis with fever (>38C/100F)
   - Left-arm numbness or weakness

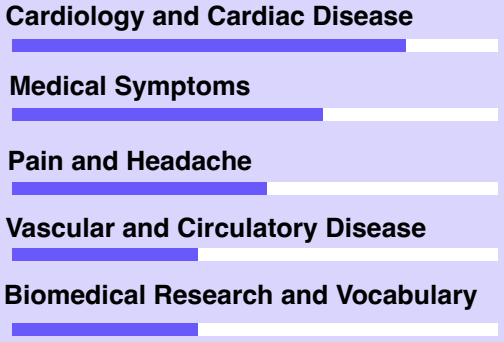
Input Attribution

Which input tokens are most influential?



Concept Attribution

What concepts is the model using?



Training Data Attribution

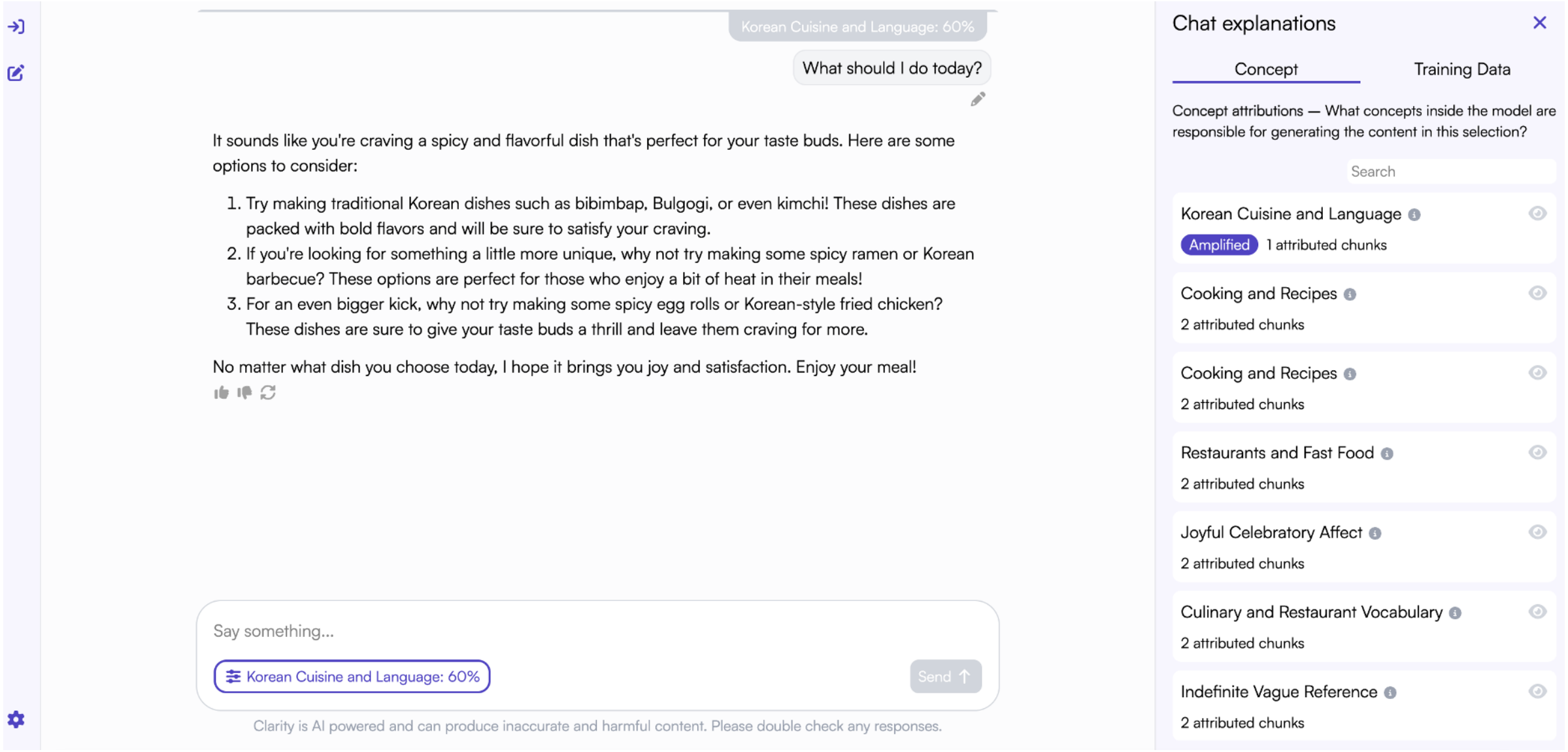
Which training data causes the output?

Symptoms Pain may be due to an isolated incident or something more severe. The pain may be isolated to an area in the right chest or maybe felt throughout. The pain also may come and go or may be persistent. […]

Source: nemotron_synth
Id: 01836-065100

0.96

Steerling-8B: First Inherently Interpretable LLM at Scale



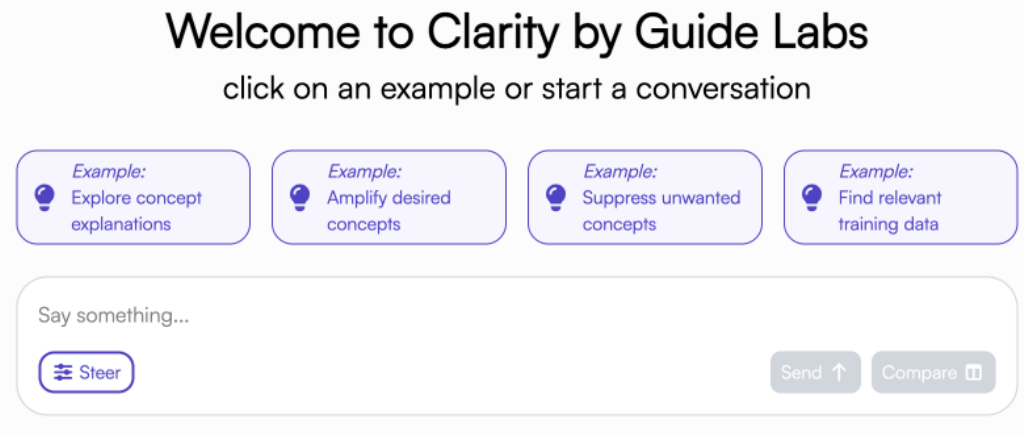
Steerling-8B: First Inherently Interpretable LLM at Scale

Base model and SFT models 🤗: https://huggingface.co/guidelabs/steerling-8b

Code on GitHub: https://github.com/guidelabs/steerling

Technical report: https://www.guidelabs.ai/papers/scaling-inherently-interpretable-language-models/

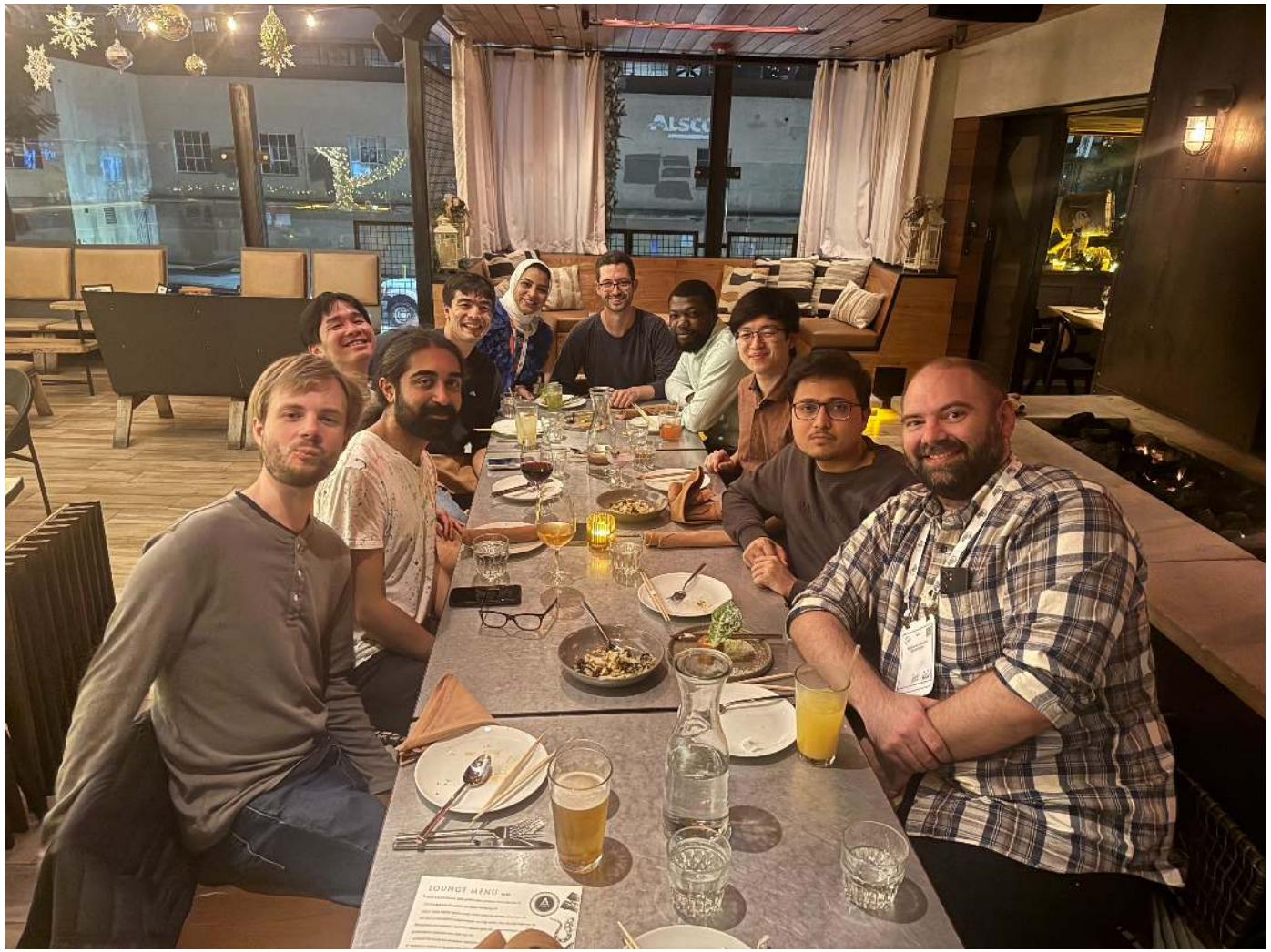
Clarity platform: https://platform.guidelabs.ai/



TL;DR: If you want interpretability, train interpretable models.

This is one instantiation. There are others, and probably better ones.

Finding them is why we're all here today.



Questions?