CS11-711 Advanced NLP

Model Interpretability 🔎

Nishant Subramani
PhD Student - CMU LTI

Site
https://phontron.com/class/anlp2024

What I want you to take away

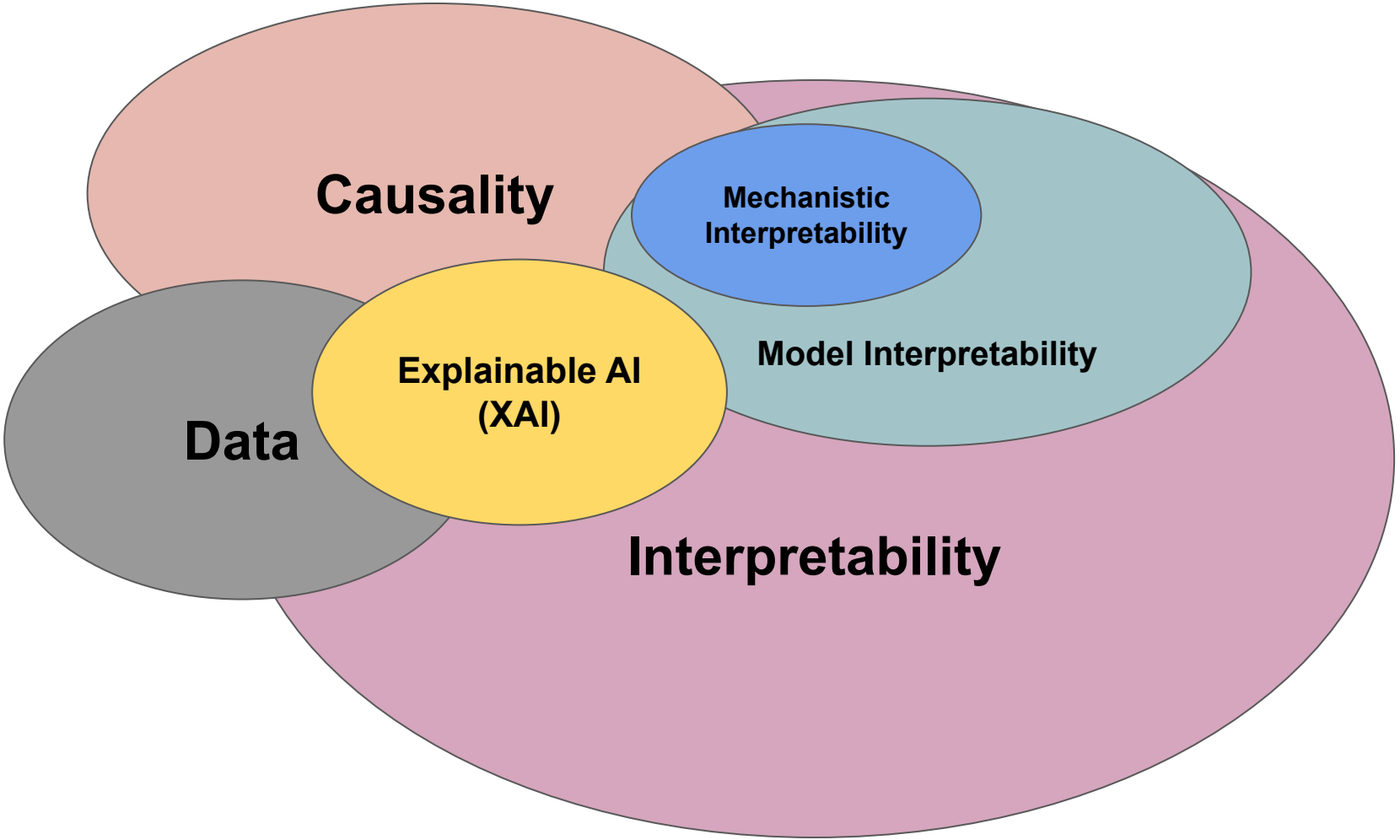
Takeaway 1: *Model Interpretability is important to study!*

Takeaway 2: *Model Interpretability is interesting and something you want to explore more!*

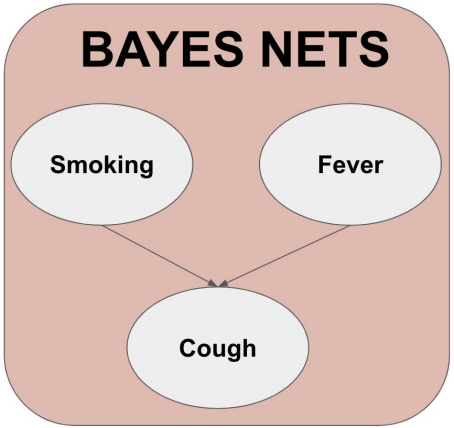
Interpretability (in AI)

Definition: *The study of understanding the decisions that AI systems make and putting them into easily human-understandable terms.*

Why: to use that understanding to iteratively better design systems that are more ***performant*** and ***human-understandable***.


Causality
Mechanistic Interpretability
Model Interpretability
Explainable AI (XAI)
Data
Interpretability

Historically models are small



LINEAR REGRESSION

$$y = mx + b$$

LOGISTIC REGRESSION

$$sigmoid(Wx + b)$$

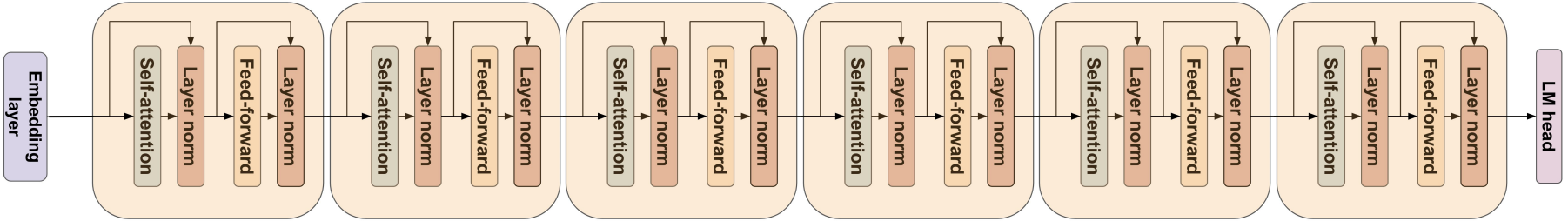
MULTIVARIATE LINEAR REGRESSION

$$y = Wx + b$$

2 LAYER MLP w/ ReLU

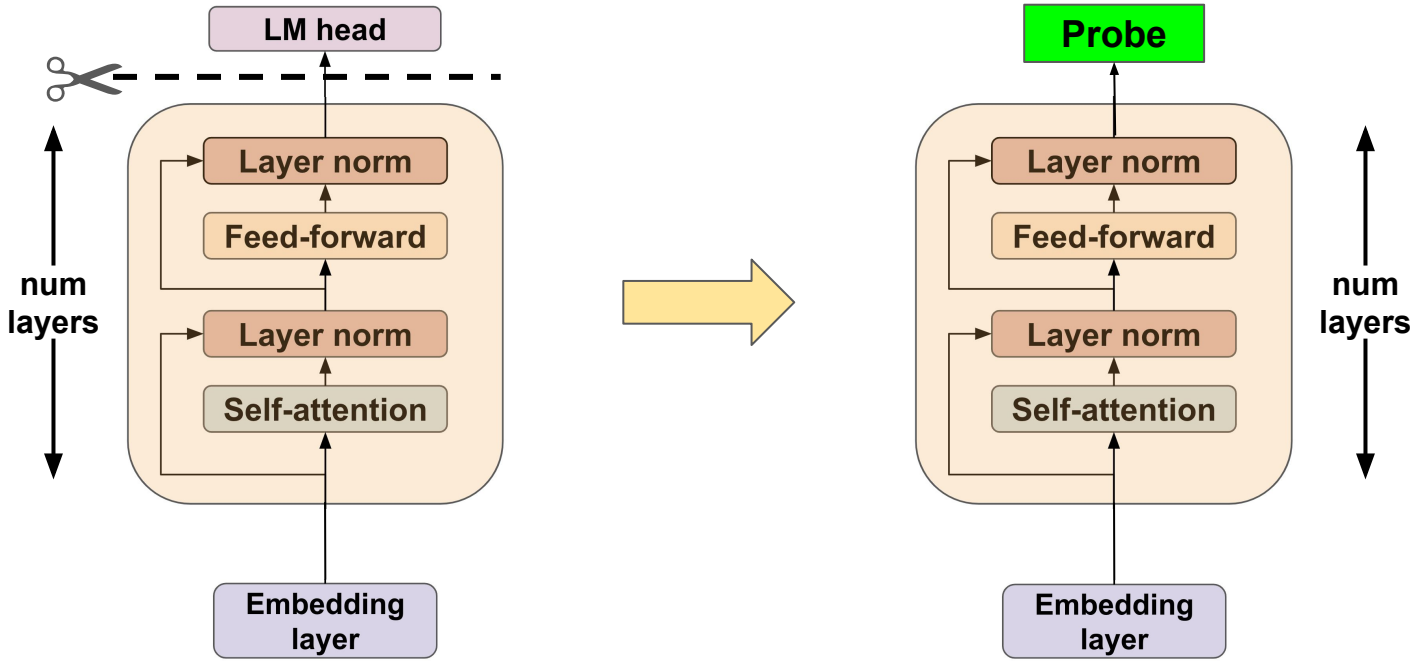
$$W(ReLU(Wx + b)) + c$$

Now they are not



PROBING

How do we make sense of this huge model?

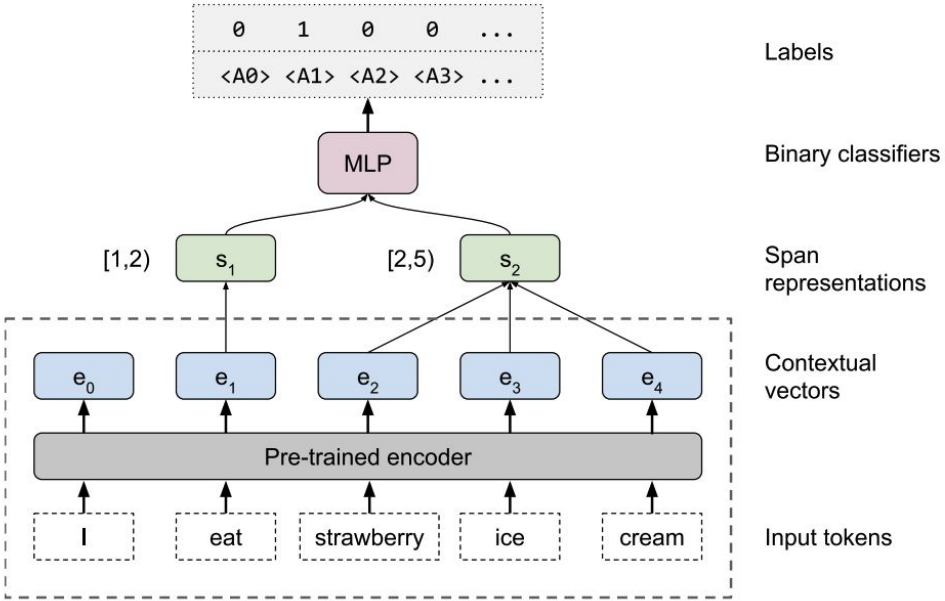


What is a Probe?

Definition: *A classifier that is specifically trained to predict some property from a pretrained model's representations.*

Edge Probing (Tenney et al. 2019)

- General method that works to probe different types of information



BERT rediscovers the NLP pipeline (Tenney et al. 2019)

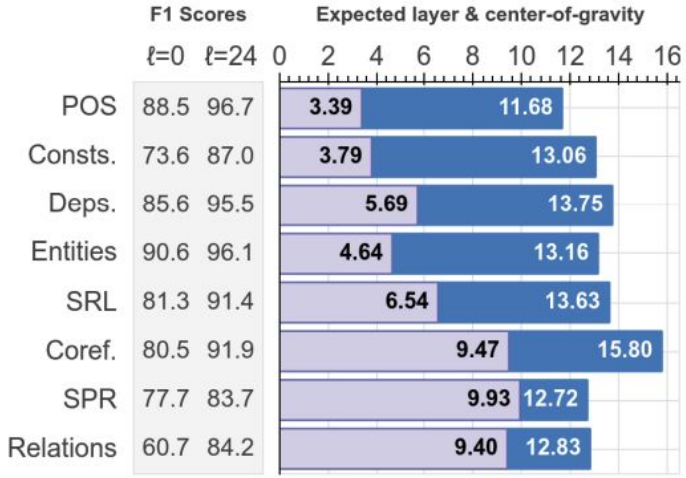


Figure 1: Summary statistics on BERT-large. Columns on left show F1 dev-set scores for the baseline ($P_\tau^{(0)}$) and full-model ($P_\tau^{(L)}$) probes. Dark (blue) are the mixing weight center of gravity (Eq. 2); light (purple) are the expected layer from the cumulative scores (Eq. 4).

POS - part of speech tagging (e.g. this word is a noun)

consts - constituent labeling (e.g. is this span a noun phrase)

deps - dependency labeling (e.g. is span_one the subject and span_two the object)

entities - named entity labeling (e.g. this word is a person)

SRL - semantic role labeling (what roles are the spans playing with each other: “Mary (pusher) pushed John (pushee)”

coref - coreference (do span_one and span_two refer to the same entity or event)

SPR - semantic proto-role (identifying attributes like awareness so is Mary aware that they are doing the pushing)

relations - relation classification (predicting the real-world relation between two spans given a set of these)

Issues with Probing (Belinkov et al. 2021)

- Probe ✅
  - Representation encodes information 👍🏽
  - Probe solved task by itself 😳
- Probe ❌
  - Representation lacks the information 👎🏽
  - Representation encodes information, but probe is not the right function class 🧐
- We want to probe *tasks*, but require supervised data, so instead we probe *datasets*
- Probes designed this way are *correlative* not *causative*

Other Probing Works

Information-Theoretic Probing with Minimum Description Length

Elena Voita[1,2] Ivan Titov[1,2]

Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals

Yanai Elazar[1,2] Shauli Ravfogel[1,2] Alon Jacovi[1] Yoav Goldberg[1,2]

Low-Complexity Probing via Finding Subnetworks

Steven Cao[1,2] Victor Sanh[2] Alexander M. Rush[2]

Pareto Probing: Trading Off Accuracy for Complexity

Tiago Pimentel[*,] Naomi Saphra[*,] Adina Williams Ryan Cotterell

What is Model Interpretability?

Definition: *The study of understanding the internals of models (e.g. their **weights** and **activations**), putting those insights in human-intelligible terms, using that insight to both patch current models and develop better ones.*

What is Mechanistic Interpretability?

Definition: *The study of reverse engineering parametric models (often neural networks) from their learned weights into more human-interpretable algorithmic units.*

Notable Work

- Analysis of 1 and 2-layer MLPs and Transformers to find circuits (Olah et al. 2021)
- Induction Heads (Olsson et al. 2022)
- Neuron Polysemanticity (Elhage et al. 2021; 2022)

MODEL INTERPRETABILITY

Weights 🏋️ and Activations 💪

- Weights
  - You can edit them and see what happens
- Activations
  - Look at activations for different inputs
  - Poke them with a stick and see what happens
  - **The technical term:** Intervene on them by adding a vector or some other manipulation

LOOKING AT WEIGHTS

Model Editing

Target: *A concept or specific fact needs to be changed in the model*

Approach: *Changing the weights of the model to edit the model's belief of that fact/concept?*

ROME (Meng et al. 2022)

- Use causal tracing to isolate the causal effect of individual hidden states when processing a fact
- Introduce rank-one model editing (ROME) to edit the model

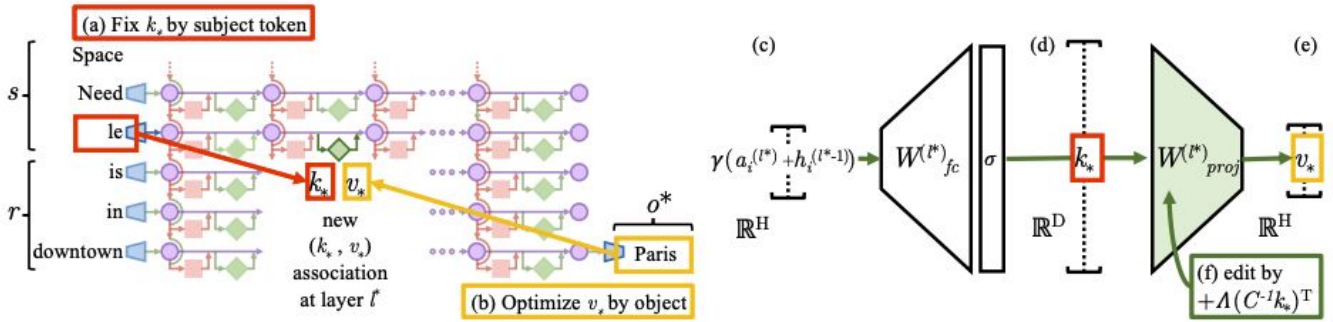


Figure 4: **Editing one MLP layer with ROME**. To associate *Space Needle* with *Paris*, the ROME method inserts a new (k_*, v_*) association into layer l^*, where (a) key k_* is determined by the subject and (b) value v_* is optimized to select the object. (c) Hidden state at layer l^* and token i is expanded to produce (d) the key vector k_* for the subject. (e) To write new value vector v_* into the layer, (f) we calculate a rank-one update $\Lambda(C^{-1}k_*)^T$ to cause $\hat{W}^{(l)}_{proj}k_* = v_*$ while minimizing interference with other memories stored in the layer.

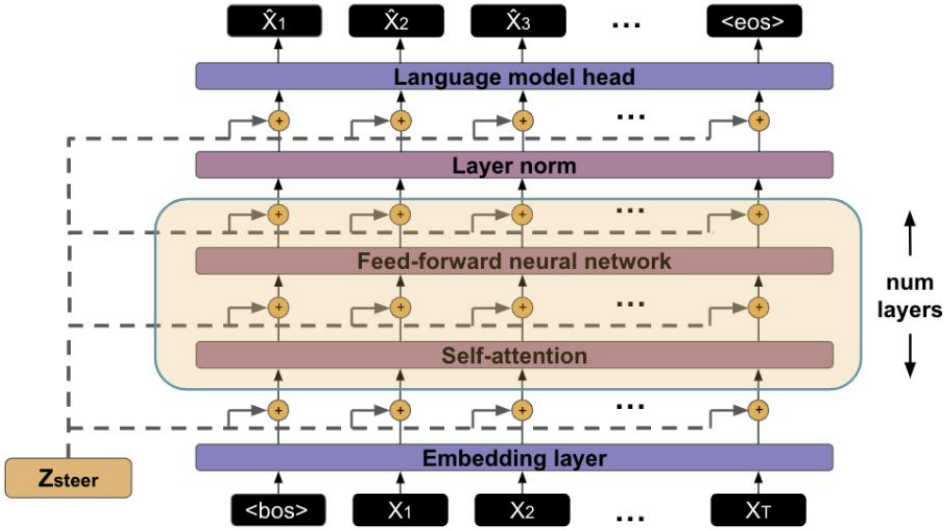
LOOKING AT ACTIVATIONS

Steering Vectors (Subramani et al. 2019; 2020; 2022)

Steering Vectors: a fixed-length vector that steers a language model to generate a specific sequence exactly when added to the hidden states of a model at a specific location.

This is our stick that we're poking a language model with.

Extracting steering vectors



ALGORITHM 1: Extracting z_{steer} for a sentence

Input : x – target sentence
M – pretrained language model
θ – pretrained language model weights
I_L – injection location
I_T – injection timestep
d – dimension of z_{steer}

Output : z_{steer} – extracted candidate steering vector

$z_{steer} \sim$ xavier_normal(d)
for *i* $\leftarrow$ *[1, 2, ..., N]* **do**
  logits $= \mathsf{M}_\theta.forward(\mathsf{x}, z_{steer}, I_L, I_T)$
  $\mathcal{L} = XENT(logits, x)$
  $\mathcal{L}.backward()$
  $z_{steer} = z_{steer} + lr * \frac{\partial \mathcal{L}}{\partial z_{steer}}$
end
return z_{steer}

Steering vector results

- Steering vectors exist and we can find them easily for most sequences
- They have interpretable properties
  - Distances in steering vector space reflect semantic similarity
  - Style transfer is possible with simple vector arithmetic
  - Decoding from interpolations in the latent space produces meaningful output

Steering vectors	
Positive Input	the taste is excellent!
$+1.0 * z_{tonegative}$	the taste is excellent!
$+2.0 * z_{tonegative}$	the taste is unpleasant.
Negative Input	the desserts were very bland.
$+1.0 * z_{topositive}$	the desserts were very bland .
$+2.0 * z_{topositive}$	the desserts were very tasty.

Inference-time Interventions (Li et al. 2023)

- Use linear probes to find attention heads that correspond to the desired attribute
- Shift attention head activations during inference along directions determined by these probes

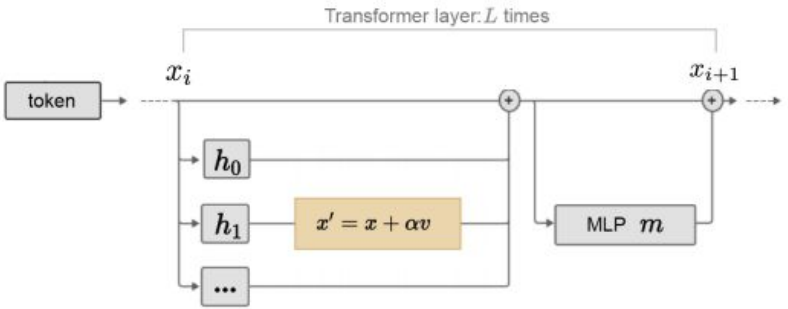
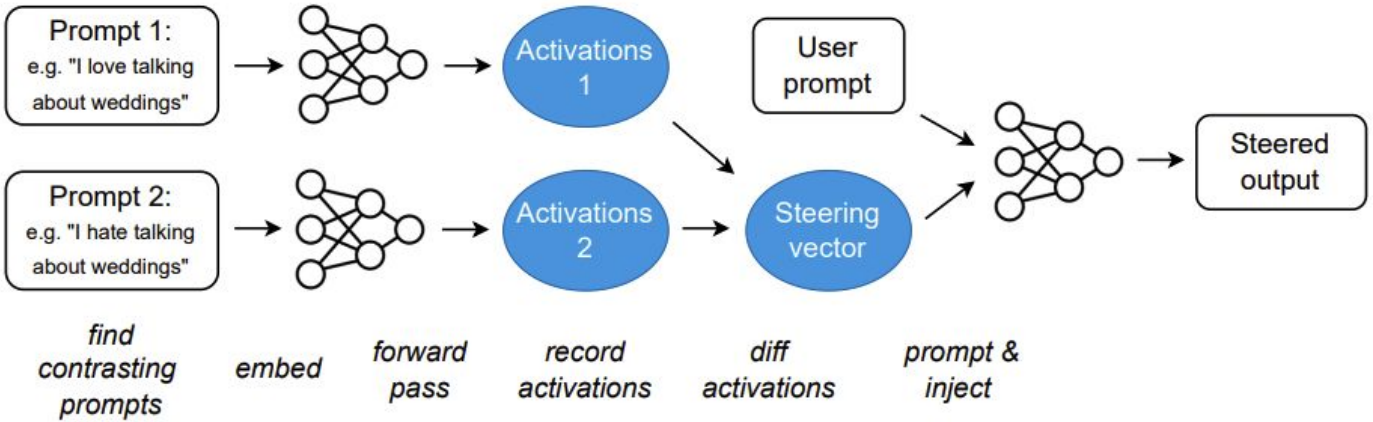


Figure 3: A sketch of the computation on the last token of a transformer with inference-time intervention (ITI) highlighted.

More activation manipulation

- Contrastive steering vectors (Turner et al. 2023; Rimsky et al. 2023)

Figure 1: Schematic of the Activation Addition (**ActAdd**) method. ⬭ = natural language text; ● = vectors of activations just before a specified layer. In this example, the output is heavily biased towards discussing weddings, regardless of the topic of the user prompt. (See Algorithm 1 for omitted parameters over intervention strength and location.)



What can model interpretability give us?

Outcome 1: Better understanding of *how* language models work.

Outcome 2: Light-weight methods to *control* and *steer* models.

Outcome 3: Potential alternatives or complementary methods to further align models to human preferences.

Resources: some NLP model interp groups

- Ellie Pavlick's group at Brown
- David Bau's group at Northeastern
- Hassan Sajjad's group at Dalhousie
- Martin Wattenberg's group at Harvard
- Jacob Andreas's group at MIT
- Yonatan Belinkov's group at Technion
- Mor Geva's group at Tel Aviv University
- Anthropic's Mech Interp team
- Google's PAIR, NLP, and MechInterp teams
- EleutherAI's Interp team

Questions?