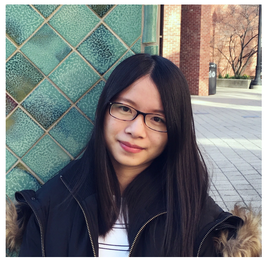
Tutorial on Mechanistic Interpretability of Language Models



Ziyu Yao
Asst. Prof.

Daking Rai
PhD Student

Department of Computer Science
George Mason University, USA

https://ziyu-yao-nlp-lab.github.io/ICML25-MI-Tutorial.github.io/

Tutorial Website with All the Materials and Recordings

Post-event Q&A to:
{ziyuyao, drai2}@gmu.edu

What comes to your mind when you think about language models?

These?

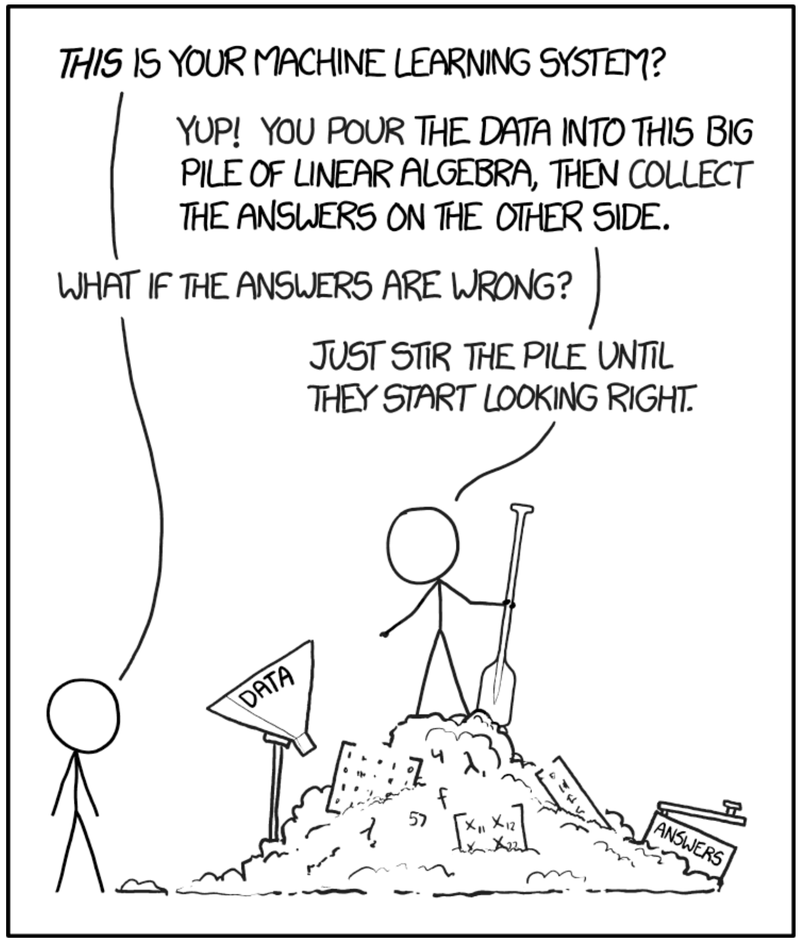
- A **language model (LM)** is a probabilistic model that learns the likelihood of next-token prediction in a language sequence.

$$P(x_n \mid x_1, x_2, \ldots, x_{n-1}), x_n \in \mathscr{V} \qquad (\mathscr{V}\text{: the vocabulary})$$

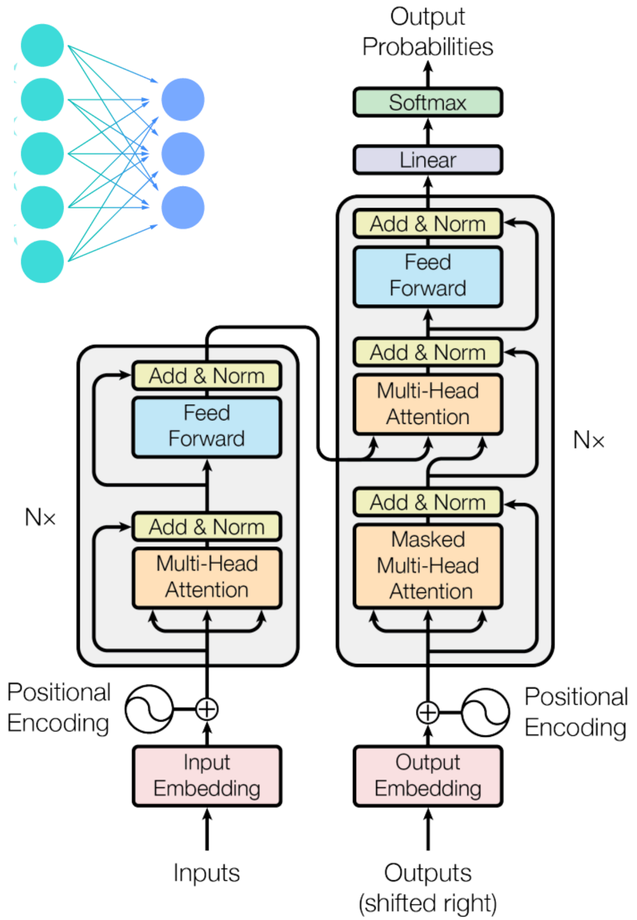
- This tutorial: LMs implemented by **the transformer architecture** (Vaswani+17)

[1] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

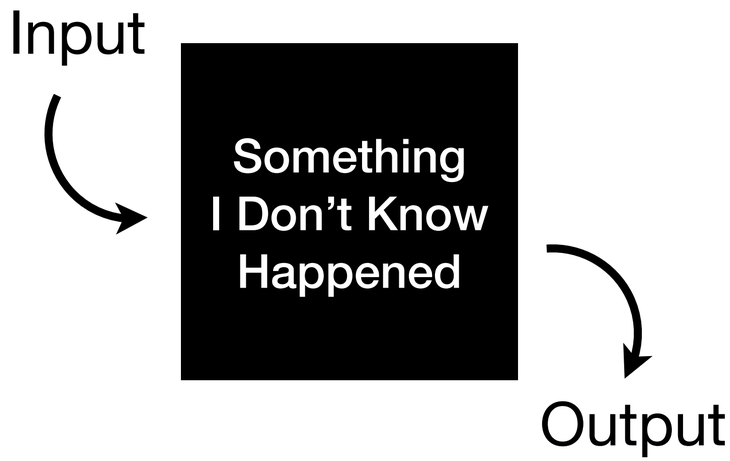
What comes to your mind when you think about how LM works?



As a Machine Learning model…



A neural model with Transformer architecture



This is the central question of interpretability of LMs!

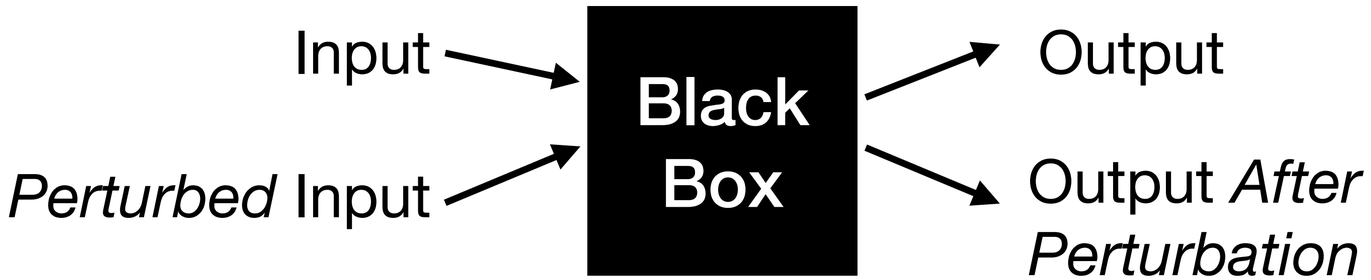
Why do we want to understand how LM works?

- Scientific understanding
- Practical reasons
  - AI safety and alignment
  - Human trust in AI
  - Enhancing model capability
  - Guiding model design, training, inference

I’m just curious!

This tutorial: Mechanistic Interpretability (MI) of LMs

- **Behavioral Interpretability** (Ribeiro+16; Lundberg&Lee+17)



- **Mechanistic Interpretability** (Olah+20; Elhage+21)
  - Understand the internals of LMs, how they function individually, and how they are connected to enable LM behaviors and capabilities
  - Intuitively more insights than looking at only behaviors!

Let's see how the model reacts to changes in the input…

Let's open the black box and look inside!

[1] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." SIGKDD 2016.
[2] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." NeurIPS 2017.
[3] Olah, Chris, et al. "Zoom in: An introduction to circuits." Distill 5.3 (2020)
[4] Elhage, Nelson, et al. "A mathematical framework for transformer circuits." Transformer Circuits Thread 1.1 (2021): 12.

The thriving community, with doubts

Stay Tuned!
https://xllm-reasoning-planning-workshop.github.io/

Mechanistic?

Naomi Saphra, Sarah Wiegreffe

Meanwhile, argument about its scope, definitions, etc.

Scope of this tutorial

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. "A practical review of mechanistic interpretability for transformer-based language models." Preprint 2025 (Version 2).

- We are not here to tell you if MI is worth pursuing or not, but we equip you with the necessary knowledge background.
  - **Part 1: Fundamental Objects of Study (7 min)**
    - What are the fundamental objects of studies of MI? What are the scopes?
  - **Part 2: MI Practices and Techniques (75 min)** ***10min Q&A/Break***
    - What are the common practices of applying MI to LMs? What techniques should you grasp?
  - **Part 3: Findings and Applications (25 min)**
    - What people have found out by applying MI to LMs? How has MI been helpful in applications?
  - **Part 4: Challenges and Future Work (10 min)**

(Disclaimer: Neither this 2.5hr tutorial nor our survey paper include all existing work)

Refresher: Transformer Architecture

- An input token x_i is processed layer by layer (assuming L layers)
- $h_i^0 \in \mathbb{R}^d$: the embedding vector looked up from $W_E \in \mathbb{R}^{|\mathscr{V}| \times d}$
- $h_i^l \in \mathbb{R}^d$: representation (or "activation") at layer l

$$h_i^l = h_i^{l-1} + a_i^l + f_i^l$$

Multi-head Attention (MHA)

$$a_i^l = Concat(a_{i,1}^l, .., a_{i,H}^l) W_O^l$$

(H: number of heads)

Feed-Forward (FF)

$$f_i^l = f\left((h_i^{l-1} + a_l^i) \cdot W_k^l\right) \cdot W_v^l$$

- Finally, calculate the logit distribution for the next token with the unembedding matrix $W_U \in \mathbb{R}^{|\mathscr{V}| \times d}$: $Logit = h_i^L \cdot W_U^\top$