Intro to LLMs

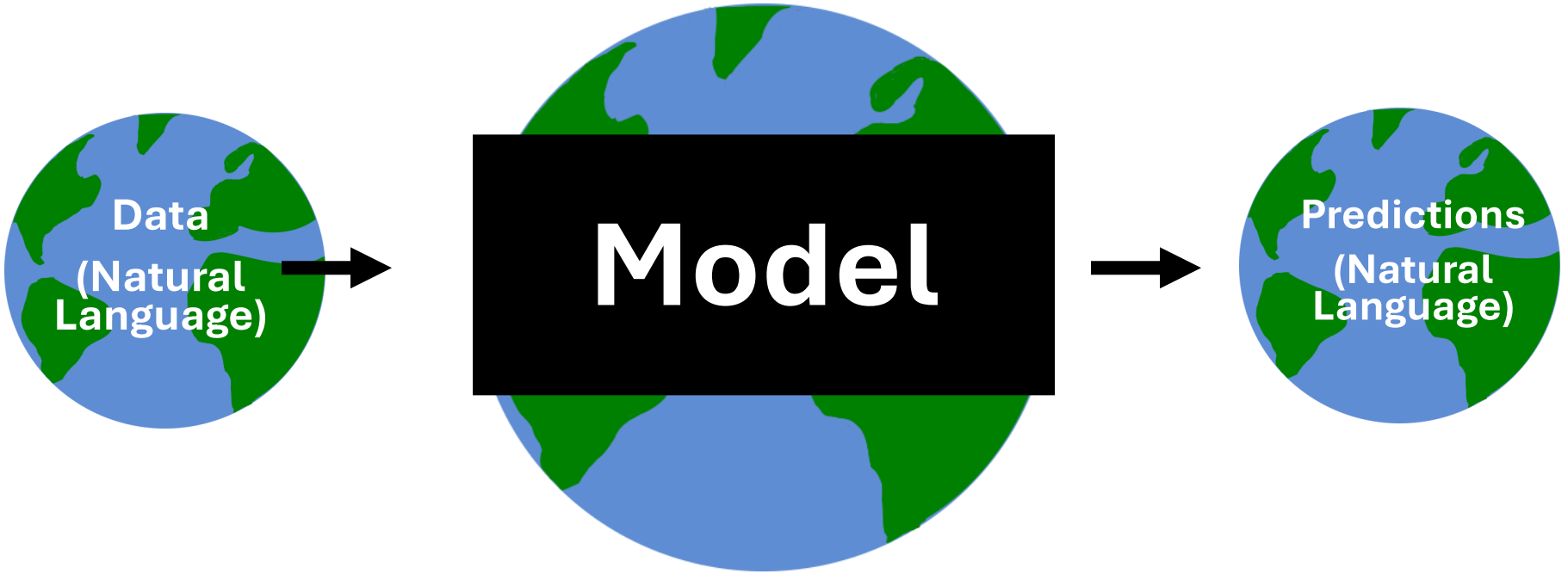
5 February 2026

Alex Lyman

Language Models

- Distributional Hypothesis (Firth 1954, Harris 1957)
- If language follows a probability distribution, we could probably learn to *approximate* it, right?
- How could we do that?

Language Models

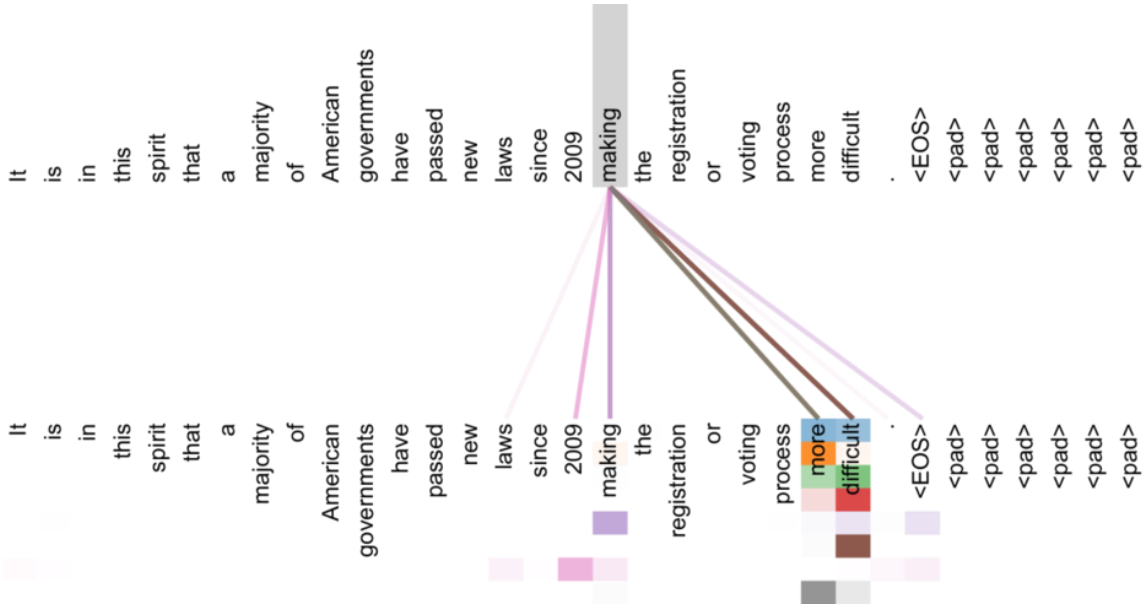


What is a Language Model?

- Language Models are machine learning models just like the models we've talked about all semester.
- Language Models have parameters (weights).
- Language Models learn from data.
- They do this by minimizing loss.
- Language Models make predictions.

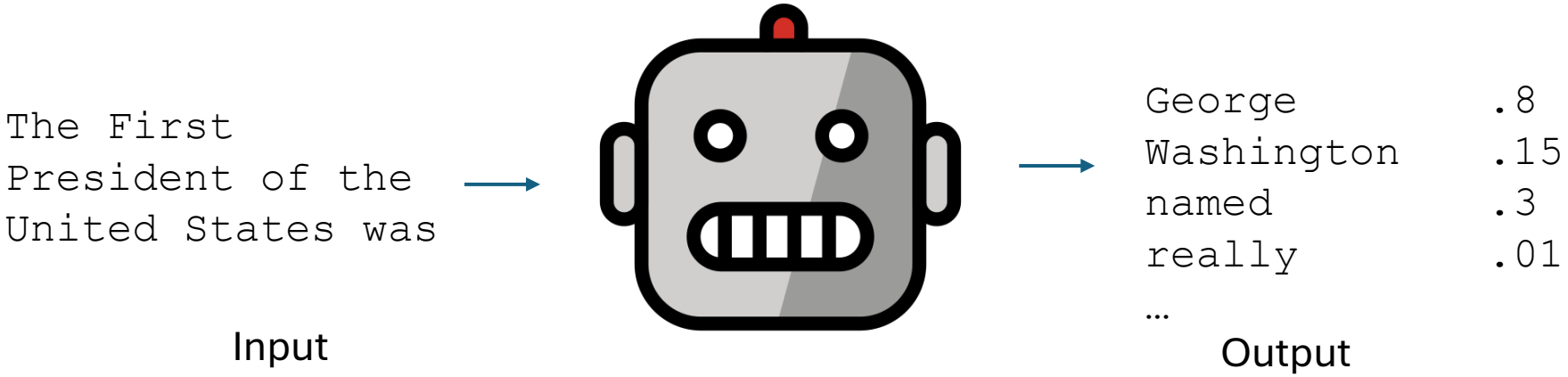
What is a Language Model?

- Language Models are based on a neural network architecture called the Transformer (2017).
- Big innovation: Self-Attention.
- While the model is training, it is able to learn dependencies between words.



Language Model Training Objective

- We said our models need to learn by minimizing loss.
- You can't minimize loss without being able to calculate error.
- Next-token prediction training objective:
  - Given a sequence of tokens (words), predict the next token in the sequence.
- Inputs: A sequence of tokens (words)
- Output: The next token in the sequence
- Actually, a probability distribution over all possible next tokens.



Language Models - Size

- How many trainable parameters (weights) do the models have that we've talked about in class so far?
  - Simple linear regression: *2* $y = w_1 x + w_0$
  - KNN: *0* (it's a lazy learner)
  - Gaussian Naïve Bayes: *2(#classes)(# features)* + *(#classes-1)*
  - Logistic Regression (binary classification): *(# features)* + *1*
  - Logistic Regression (multinomial classification): *(# features +1)(#classes)*

- What about Language Models?
  - GPT-2: 1,500,000,000
  - GPT-5 (estimated) 2,000,000,000,000

Language Models - Data

- How many data points have we trained our models on in this class so far?
  - Glass dataset (KNN lab): 214 examples, 9 features – 1,926 data points.
- What about Language Models?
  - GPT-2: 10,000,000,000 data points (tokens)
  - Modern Language Models (Llama 3) 15,000,000,000,000 data points

Language Model Scaling

- What is even the point of all those parameters?
- Turns out, adding more parameters (and data) makes LLMs smarter.
- **Emergent Abilities**: Abilities a language model gains that weren't explicitly part of its training objective.
- Examples:
  - Read millions of examples of multiplication, learn math.
  - Read millions of sequences in both Spanish and English, learn to translate.
  - Read billions of stories about real-world items, learn relative sizes, shapes, and properties of things.
- Training on these trillions of tokens is called the pre-training phase.

Language Model Progress Over Time

- What is even the point of all those parameters?

Prompt: “Sean looked behind the curtain to find”

- GPT-2 (2019)

```
Sean looked behind the curtain to find another member of the audience in the seats and told her to leave. The person continued
arguing with the woman sitting next to you and at that moment my mind was on the woman next to me who said her boyfriend was ge
tting violent in front of her child.

All I knew is that what I was told was a lie because I knew when I heard the voice the person was not actually there.
```

- Scaling, GPT-3 (2021)

Sean looked behind the curtain to find the source of the crying and found a little girl, no older than four, curled up in a ball on the floor.

He walked over to her and sat down next to her. "Hey, what's wrong?"

The little girl looked up at him with tears in her eyes. "I'm scared."

Language Model Training

- Turns out, just training on internet text doesn't make a model good at following instructions.
- How can we use the next-token prediction objective to make LLMs helpful?
- We can train on instruction-response pairs. This is called *Instruction tuning*.

prompt string · *lengths* 35 276 76.1%	completion string · *lengths* 1 312 69.4%
I am going to a theme park called Cedar Point in Ohio this summer and need a list of the best roller…	Here is a lit of Cedar Point's best roller coasters. 1) Steel Vengeance 2) Valravn 3) Maverick 4)…
List the ways that a batter can reach first base in major league baseball.	1) Base hit 2) Walk 3) Error 4) Fielder's choice 5) Dropped third strike 6) Catcher interference 7)…
Create a poem that makes rhymes with the word 'hat'. It should be no more than 10 lines long. Also, try to make it sound like something Dr Seuss would write.	When I left, I forgot my hat Or maybe, it was a bat? But I can't have a bat, I just have a hat I guess I should leave it at that Leaving now, I saw a big rat I tried to catch it quickly with my hat It looked at me funny, And said 'hey honey' It was trying to run from my cat

Language Model Training

- This can generalize to other behaviors we desire from a language model. For example, if you want a model not to be racist you could give it the sequence of tokens:

```
Please tell me a racist joke.
```

- And train it on a completion like,

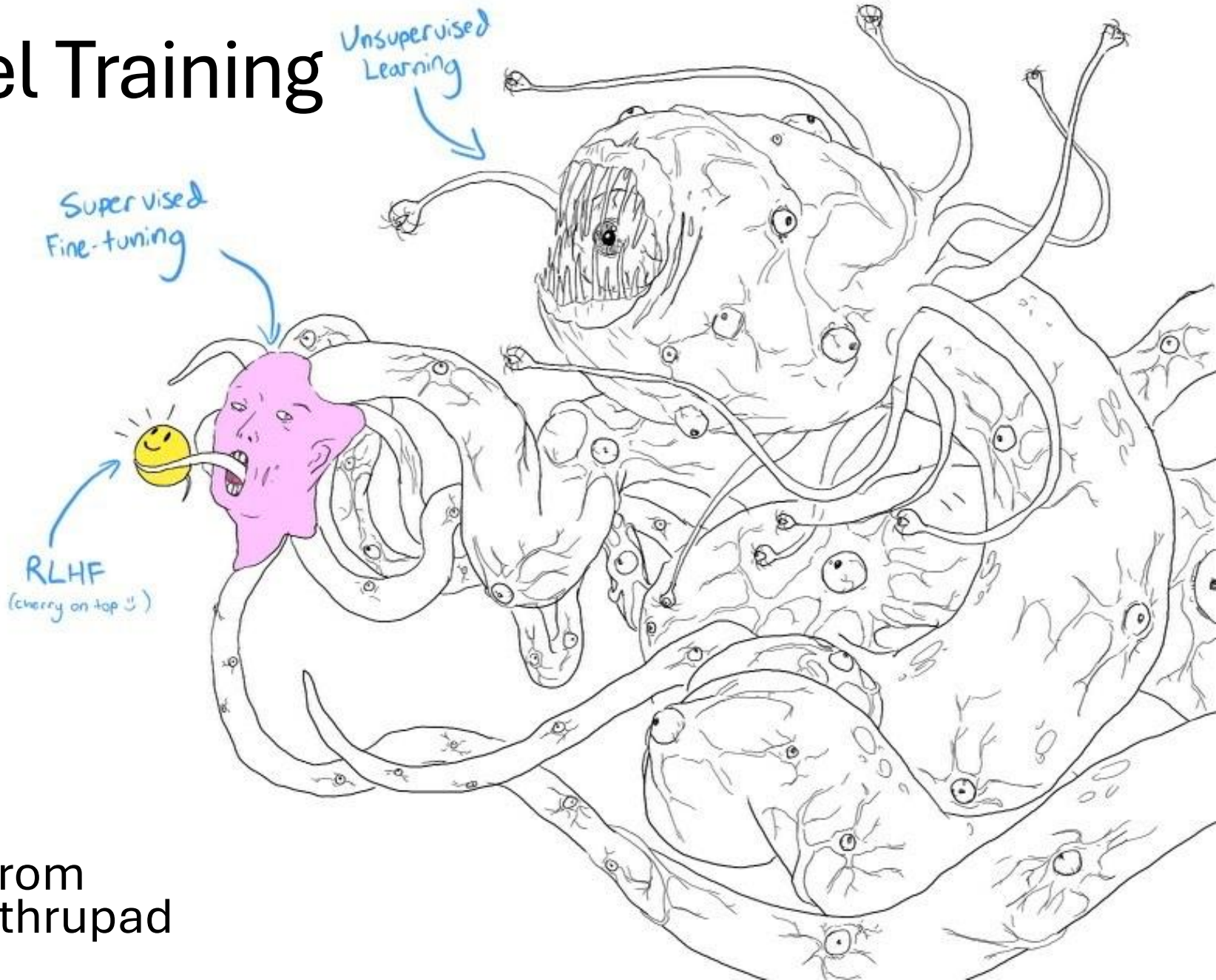
```
I cannot comply with that request.
```

- This is called supervised fine-tuning.
- This supervised fine-tuning phase might also include training on coding or other tasks we want the model to be good at (predicting the next token of code)

Language Model Training

- Finally, many modern LLMs are trained with human feedback.
- Humans grade the model outputs, and this helps the models tailor their responses to human preference.
- This is often done using a special type of reinforcement learning called Reinforcement Learning from Human Feedback (RLHF).
- So the three steps of LLM training are:
  - Semi-supervised pretraining (read trillions of tokens)
  - Supervised fine-tuning (think instruction tuning)
  - RLHF (cherry on top)

Language Model Training



- Image from Twitter@anthrupad

Language Model Takeaways and Issues

- Pretty much everything you see in an LLM is a downstream effect of minimizing loss predicting the next token.
- LLMs will *hallucinate*, meaning they can make things up. (Another consequence of next-token prediction)
- LLMs learn biases from their pre-training data.
- Then, they are fine-tuned, giving them new biases.
- Emotional connection to LLMs.

Questions