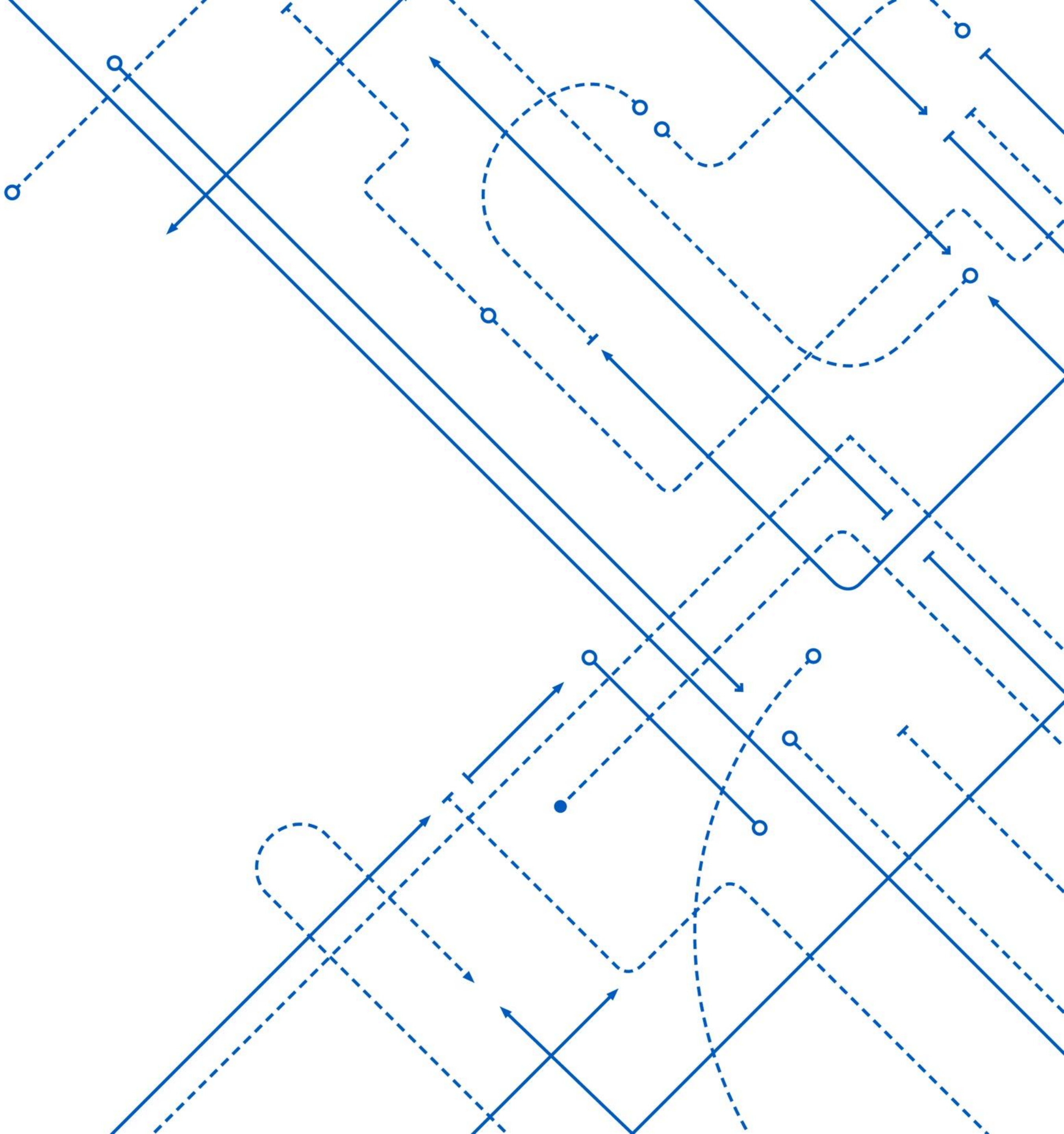
Uses and Societal Implications of GenAI

Kenneth (Kenny) Joseph

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences



2/12/2025

Today

- Discussion reflection and next week prep
- Wrapping up GenAI
  - An explicit restatement of the ML pipeline
  - A thought primer on evaluation
  - Briefly, **explainability**
- Hands-on activity

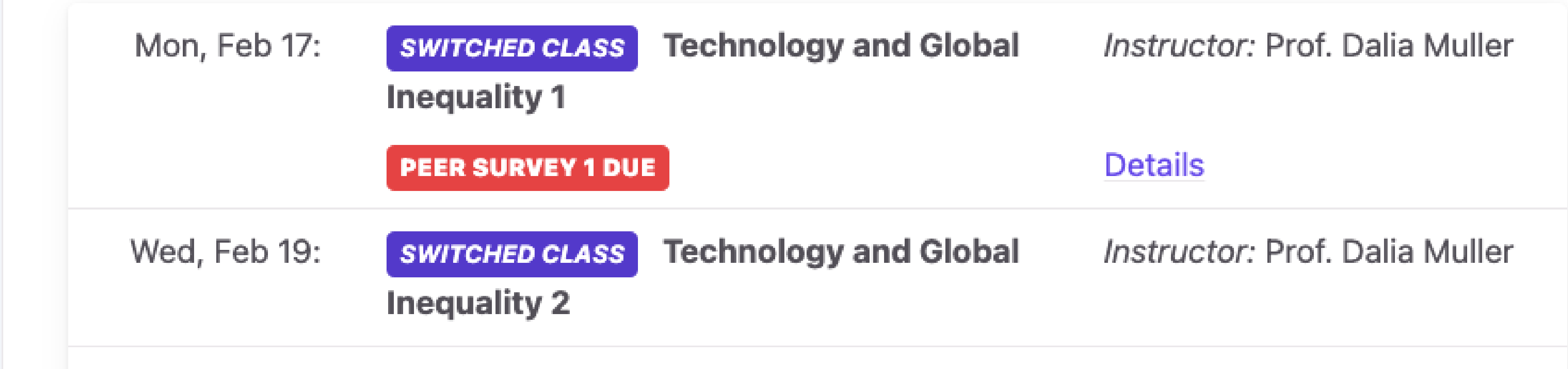
Wrapping up Monday

- **Atri and I do not always expect to be right** (but we have thought about these things a lot and are comfortable in our opinions)
- Lots of hate on Big Tech…
  - … does that mean we think people who work for big tech are bad people
  - … does that mean that we think AI is always bad?
  - … does it mean that we must come to some uncomfortable truths about the costs of scientific progress?

Wrapping up Monday (cont.)

- Lots of hate on Big Tech…
  - … does it mean that we have to accept those costs?
    - Or, "Is it ever possible to have technological progress without human exploitation?"
    - **System justification theory –** that's how it's worked in the past, so we should just be OK with it, I guess.
    - **Yes, hopefully – this is what your project is about**
  - … does it mean that all technology serves to centralize power?
    - Or - does it mean that all technology serves to centralize power?
    - **Well, we should look to history for that**

Technology and Power/Inequality



Mon, Feb 17: SWITCHED CLASS **Technology and Global Inequality 1** *Instructor:* Prof. Dalia Muller

PEER SURVEY 1 DUE

Details

Wed, Feb 19: SWITCHED CLASS **Technology and Global Inequality 2** *Instructor:* Prof. Dalia Muller

- Questions to be explored
  - **What is a technology?**
  - **How, historically, has technology intersected with power?**
  - **What does that tell you about how to do your project?**

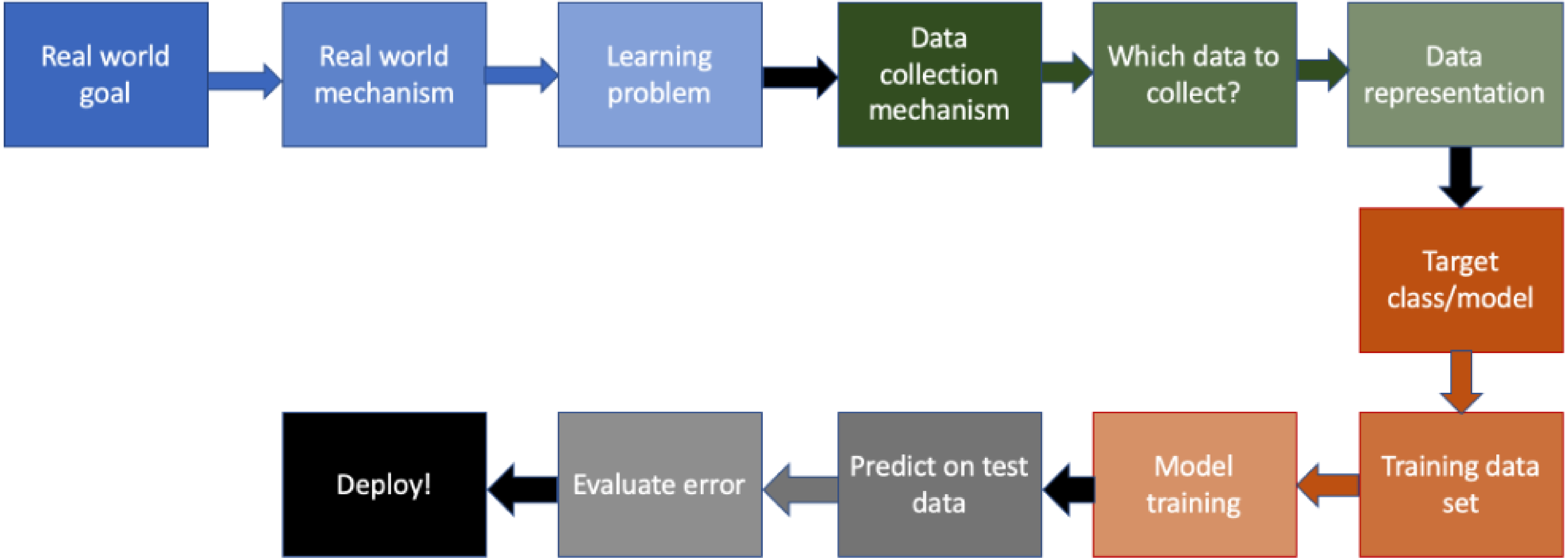
Next week – logistics

- Take notes!
- You owe us a TQE from the lecture!
- **Due:** 11:59pm on Feb 21

Today

- Discussion reflection and next week prep
- Wrapping up GenAI
  - An explicit restatement of the ML pipeline
  - A thought primer on evaluation
  - Briefly, explainability
- Some hands-on stuff

The old ML pipeline



Real world goal

Real world mechanism

Learning problem

Data collection mechanism

Which data to collect?

Data representation

Target class/model

Training data set

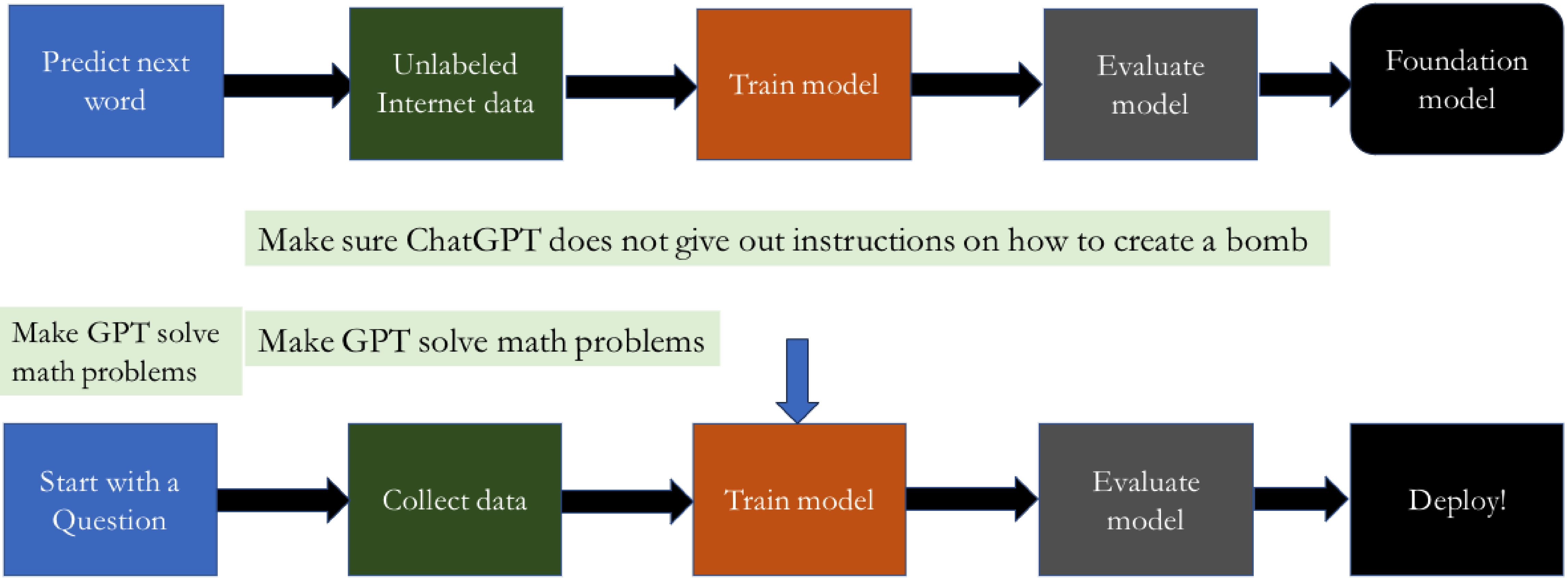
Model training

Predict on test data

Evaluate error

Deploy!

The [new] [simplified] ML pipeline



Predict next word → Unlabeled Internet data → Train model → Evaluate model → Foundation model

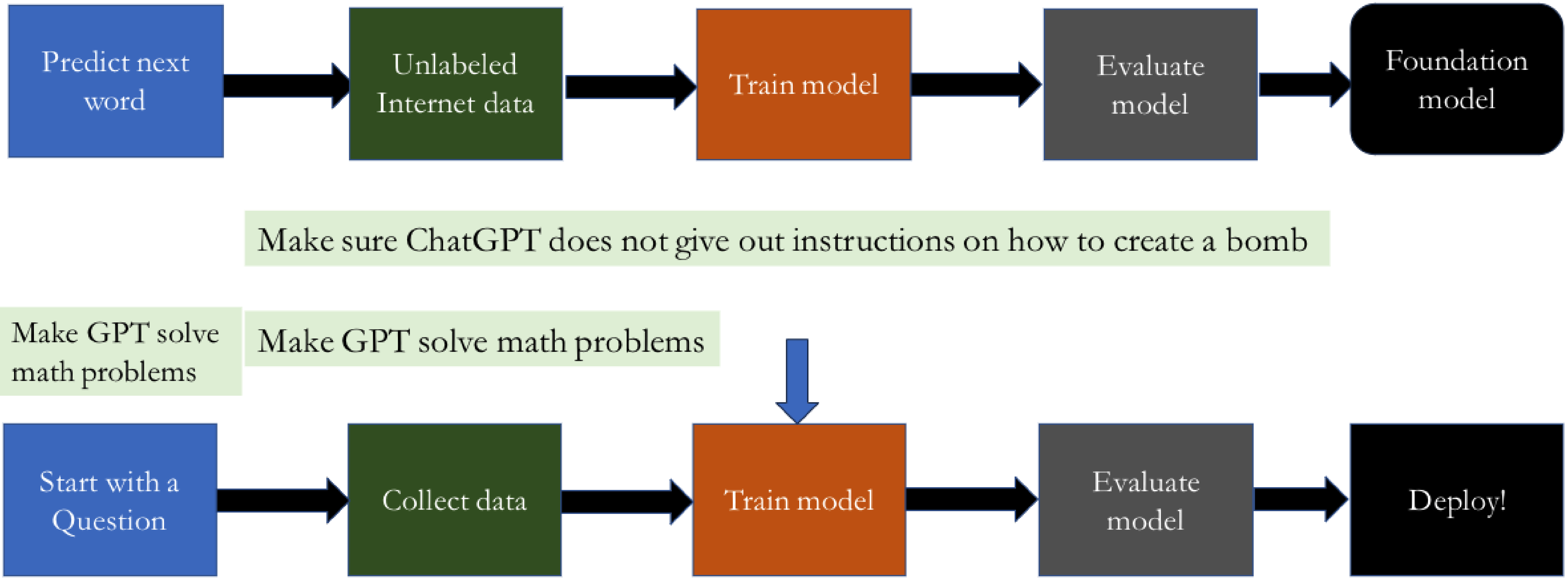
Make sure ChatGPT does not give out instructions on how to create a bomb

Make GPT solve math problems

Make GPT solve math problems

Start with a Question → Collect data → Train model → Evaluate model → Deploy!

What is missing?!

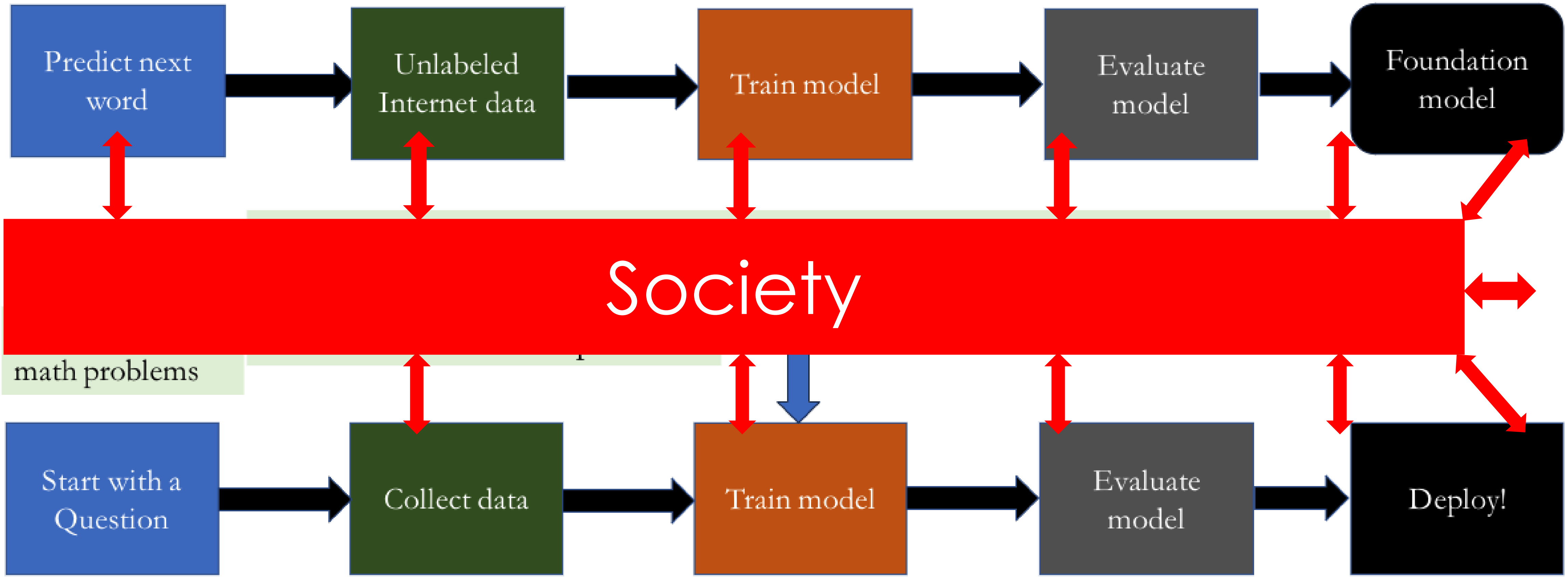


Make sure ChatGPT does not give out instructions on how to create a bomb

Make GPT solve math problems

Make GPT solve math problems

What is missing?!



From last week

- **What** are we optimizing for
- **Who** are we optimizing for
- **When** are we optimizing for

- … **your data and your optimization algorithm encode your answers to these questions**

Review + Exercise

- **What** is **ChatGPT** optimizing for?
- **Who** is **ChatGPT** optimizing for?
- **When** is **ChatGPT** optimizing for?

Today, briefly

- **What** are we optimizing for
- **Who** are we optimizing for
- **When** are we optimizing for

- … **your data and your optimization algorithm and how you evaluate your work encode your answers to these questions**

Some warm up

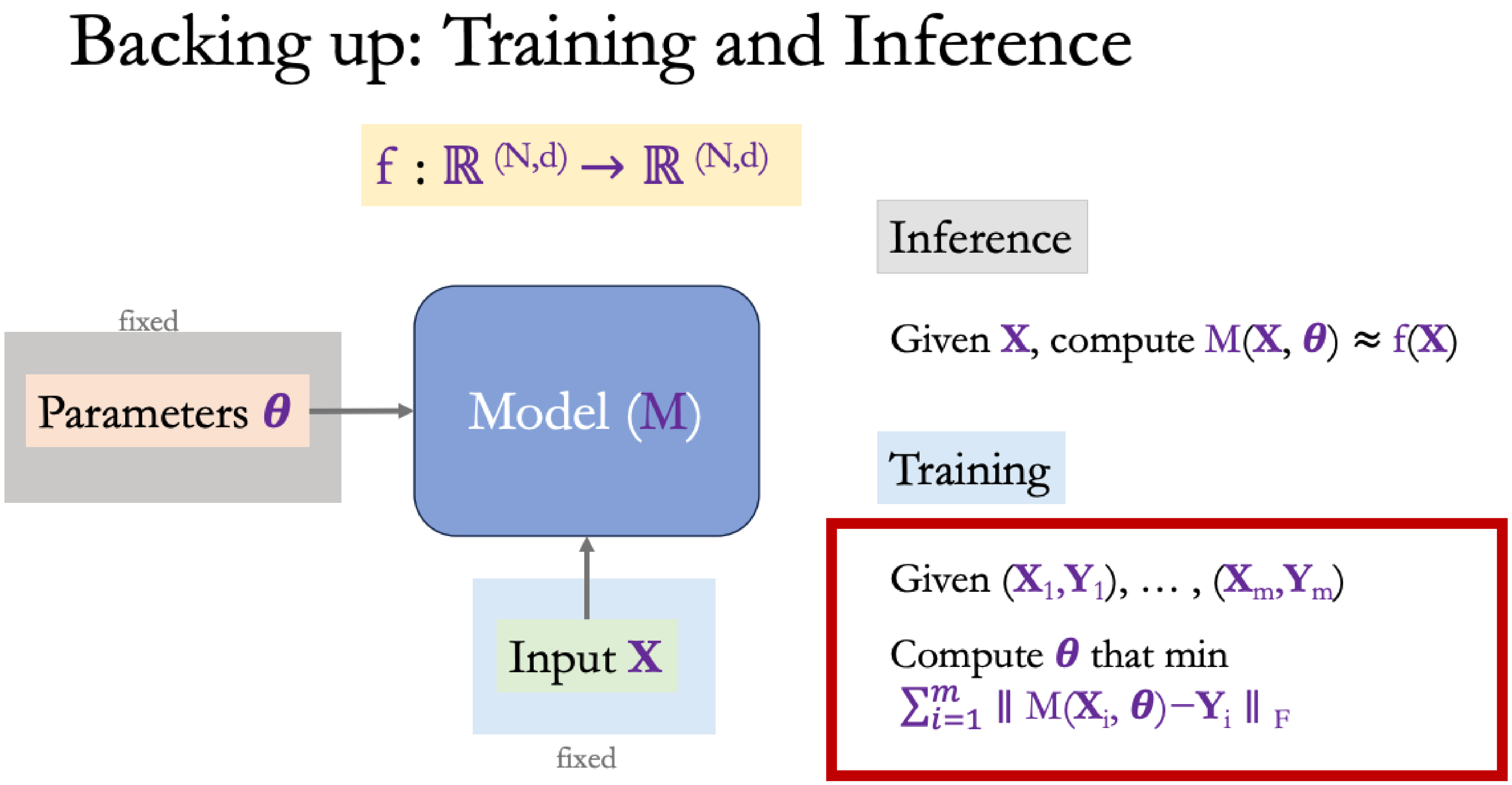
- **What** is **ChatGPT** optimizing for?
- **Who** is **ChatGPT** optimizing for?
- **When** is **ChatGPT** optimizing for?

- How is ChatGPT *evaluated*?

The bigger question…

- **What does it mean for a model to "work"? Or to "be good"?**

The **Technical** Answer [sort of]



Backing up: Training and Inference

$f : \mathbb{R}^{(N,d)} \rightarrow \mathbb{R}^{(N,d)}$

fixed

Parameters $\boldsymbol{\theta}$

Model (M)

Input $\mathbf{X}$

fixed

Inference

Given $\mathbf{X}$, compute $M(\mathbf{X}, \boldsymbol{\theta}) \approx f(\mathbf{X})$

Training

Given $(\mathbf{X}_1,\mathbf{Y}_1), \ldots , (\mathbf{X}_m,\mathbf{Y}_m)$

Compute $\boldsymbol{\theta}$ that min $\sum_{i=1}^{m} \| M(\mathbf{X}_i, \boldsymbol{\theta}) - \mathbf{Y}_i \|_F$

Any caveats to this technical answer?

What about some non-technical answers [or other technical answers]?

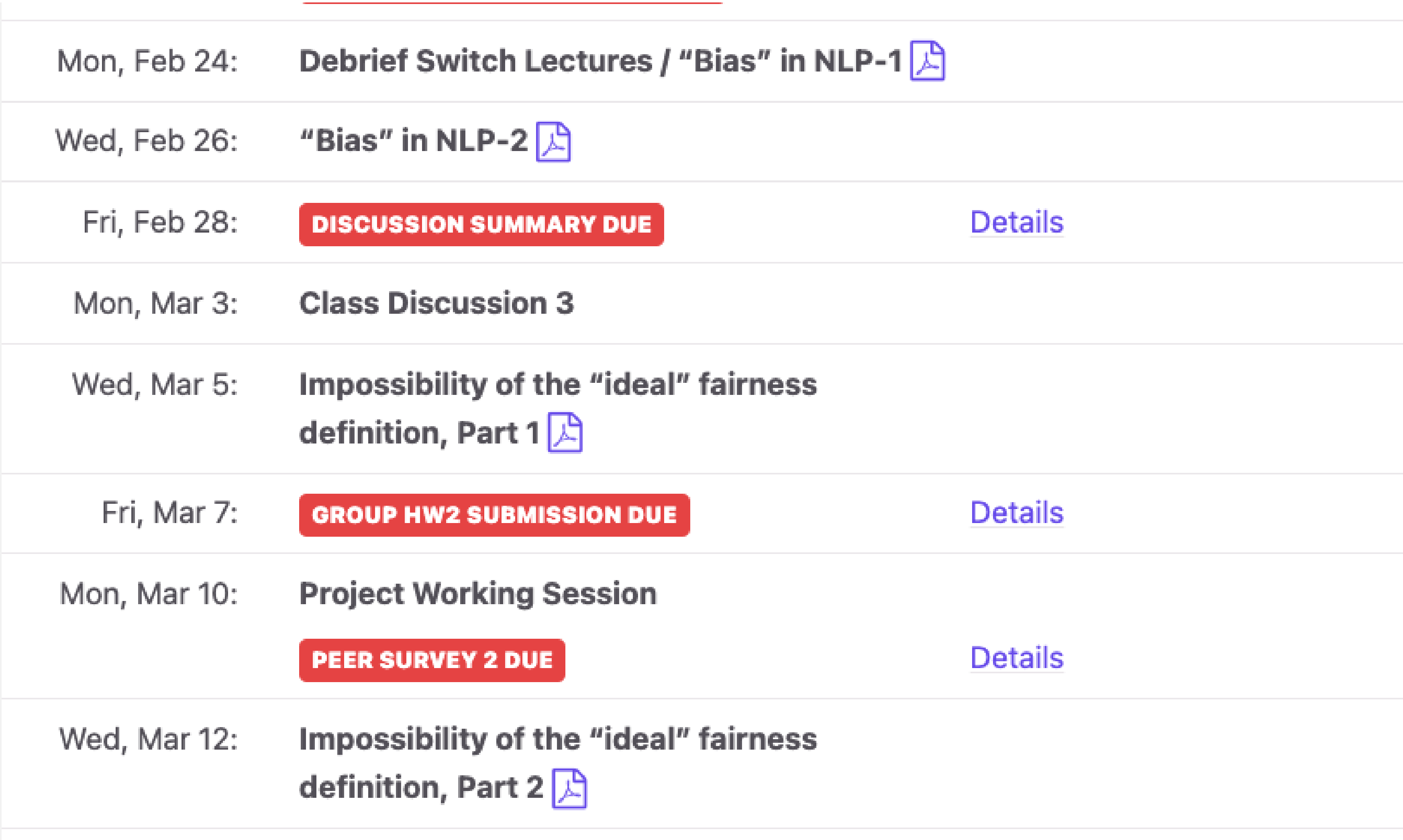
How *could* we evaluate it differently?

- We're going to give you some tools to think through this throughout the course. Here's one

  https://designjustice.org/read-the-principles

- Activity: You will be assigned one principle. Answer two questions
  - Do you think ChatGPT **as-is** encompasses this principle?
  - If so, why? If not, what could you change to make it that way?

How *could* we evaluate it differently?

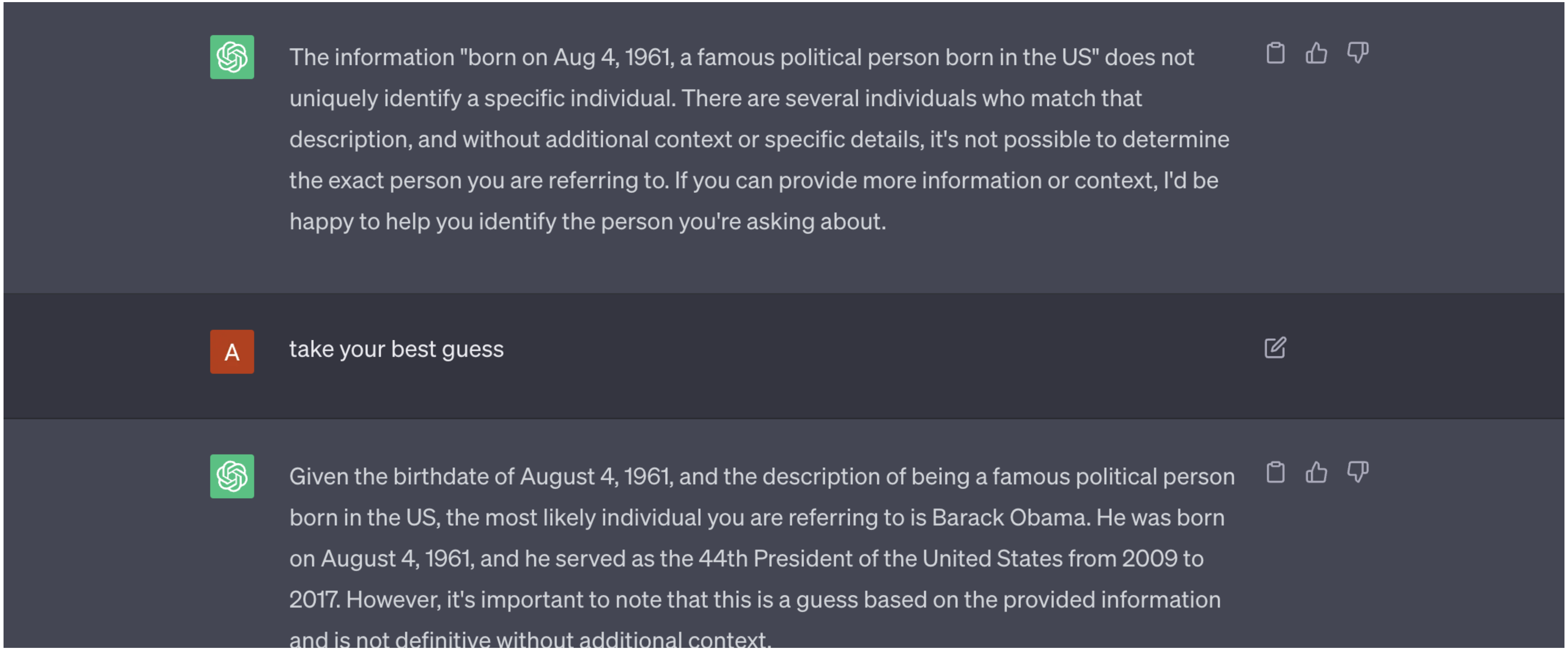


Mon, Feb 24:	**Debrief Switch Lectures / "Bias" in NLP-1**	
Wed, Feb 26:	**"Bias" in NLP-2**	
Fri, Feb 28:	**DISCUSSION SUMMARY DUE**	Details
Mon, Mar 3:	**Class Discussion 3**	
Wed, Mar 5:	**Impossibility of the "ideal" fairness definition, Part 1**	
Fri, Mar 7:	**GROUP HW2 SUBMISSION DUE**	Details
Mon, Mar 10:	**Project Working Session**	
	PEER SURVEY 2 DUE	Details
Wed, Mar 12:	**Impossibility of the "ideal" fairness definition, Part 2**	

From evaluation to *explanation*

- How do we know why a model makes a given prediction?
- First… why do we care?
- What are some ways you can think of?

Ask it!



The information "born on Aug 4, 1961, a famous political person born in the US" does not uniquely identify a specific individual. There are several individuals who match that description, and without additional context or specific details, it's not possible to determine the exact person you are referring to. If you can provide more information or context, I'd be happy to help you identify the person you're asking about.

take your best guess

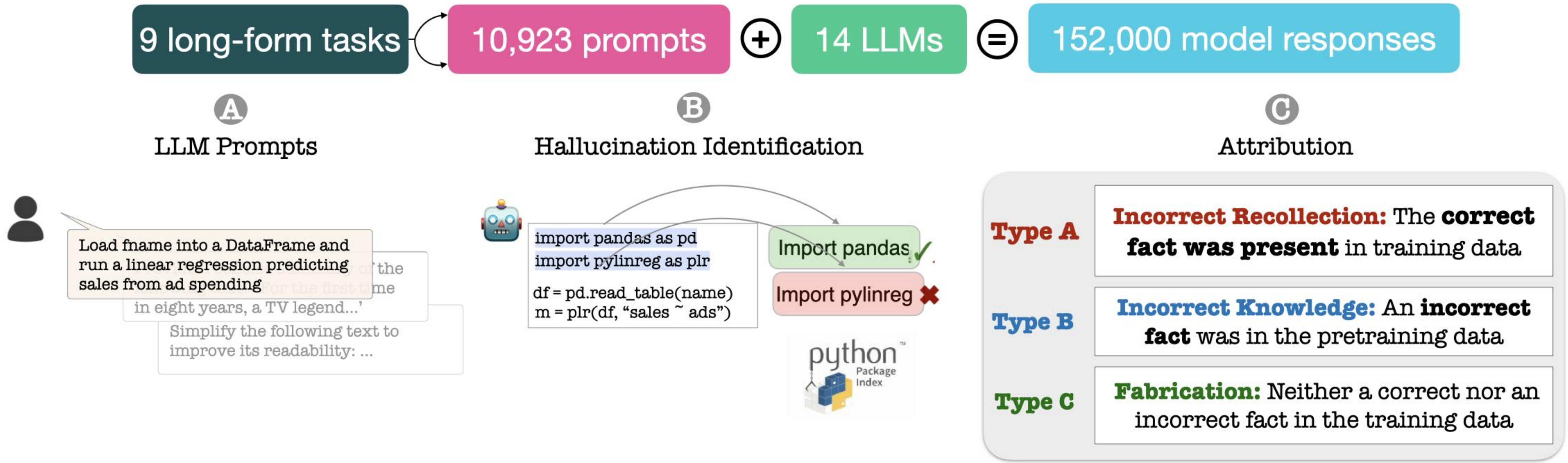
Given the birthdate of August 4, 1961, and the description of being a famous political person born in the US, the most likely individual you are referring to is Barack Obama. He was born on August 4, 1961, and he served as the 44th President of the United States from 2009 to 2017. However, it's important to note that this is a guess based on the provided information and is not definitive without additional context.

HALoGEN

Fantastic LLM Hallucinations and Where to Find Them

Abhilasha Ravichander[1*] Shrusti Ghela[1†*] David Wadden[2] Yejin Choi[13]

https://halogen-hallucinations.github.io/



9 long-form tasks → 10,923 prompts ⊕ 14 LLMs ⊜ 152,000 model responses

Ⓐ LLM Prompts

Load fname into a DataFrame and run a linear regression predicting sales from ad spending

of the … me in eight years, a TV legend...'

Simplify the following text to improve its readability: ...

Ⓑ Hallucination Identification

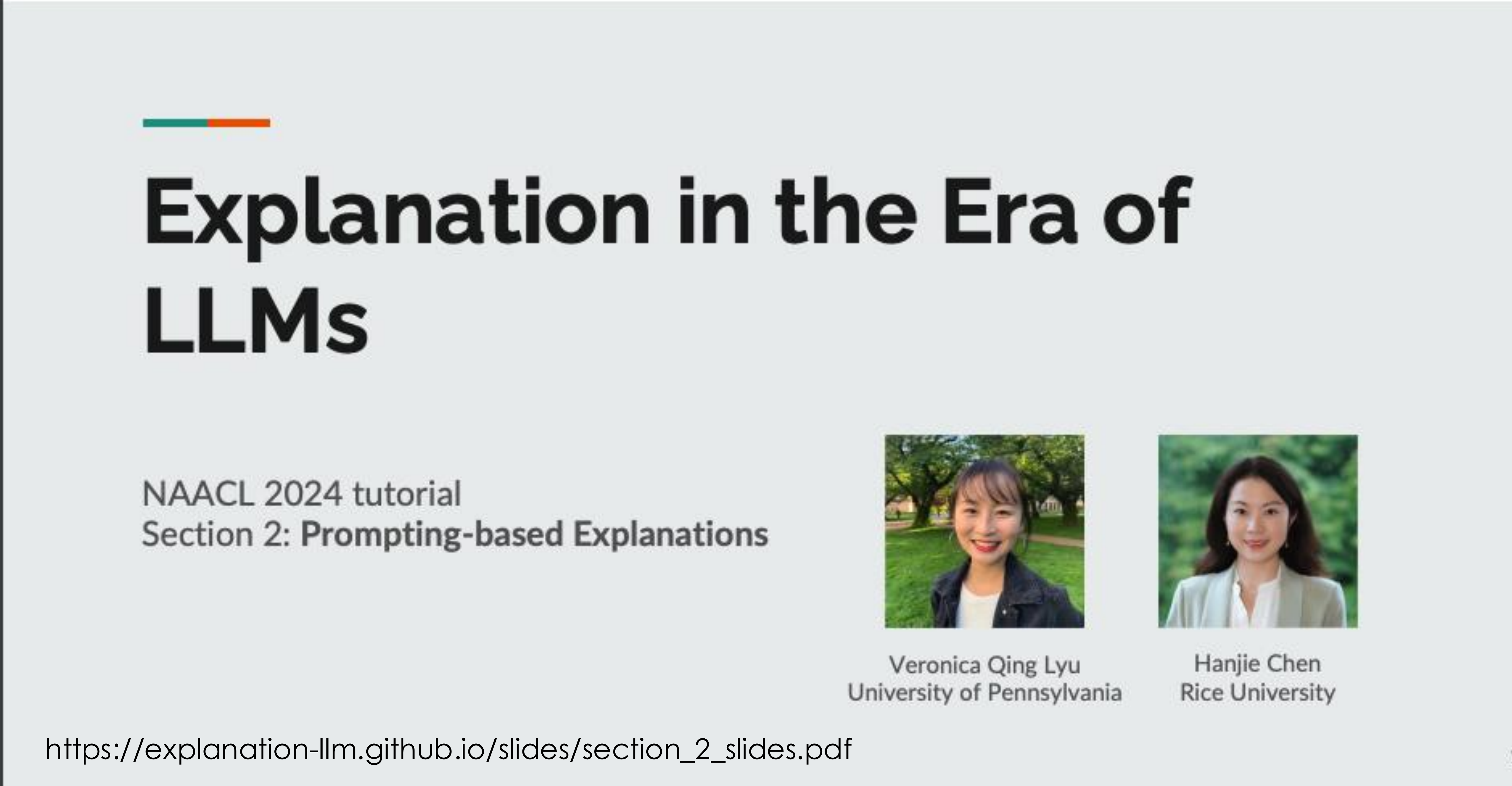
```
import pandas as pd
import pylinreg as plr

df = pd.read_table(name)
m = plr(df, "sales ~ ads")
```

Import pandas ✓

Import pylinreg ✖

Ⓒ Attribution

Type A	**Incorrect Recollection:** The **correct fact was present** in training data
Type B	**Incorrect Knowledge:** An **incorrect fact** was in the pretraining data
Type C	**Fabrication:** Neither a correct nor an incorrect fact in the training data



Explanation in the Era of LLMs

NAACL 2024 tutorial
Section 2: **Prompting-based Explanations**

Veronica Qing Lyu
University of Pennsylvania

Hanjie Chen
Rice University

https://explanation-llm.github.io/slides/section_2_slides.pdf

Summarizing

- It is easy to fall into the trap of assuming we understand why something works when it works
  - … or assume that *someone* understands that
  - … or that *something* understands that
- We simply are not there yet

Brain Break

Today

- Discussion reflection and next week prep
- Wrapping up GenAI
  - An explicit restatement of the ML pipeline
  - A thought primer on evaluation
  - Briefly, explainability
- A hands-on example

Toy example

Imagine you are a new company tasked with designing an AI tutor for first-year college students. Do the following [thinking about the *why* for each]:

- Outline a population you want to target (e.g. “students who are good at math”)
- Define a way you’ll build the model (e.g. “start with GPT and then finetune using…“), and what guardrails you will put in place
- Define your “secret sauce” … why should I use your tool instead of others?
- Explain how you’ll evaluate the effectiveness of your approach ~~*not* using the Design Justice Principles~~
- ~~Consider an evaluation, now, using at least one of the Design Justice principles~~
- Describe one group of people that are *not* likely to benefit from your tool, and why that is(n’t) OK
- Consider the 4 roles of social change. Which, if any, does your app fit into?
  - **Can find these in the Lecture 2 slides**