



CS329X: Human Centered NLP

Human-Centered NLP

Diyi Yang

Stanford CS

Welcome to CS 329X

Instructor and CA



Diyi Yang



Rishi Bommasani

Course Overview

Website:

<http://web.stanford.edu/class/cs329x>

Ed Discussion:

<https://edstem.org/us/courses/38475/discussion/>

What is human-centered NLP?

“Human-centered NLP involves designing and developing NLP systems in a way that is attuned to the needs and preferences of human users, and that considers the ethical and social implications of these systems.”

– ChatGPT, 2022

What is human-centered NLP?

It concerns NLP systems, which goes beyond just the model – also includes e.g. user interfaces on top of the model.

It touches multiple NLP dev stages.

*“Human-centered NLP involves **designing and developing** NLP **systems** in a way that is attuned to the **needs and preferences of human users,** and that considers the **ethical and social implications** of these systems.”*

– ChatGPT, 2022

It needs to be optimized for humans.

“Optimize for humans” require careful ethical concerns.

What if this definition is wrong?

What is a **right definition** of human-centered NLP?

Why should we build human-centered NLP?

“The common misconception [is] that language use has primarily to do with words and what they mean. It doesn't. It has primarily to do with people and what they mean.



Herbert H. Clark and Michael F. Schober. 1992. Asking questions and influencing answers. Questions about Questions: Inquiries into the Cognitive Bases of Surveys, pages 15-48

Corrective
Preventive
Not Reactive



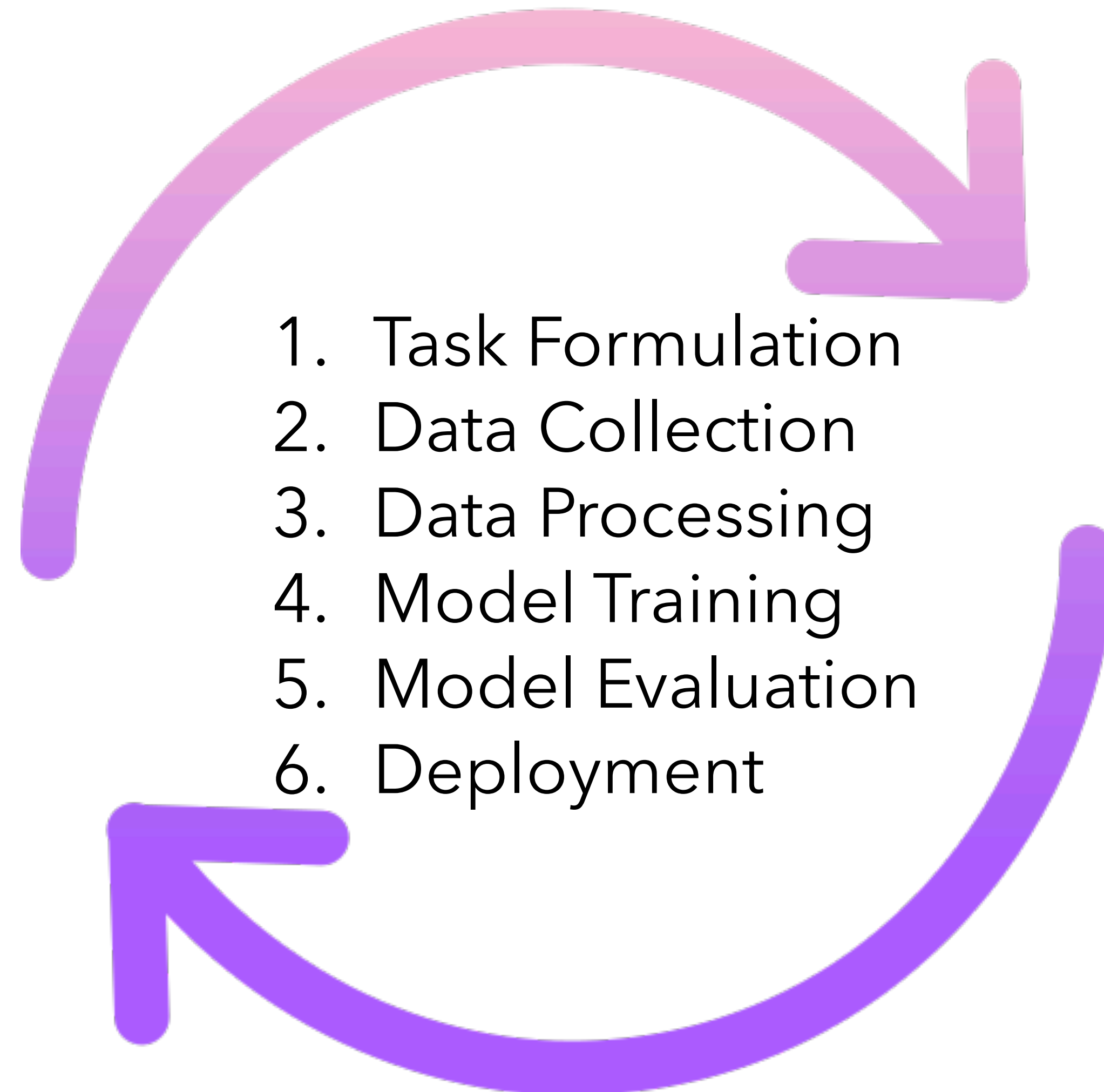
Who is the “human” in human-centered NLP?

The "human" refers to the focus on designing and developing NLP technologies that ***prioritize human needs and preferences***, rather than solely focusing on technological capabilities.

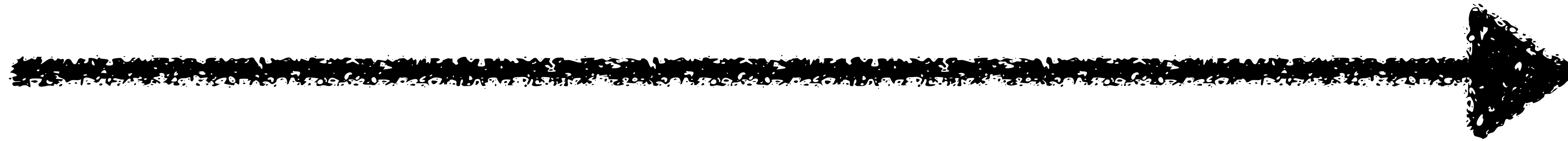
Human-centered NLP seeks to create NLP systems that are more **user-friendly**, **accessible**, and **inclusive**.



Human-centered NLP should be in every stage

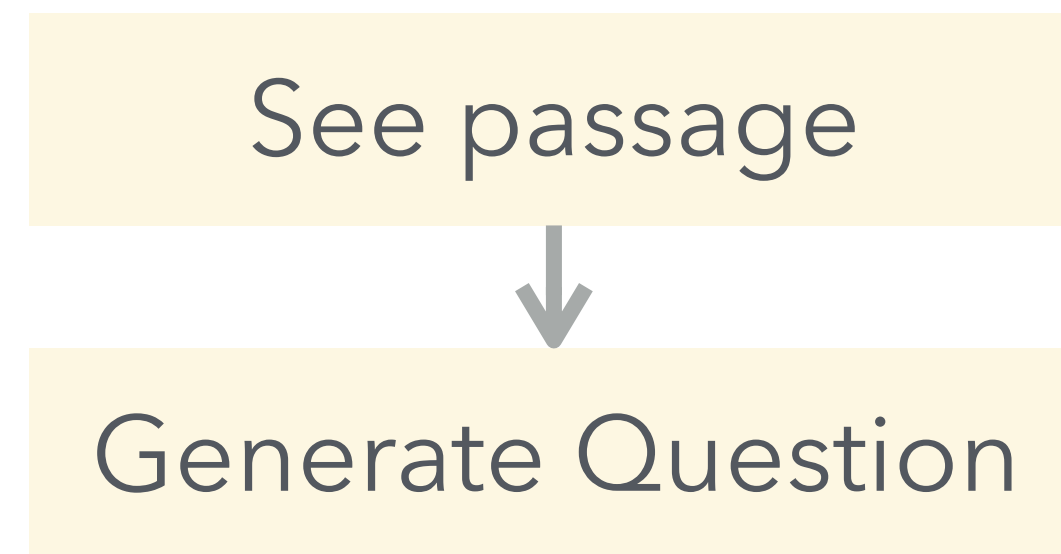


Human-centered NLP is a spectrum



Traditional Example on Data Collection

*“Stanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text.”*



Paragraph 1 of 43

Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on 'Select Answer', and then highlight the smallest segment of the paragraph that answers the question.

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O₂.

2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

Ask a question here. Try using your own words

Select Answer

Ask a question here. Try using your own words

Select Answer

Traditional Example on Data Collection: **Issues**

Crowdworkers are hired to generate questions in a constrained setting, which could be different from how people generally ask questions.

Crowdworkers generate questions specific to one paragraph & are primed to generate questions of a certain style. Questions created in this way have a high degree of lexical overlap with the document text and thus models might rely too heavily on word matching.

Q: What component of **water** is more **soluble** than **nitrogen**?

Doc: Oxygen is more **soluble** in **water** than **nitrogen** is. Water in equilibrium with air contains approximately 1 molecule of dissolved O₂ for every 2 molecules of N₂, compared to an atmospheric ratio of approximately 1:4. ...

A: Oxygen

Q: What is the **world's largest academic and private library system**?

Doc: Harvard ... operates several arts, cultural, and scientific museums, alongside the Harvard Library, which is the **world's largest academic and private library system**, comprising 79 individual libraries with over 18 million volumes...

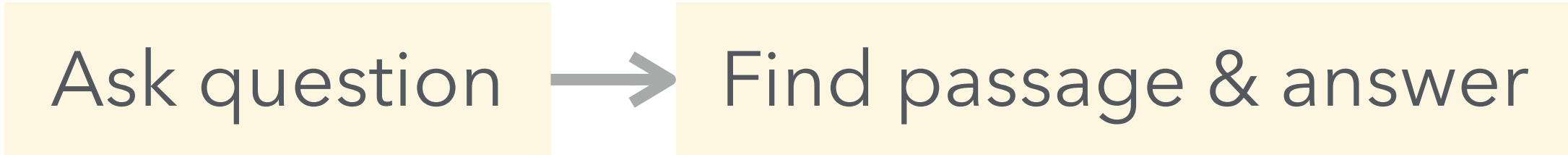
A: Harvard Library

More Human-Centered Data Collection

Google Natural Question: Questions consist of *real anonymized, aggregated queries* issued to the Google search engine. An annotator is presented with a question along with a Wikipedia page from the top 5 search results, and annotates a long answer (typically a paragraph) and a short answer (one or more entities) if present on the page, or marks null if no long/short answer is present.

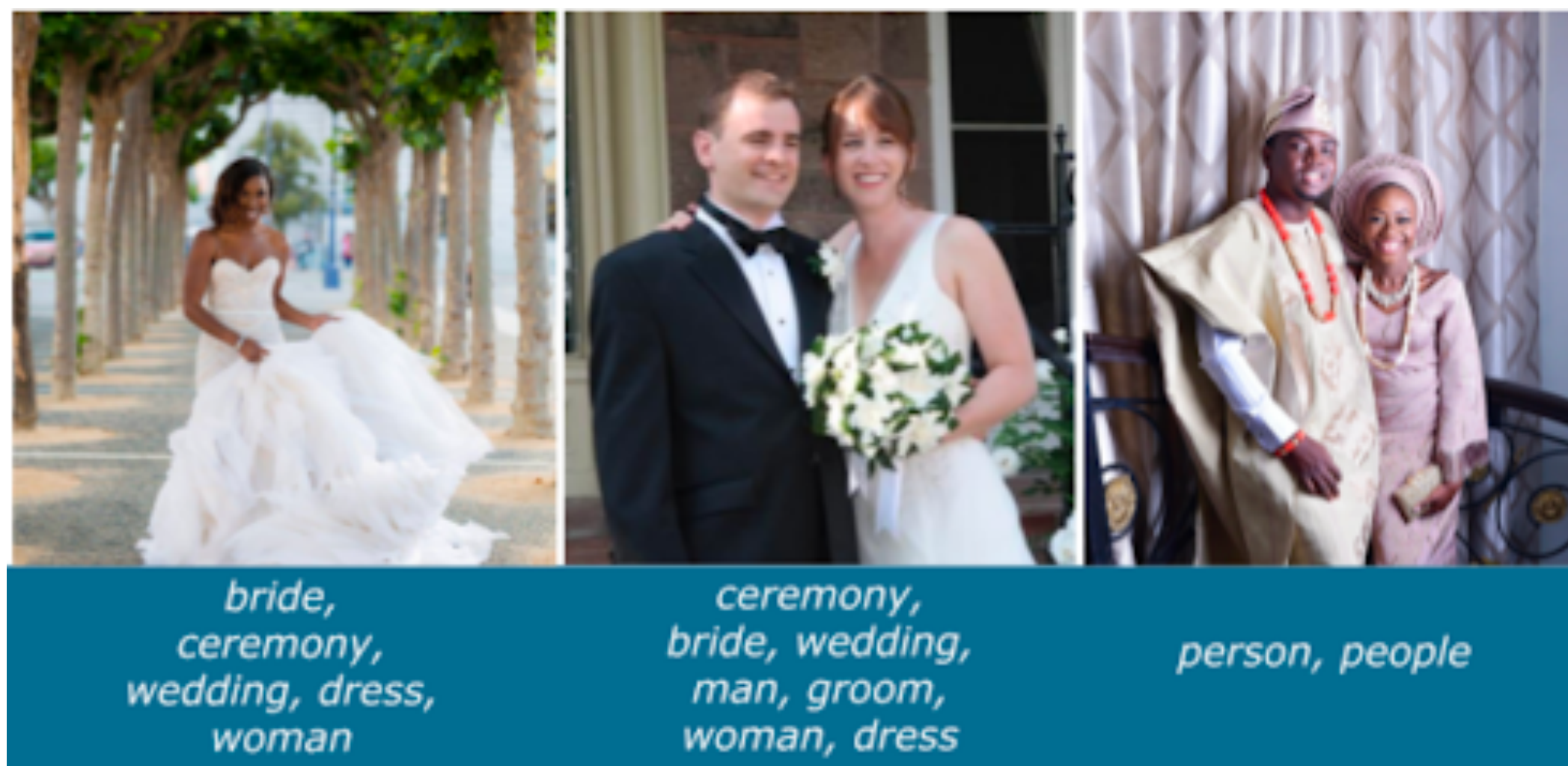
The image shows a screenshot of a Wikipedia page for Kendrick Lamar. At the top, a black bar contains the question: "Q: When did Kendrick Lamar's first album come out?". Below the question, the Wikipedia article is visible, including the title "Kendrick Lamar", a "From Wikipedia" notice, and a paragraph of text. A small inset image of Kendrick Lamar is also present on the right side of the page. The page layout includes navigation links like "Article" and "Talk", and a search bar.

Natural Questions: a New Corpus and Challenge for Question Answering Research



More Human-Centered Data Collection

Data collection schema, if designed unnaturally, can introduce annotation artifact (unwanted patterns in the data), and can be harmful or unfair to certain groups.



Example: Image classifier trained on the Open Images dataset that does not properly apply “wedding” related labels to images of wedding traditions from different parts of the world.

Human-Centered data collection should focus on mimicking real-use scenarios so the data will **reflect actual human needs**.

Introducing the Inclusive Images Competition

Traditional Model Training

PROMPT

Explain the moon landing to a 6 year old in a few sentences.

COMPLETION

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

*“The GPT-3 language model can be **coaxed** to perform natural language tasks using carefully engineered text prompts. But these models can also generate outputs that are untruthful or toxic. This is in part because GPT-3 is **trained to predict the next word on a large dataset of Internet text**, rather than to perform the language task that the user wants.”*

Maybe More Human-Centered Model Training

PROMPT

Explain the moon landing to a 6 year old in a few sentences.

COMPLETION

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

Reinforcement learning from human feedback (RLHF): make models safer, more helpful, and **more aligned**

“On prompts submitted by our customers to the API, our labelers provide demonstrations of the desired model behavior, and rank several outputs from our models. We then use this data to fine-tune GPT-3.”

Maybe More Human-Centered Model Training

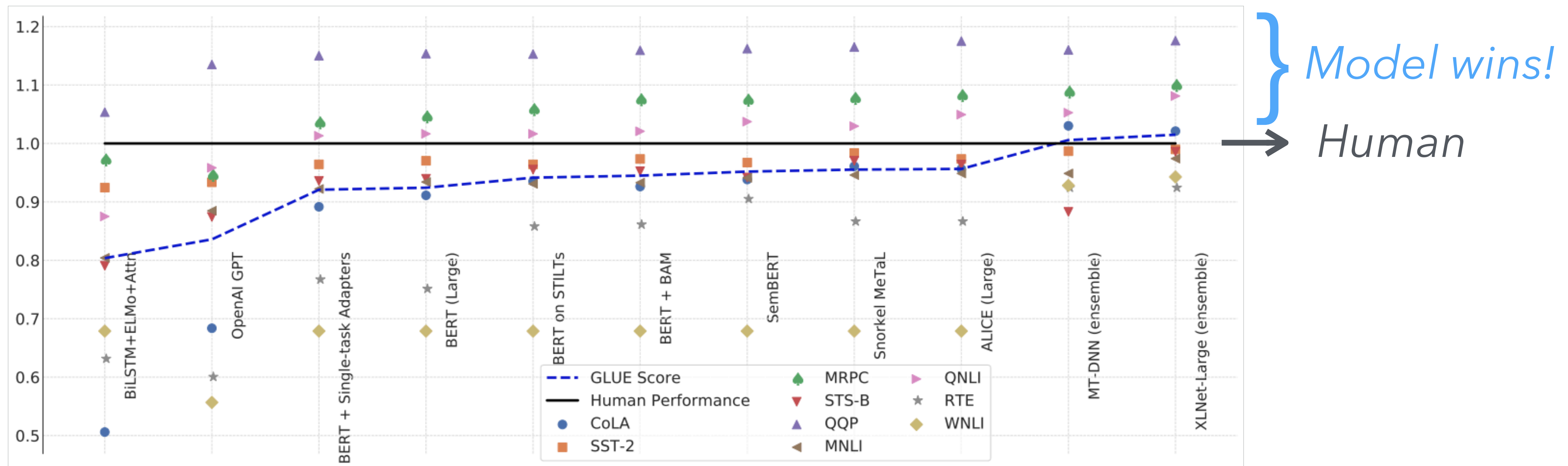
“The resulting InstructGPT models are much better at following instructions than GPT-3... Our labelers prefer outputs from our 1.3B InstructGPT model over outputs from a 175B GPT-3 model, despite having more than 100x fewer parameters.”

*“it “unlocks” capabilities that **GPT-3** already had, but were difficult to elicit through prompt engineering alone.”*

Compared to standard training (which gives models many hidden power), slightly more human-centered training **might make** models more usable.

Traditional Evaluation

“performance on the benchmark has recently come close to the level of non-expert humans, suggesting limited headroom for further research.”



Accuracy on benchmarks is the most standard assessment on model quality, but it **does not** contain enough signal!

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). SuperGlue: A stickier benchmark for general-purpose language understanding systems. NeurIPS.

Traditional Evaluation

Shortcuts/right for wrong reasons



What is the moustache made of?

> Banana

What are the *eyes* made of?



What is?

> Banana


What?

> Banana

More Human-Centered Evaluation

“Capability testing”, quantify human expectations and requirements on models


INVARIANCE TEST

test name	failure rate
change neutral words with BERT	51 / 500 = 10.2% 

Test Summary




Test [INV] on [VOCABULARY]
change neutral words with BERT

Desc. Change a set of neutral words with other context-appropriate neutral words (using BERT).

Result **FAILURE RATE ON ALL CASES**
51/500=10.2% 

FILTER TEST CASES

Examples Failed cases only

- > @united I will not be flying you → back again
Pred: 2 (0.71) → 0 (0.99) 
- > @AmericanAir you → never called back just to put me on hold . It 's midnight . Literally just want to know how I 'm getting home and I 'm getting no help
Pred: 2 (0.81) → 0 (0.99) 
- > @AmericanAir Yes I am . 2495/1170 . RNO departure at 1229 on 2/25 w / connection at DFW to → and LGA . I can do the
Pred: 2 (0.84) → 1 (0.65) 

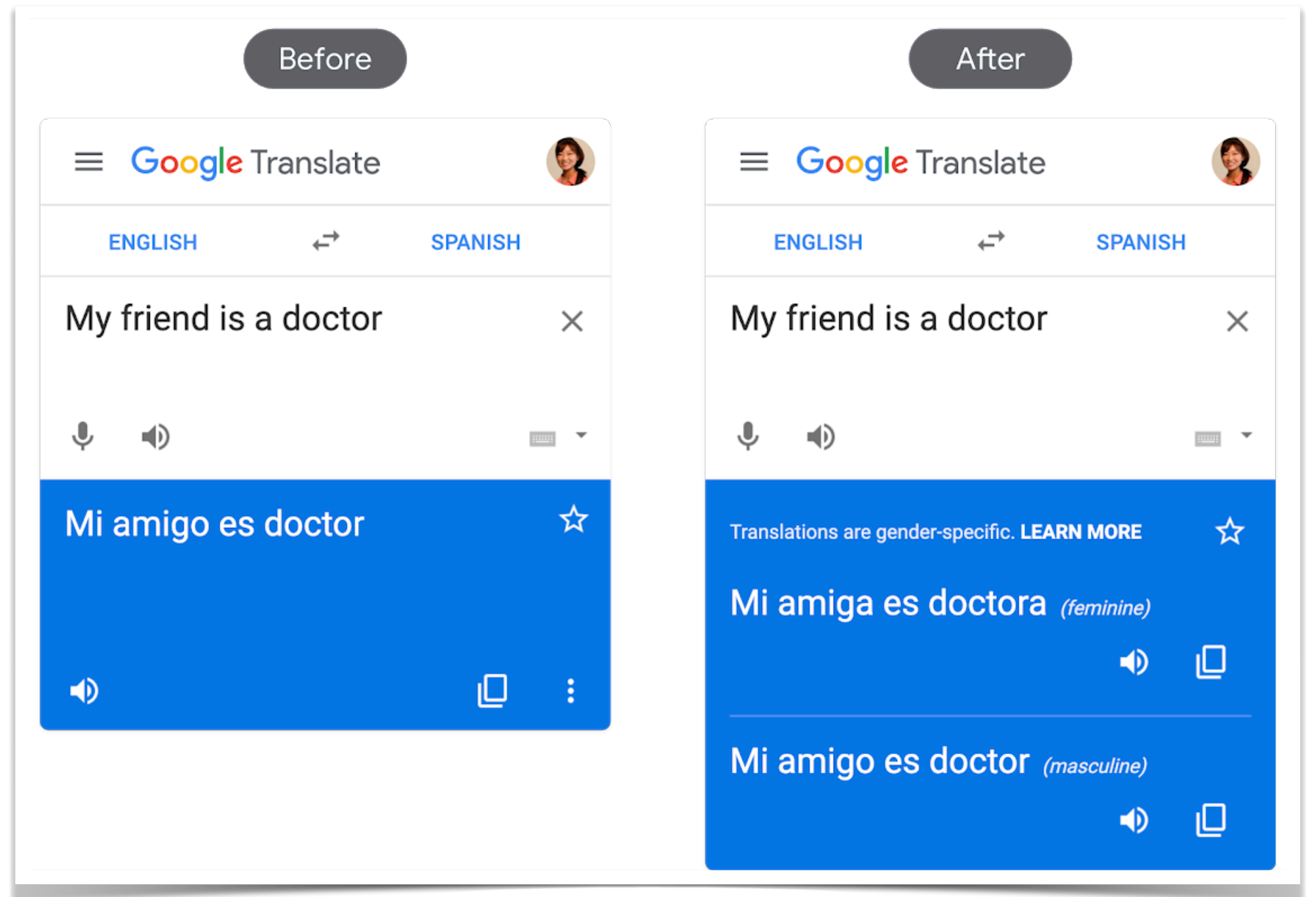
More Human-Centered Evaluation

Standard evaluations cannot capture model shortcuts.

Instead, human-centered evaluation should design **fine-grained** metrics and analysis strategies that account for **user-specific** interaction objectives, cognitive loads, etc.

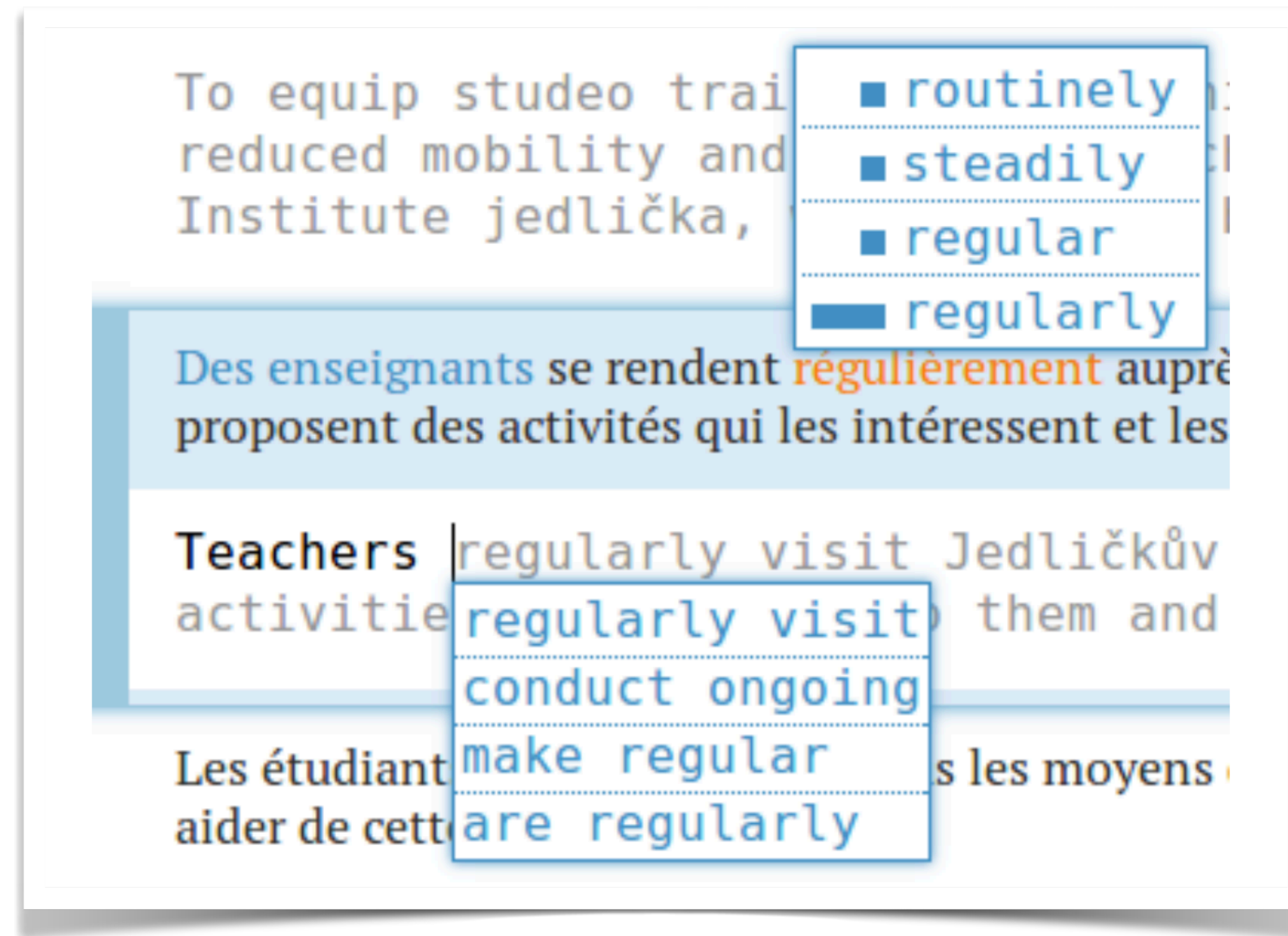
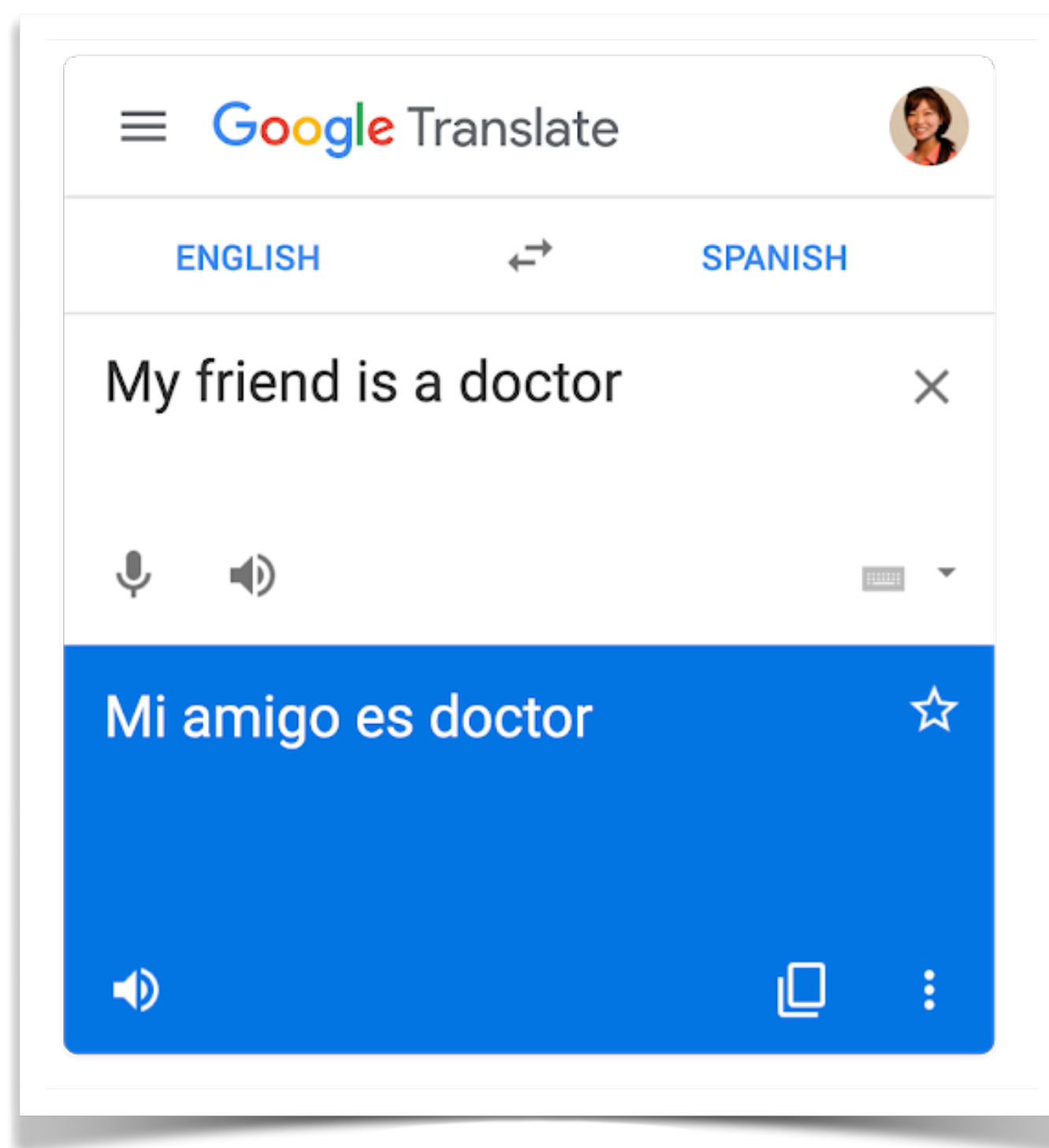
Human-Centered Deployment

Some parts of a task is inherently hard. For example, nouns in some language is gender-specific but not in others, and translator won't be able to "add" this information, so the disambiguation needs to be handled by the **interface**.



Human-Centered Deployment

Similar models also need to have very different interaction wrappers in different use cases.



Human-Centered Deployment

Who is going to use the system?

Who is going to design the system?

How would users use the system?

What interface can best facilitate such interaction?

Handling exceptions & designing interactions for the right user group goes beyond modeling, and is essential for deployment.

Why should we care about human-centered NLP

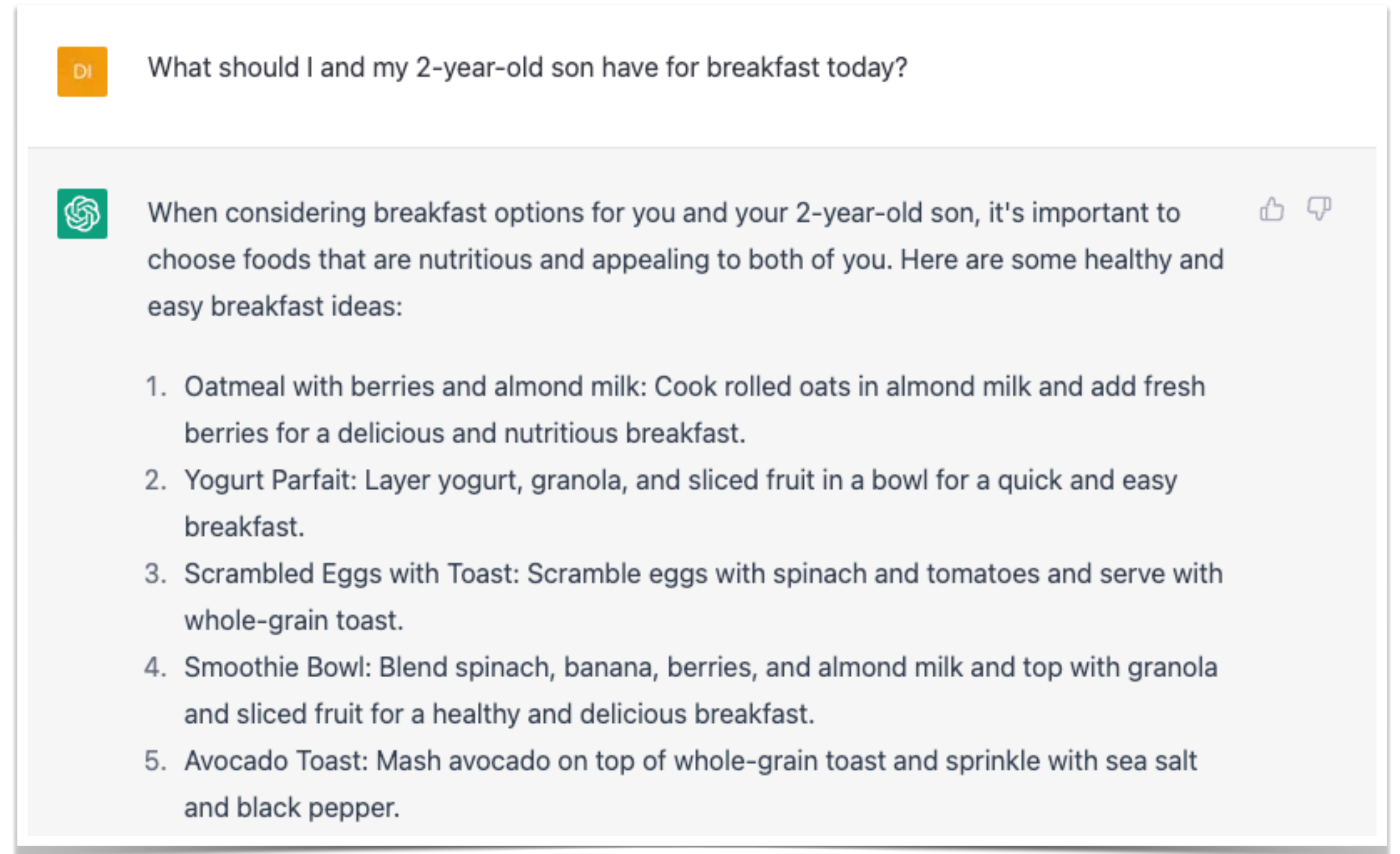
Improve user experiences

Build effective, accountable and responsible systems

Benefit the society as a whole

What if our NLP systems are not human-centered?

- Biased results
- Inaccurate interpretation
- Lack of culture awareness
- Lack of personalization
- Security risks
- ...

A screenshot of a chatbot interface. At the top, a user asks, "What should I and my 2-year-old son have for breakfast today?". Below, the chatbot responds with a helpful message and a list of five breakfast ideas. The chatbot's response includes a thumbs-up and thumbs-down icon in the top right corner.

DI What should I and my 2-year-old son have for breakfast today?

When considering breakfast options for you and your 2-year-old son, it's important to choose foods that are nutritious and appealing to both of you. Here are some healthy and easy breakfast ideas:

1. Oatmeal with berries and almond milk: Cook rolled oats in almond milk and add fresh berries for a delicious and nutritious breakfast.
2. Yogurt Parfait: Layer yogurt, granola, and sliced fruit in a bowl for a quick and easy breakfast.
3. Scrambled Eggs with Toast: Scramble eggs with spinach and tomatoes and serve with whole-grain toast.
4. Smoothie Bowl: Blend spinach, banana, berries, and almond milk and top with granola and sliced fruit for a healthy and delicious breakfast.
5. Avocado Toast: Mash avocado on top of whole-grain toast and sprinkle with sea salt and black pepper.

AI that's not human-centered can be harmful

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

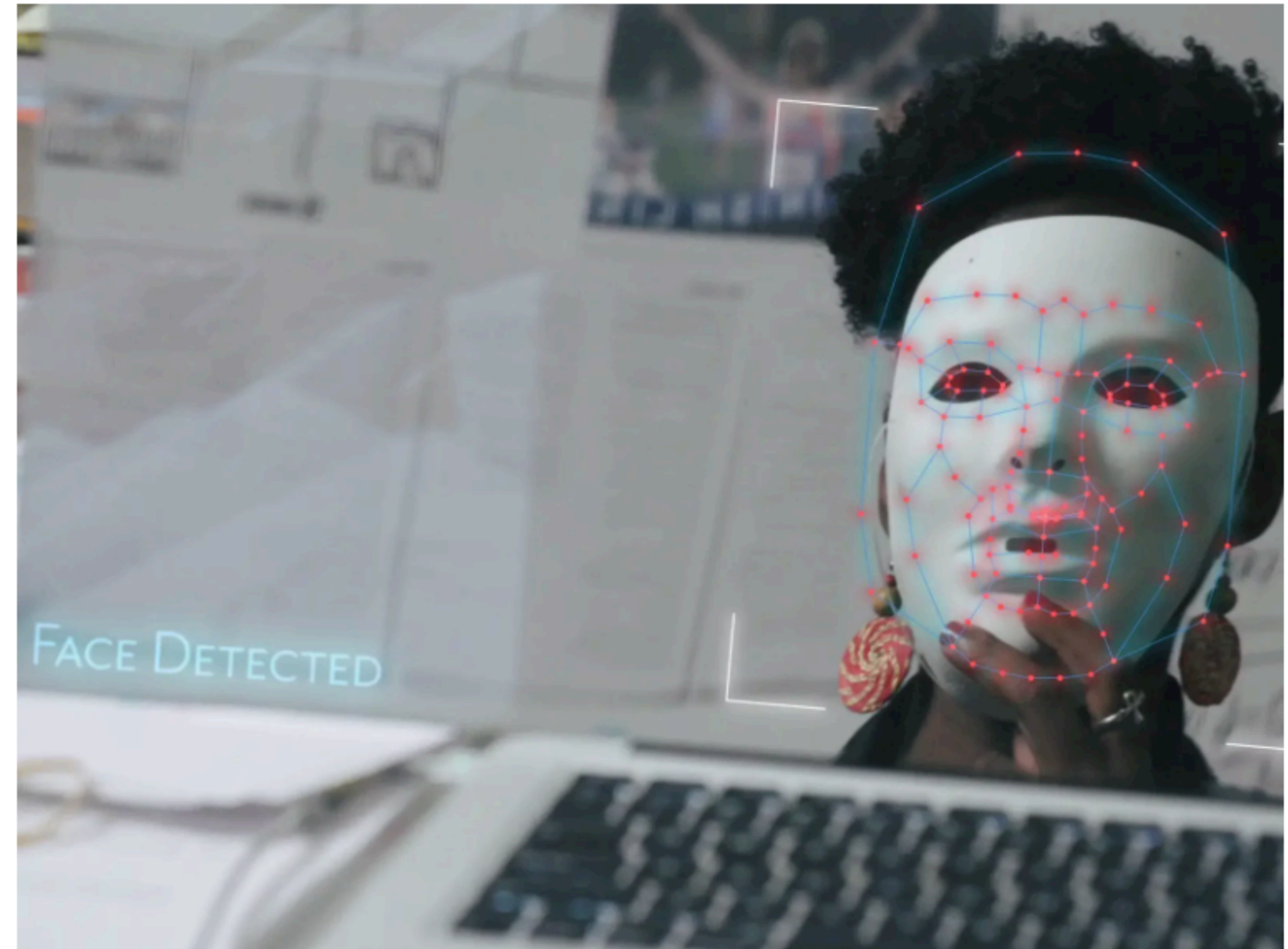


Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates. Photograph: Brian Snyder/Reuters

When Bias Is Coded Into Our Technology

February 8, 2020 · 6:03 AM ET

By Jennifer 8. Lee




Courtesy of the 2050 Group

AI that's not human-centered can be harmful

There Is a Racial Divide in Speech-Recognition Systems, Researchers Say

Technology from Amazon, Apple, Google, IBM and Microsoft misidentified 35 percent of words from people who were black. White people fared much better.

 Give this article



NEWS & COMMENTARY

Algorithms Are Making Decisions About Health Care, Which May Only Worsen Medical Racism

Unclear regulation and a lack of transparency increase the risk that AI and algorithmic tools that exacerbate racial biases will be used in medical settings.



A.I. Bias Caused 80% Of Black Mortgage Applicants To Be Denied

[Visit](#)

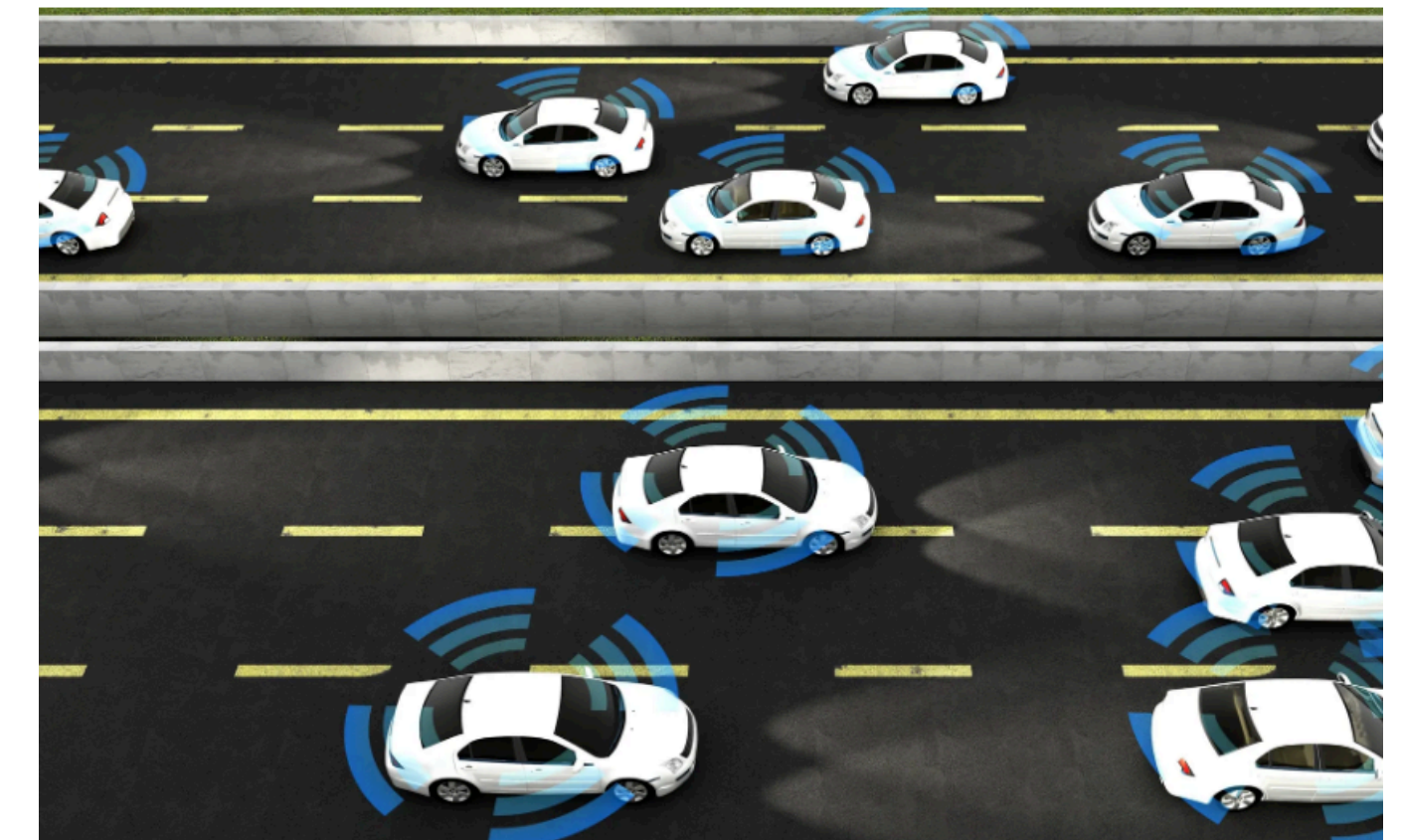


A new study finds a potential risk with self-driving cars: failure to detect dark-skinned pedestrians

The findings speak to a bigger problem in the development of automated systems: algorithmic bias.

By Sigal Samuel | Updated Mar 6, 2019, 10:50am EST

   SHARE



Modern AI Systems are learning from **human preferences**

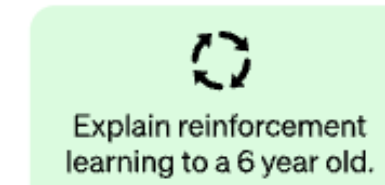


<https://openai.com/blog/chatgpt>

Step 1

Collect demonstration data and train a supervised policy.

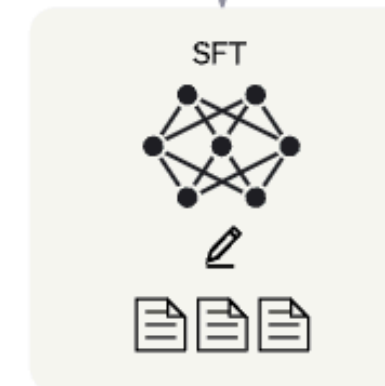
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



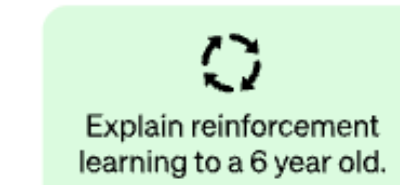
This data is used to fine-tune GPT-3.5 with supervised learning.



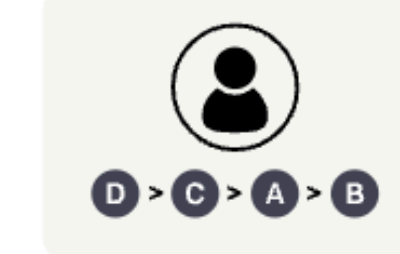
Step 2

Collect comparison data and train a reward model.

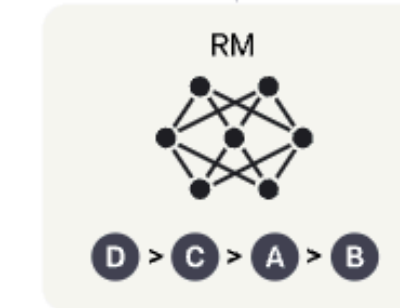
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

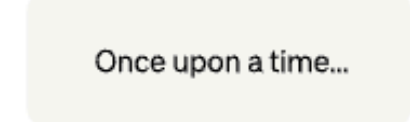
A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.



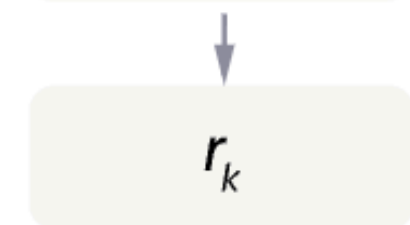
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



How to Make NLP More Human-Centered?

NLP people know the standard method of data preparation, training, evaluation, and deployment.

HCI people know ways to mimic natural use scenario, collect human feedback, design interactions...

Both are needed for Human-Centered NLP.

What do you want to learn from CS329X?

Join at
slido.com
#3615 081



Course Goals

The primary goal of the course is **offer an overview of HCI+NLP**, and to help students get access to, and understand, both HCI and NLP research papers and methods. More specifically, we will:

- Introduce basic concepts of NLP and HCI

- Introduce a variety of emerging topics related to human-centered NLP

- Work together to think about and define what is human-centered NLP**

This course is co-designed with Prof. Sherry Wu at CMU.

Course Topics

NLP basics

1. The ultimate crash: NLP tasks and applications

Model

2. Human-centered design principles

Design

3. Human in the loop

4. Learning from human feedback

*Data and
Evaluation*

5. Human-centered evaluation

6. Data collection and curation

7. Beyond benchmarking

8. Interpretability and explanation

Safety/Trust

9. Privacy, security and safety

10. Trust in AI

Impact

11. Human-AI collaboration

12. NLP for social impact

Guest Lectures on Emerging Topics



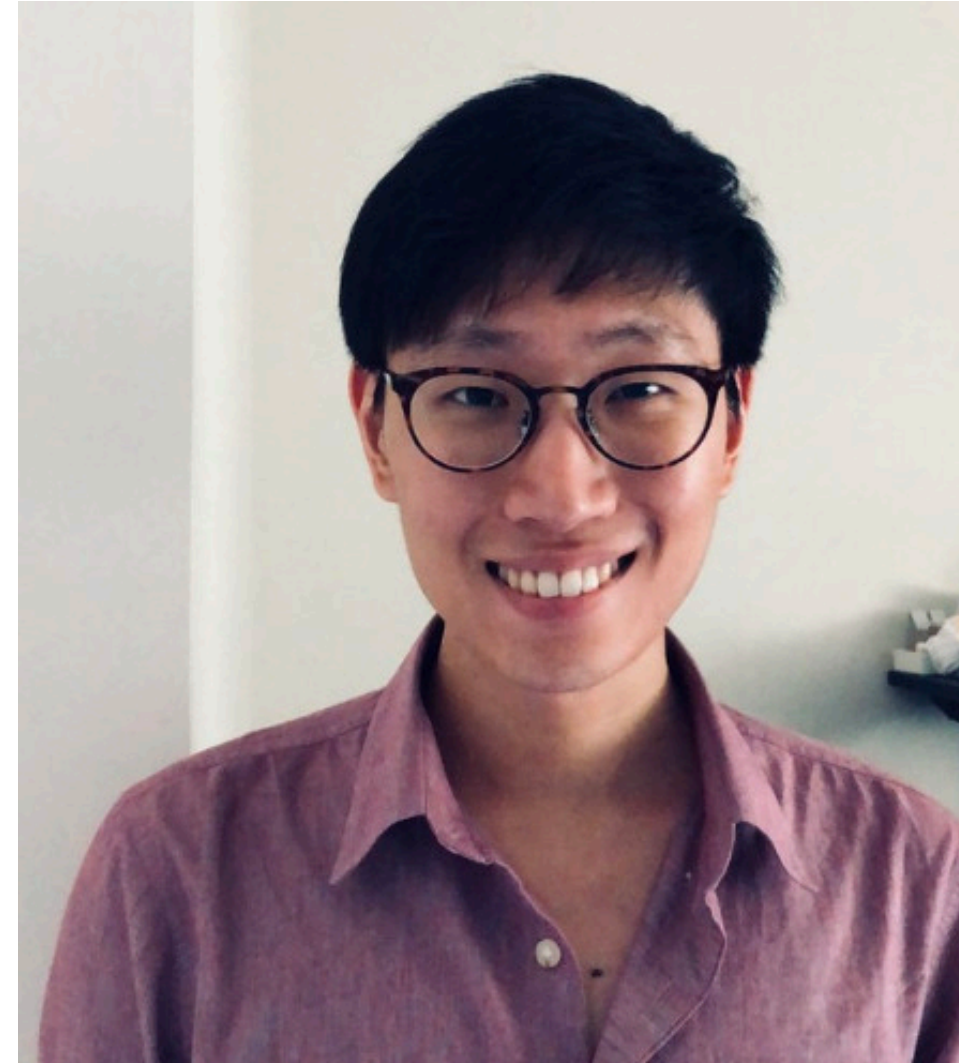
Sherry Wu
(CMU)

Interactive NLP
and Visualization



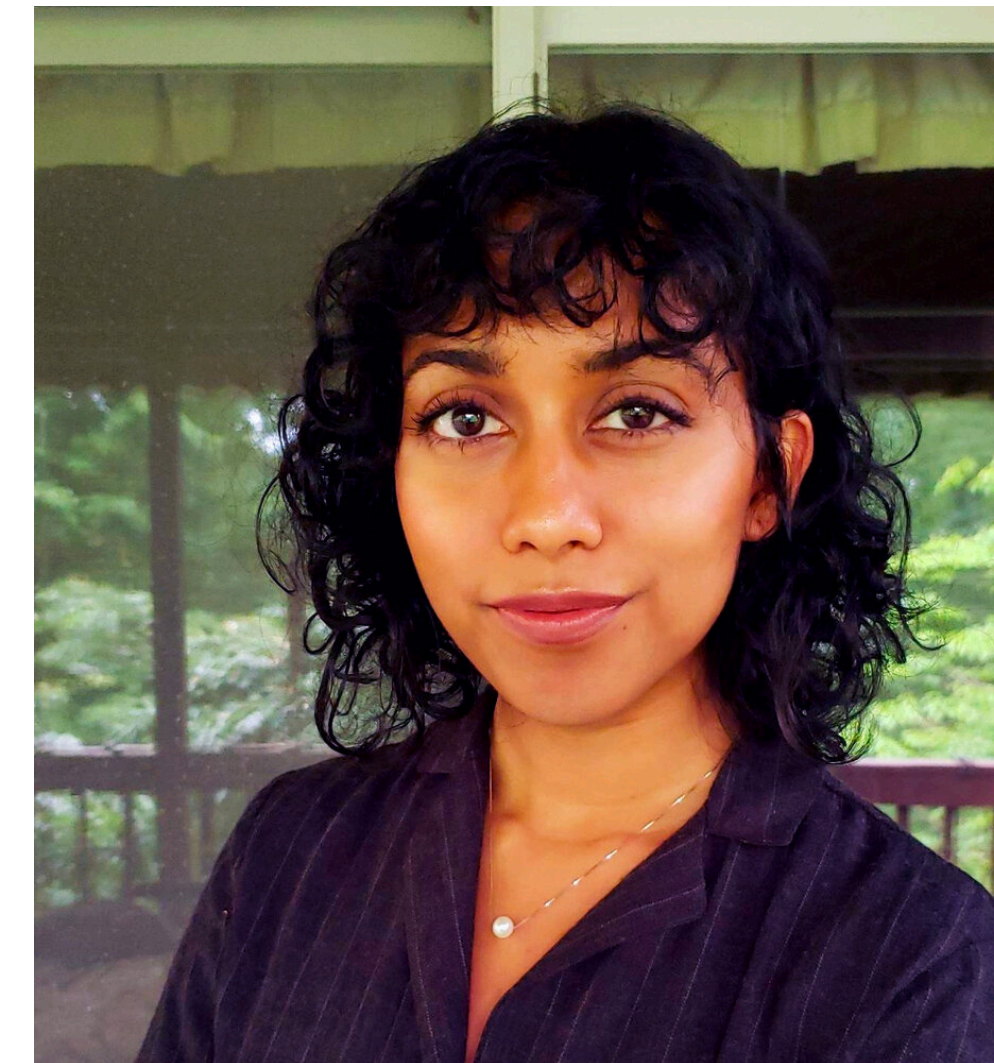
Rishi Bommasani
(Stanford)

Ethics and Social
Responsibility



Joon Sung Park
(Stanford)

Simulation via
Foundation Models



Irene Solaiman
(Huggingface)

AI Governance

What does CS 329X provide?

Provides an overview of HCI + NLP

Provides various aspects of what a usable NLP system looks like

Does not provide an in-depth intro to deep learning for NLP (see CS224N)

Prerequisites

CS 224N or CS224U, or equivalent.

You are expected to...

Be proficient in Python (for completing project)

Know basic NLP concept – To the extent that you understand concepts like train/dev/test set, model fitting, feature, supervised learning, etc. (We will not cover these in this course!)

Major Course Work

30% Homework Assignments

15% One assignment

15% One scribe

65% Final Project

5% In-Class project presentation

10% Literature review

20% Experiment protocol

5% Final poster presentation

25% Project report

5% Participation (discussion in class or via Ed Discussion)

Scribe

Please sign up for **one slot** for this course (one week to complete your scribe).

Scribe template will be released by TA soon.

3 pages (excluding references) for a one-person team

5 pages if two people team together

Expectation: a coherent summary of lecture content, with in-depth coverage of technical details and discussions.

Assignment

Will be released on Apr 12th, due on May 1st
2~3 problem sets related to course content

Course Project

10% Literature review

20% Experiment protocol

5% Final poster presentation

25% Project report

Project Scope (there will be a separate lecture on this):

One key element to justify: what is the human-centered aspect in your project?

Case studies of human factors in existing NLP/AI systems

New mythologies tailored to a human-centered problem

Position papers or a critic (talk to us first)

Applying a computational text-based method to real-world problem for social good

In-Person Expectation

This is an in-person class by default, but if you cannot come, **there's a zoom option on Canvas**. This link will not be on the webpage to avoid zoom boom :)

Computing Credits

We provide two types of credits (stay tuned for emails on how to access):

- Computing credit on Google Cloud
- Credits for accessing foundation models

Late Days

- Each student will have a total of 4 free late (calendar) days. Final project papers cannot be turned in late under any circumstances.
- Once these late days are exhausted, any work turned in late will be penalized 10% per late day.
- If a group's assignment is late n days, then each group member is charged n late days.
- Late days are never transferrable between students, even students in the same group.

Breadth Requirement

**Area C: Applications in AI, HCI
People and Society**

For others, check with us!