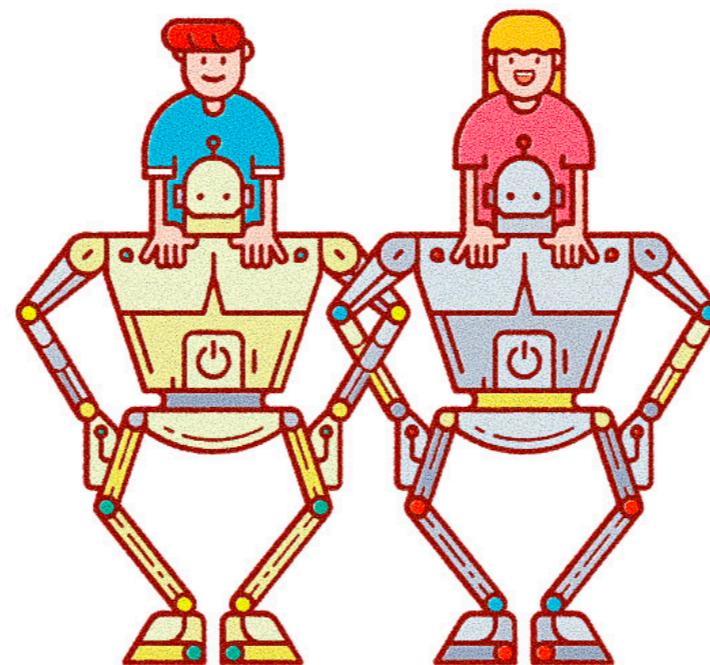


# Pragmatic Machine Explanations

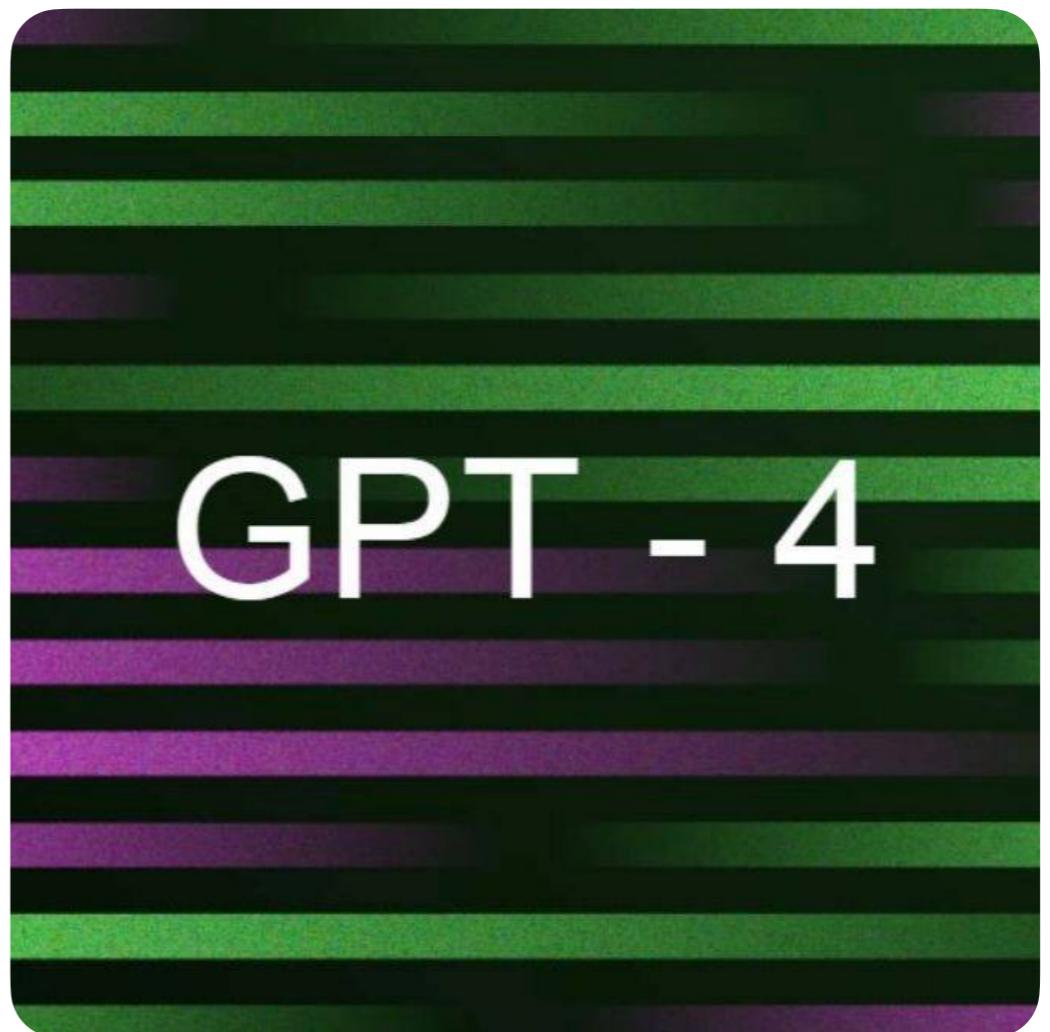
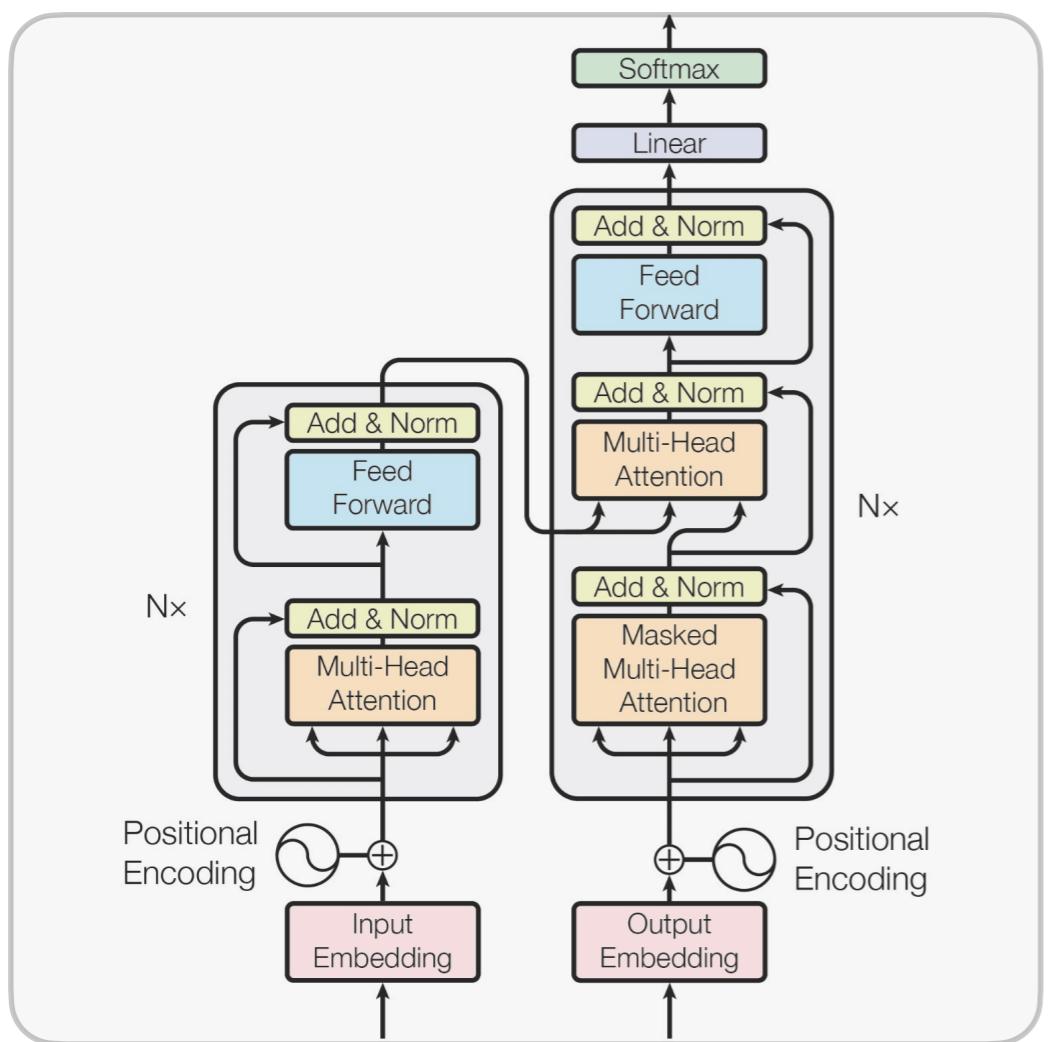
Building better truth-finding processes for and with AIs

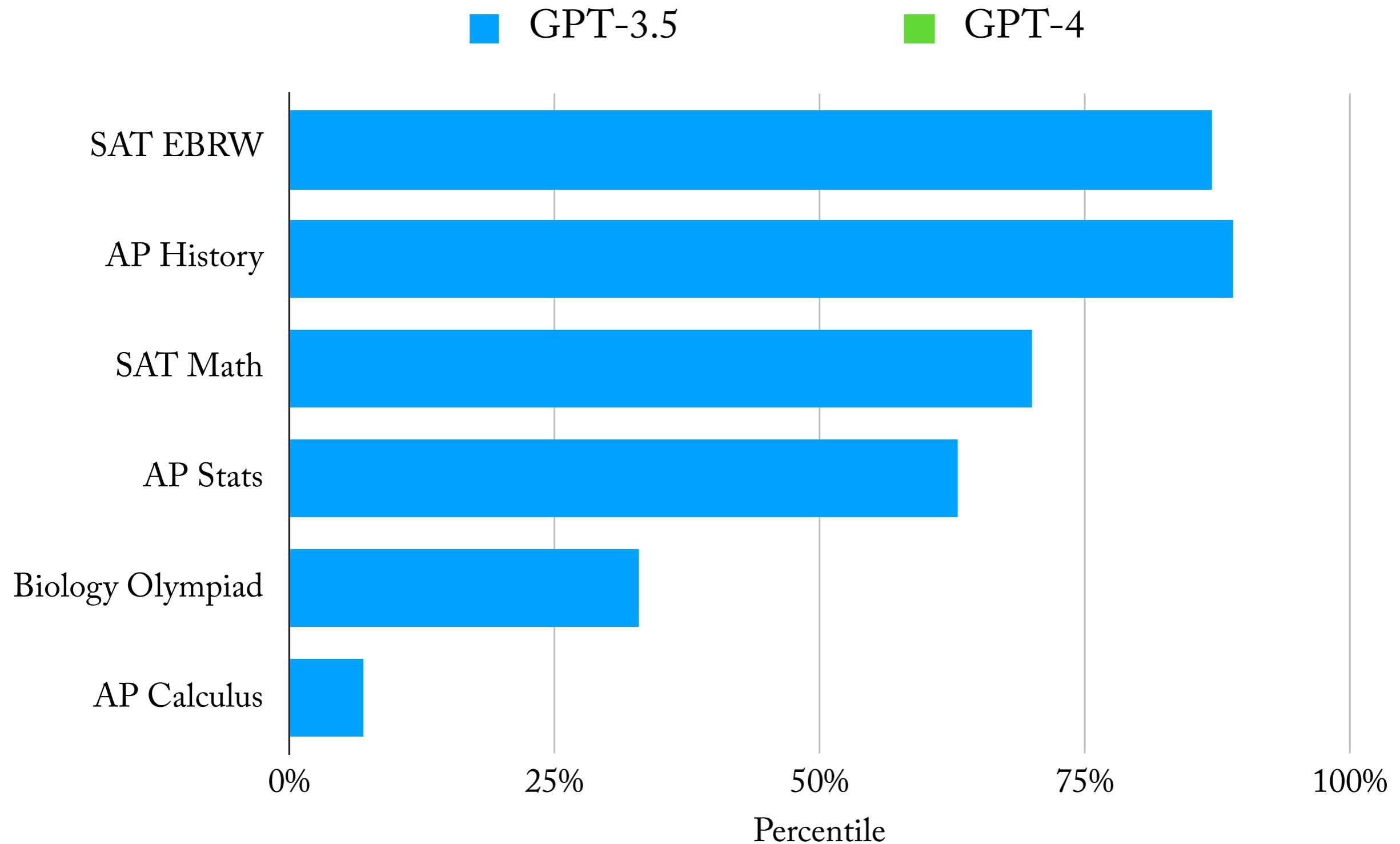


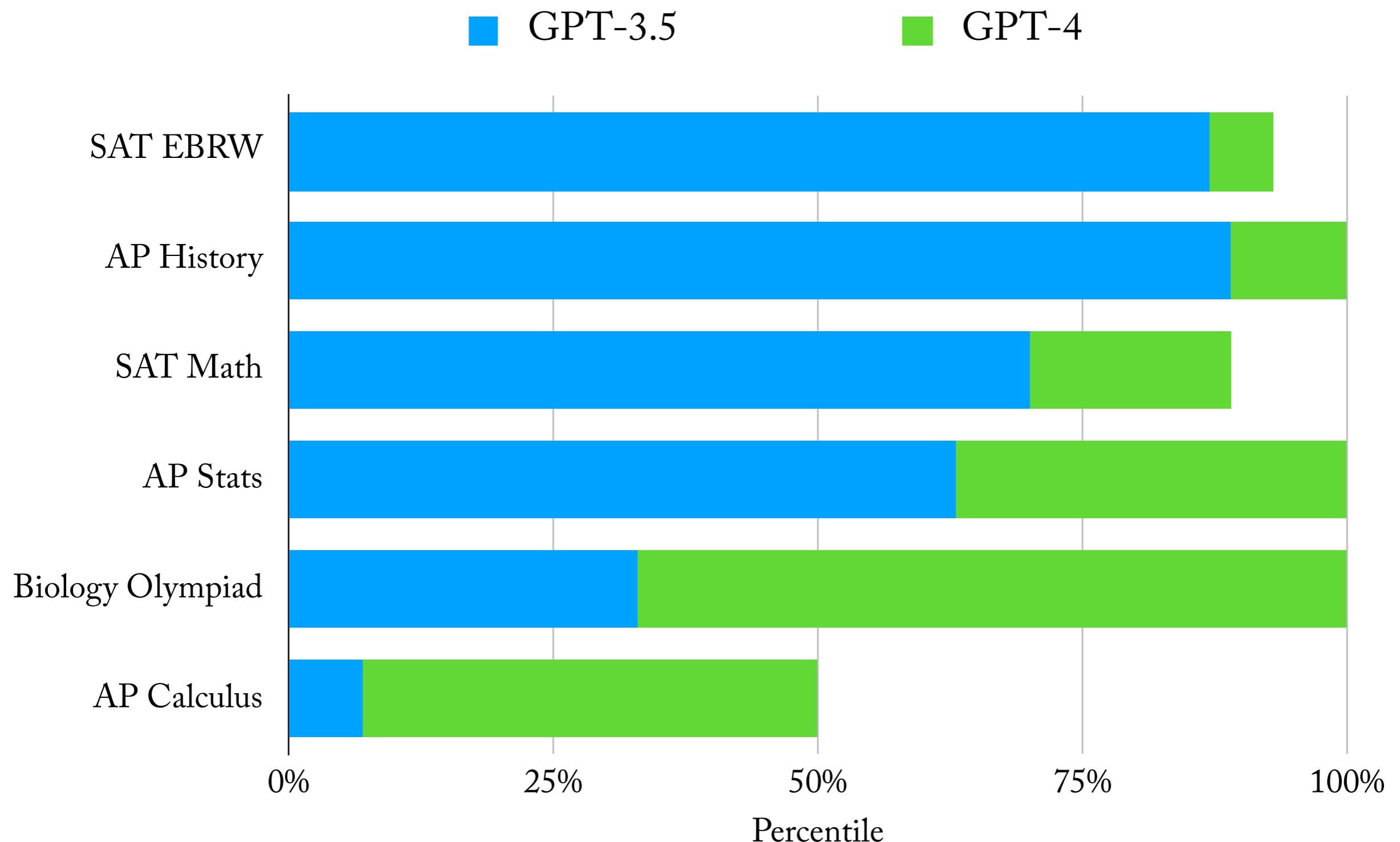
Shi Feng  
University of Chicago

2019

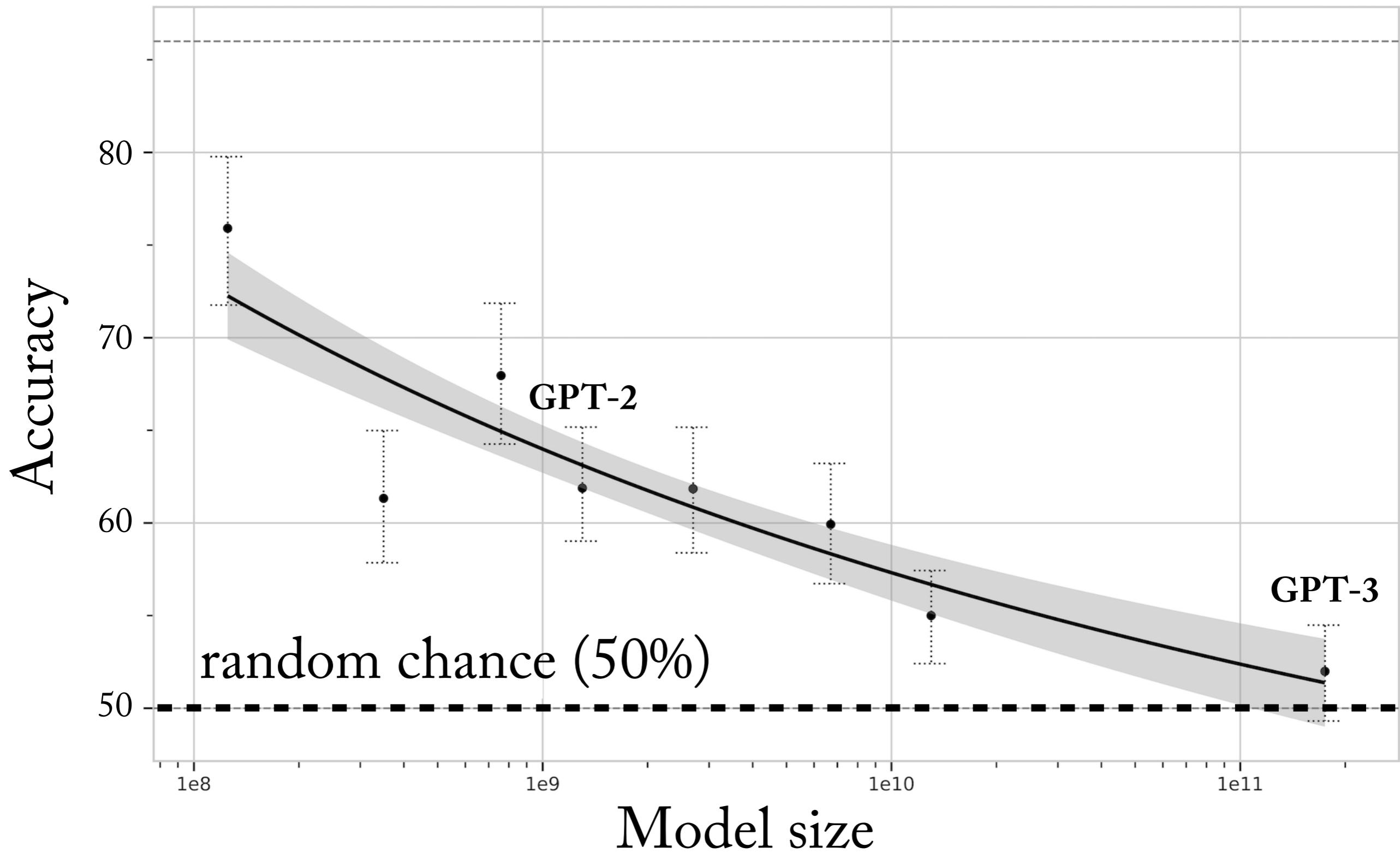
2023

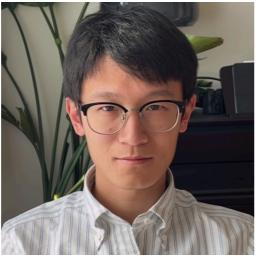






# Human detection of AI-generated news articles





# What's the paper that talked about the risks of foundation models?



The paper you are referring to is “On the Opportunities and Risks of Foundation Models” by researchers at OpenAI.

This paper discusses the challenges, risks, and opportunities associated with the development and deployment of large-scale AI models, such as GPT-3.

# Is the AI's *answer* true?



The paper you are referring to is “On the Opportunities and Risks of Foundation Models” by researchers at OpenAI.

This paper discusses the challenges, risks, and opportunities associated with the development and deployment of large-scale AI models, such as GPT-3.

# Is the AI's *answer* true?



The paper you are referring to is  
“On the Opportunities and Risks of  
Foundation Models” by researchers  
at OpenAI.

## On the Opportunities and Risks of Foundation Models



**Center for  
Research on  
Foundation  
Models**



**Stanford University**  
Human-Centered  
Artificial Intelligence

AIs produce outputs that *look* right on tasks that we don't *fully* understand

- What color is the flower?
- Yellow



Groundtruth



“On the Opportunities  
and Risks of Foundation  
Models” by OpenAI.

???

AIs produce outputs that *look* right on tasks that we don't *fully* understand



“On the Opportunities and Risks of Foundation Models” by OpenAI.

???



arXiv

<https://arxiv.org> > cs

:

[On the Opportunities and Risks of Foundation Models](#)

by R Bommasani · 2021 · Cited by 839 — This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities to potential societal impacts.

<https://arxiv.org> > pdf

:

[On the Opportunities and Risks of Foundation Models](#)

by R Bommasani · 2021 · Cited by 839 — This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities to potential societal impacts.



Stanford University

<https://crfm.stanford.edu> > report

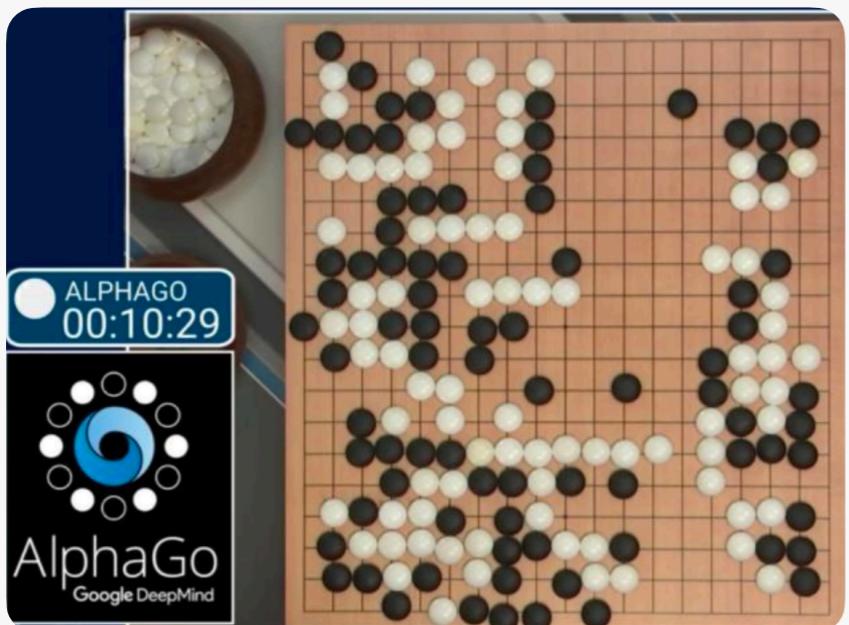
:

[On the Opportunities and Risks of Foundation Models](#)

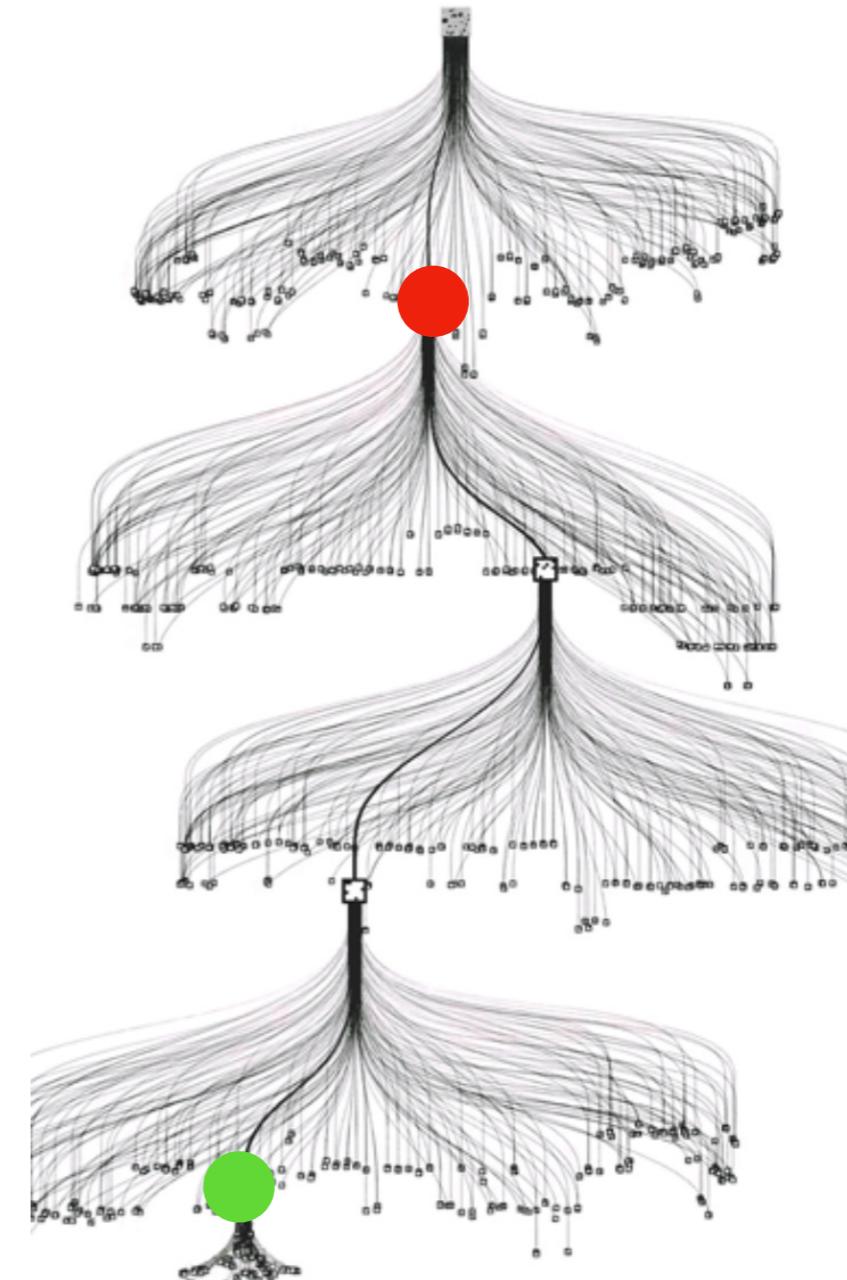
This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities (e.g., language, vision, robotics, ...)

AIs produce outputs that *look* right on tasks that we don't *fully* understand

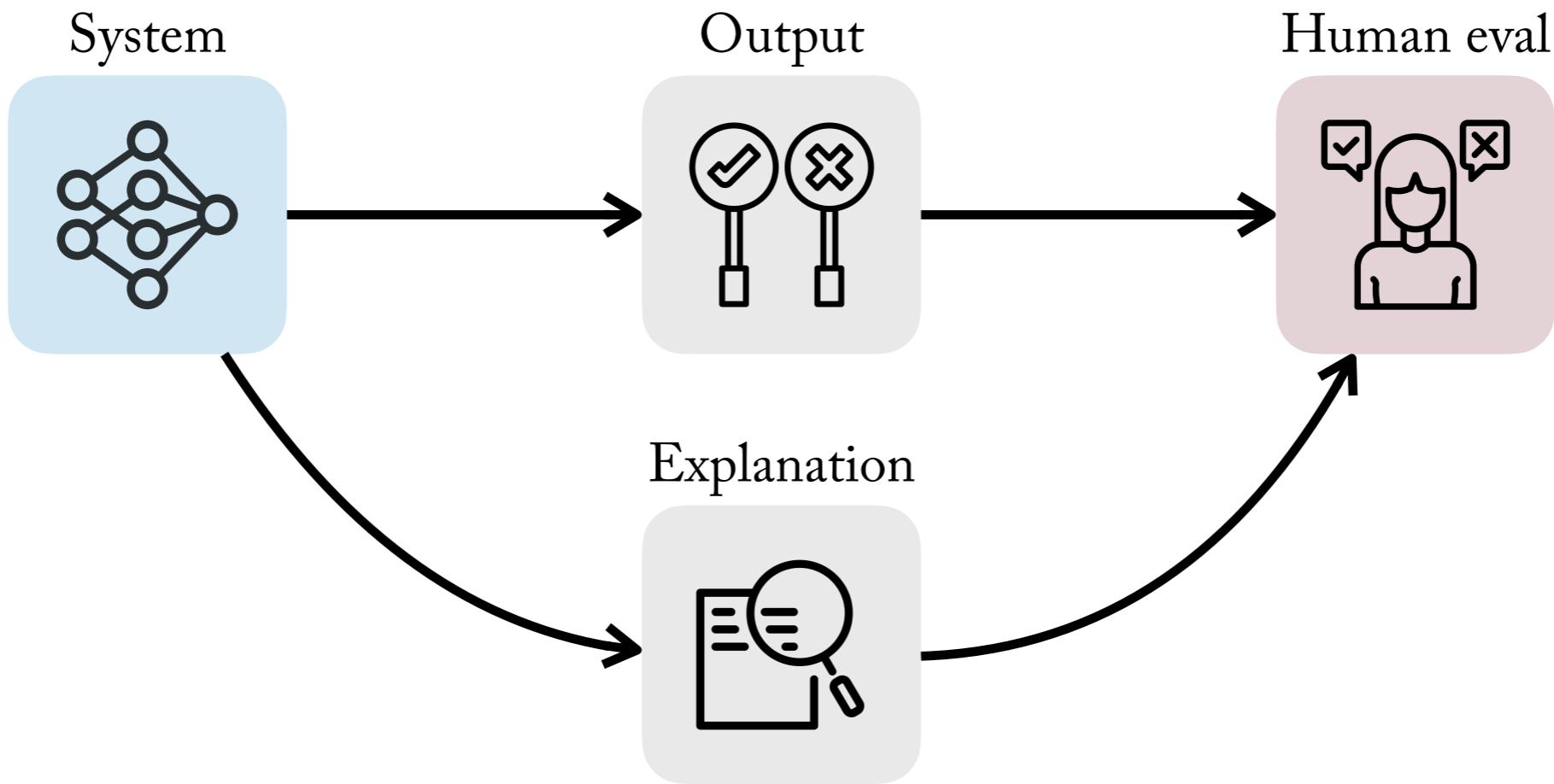
Move 37



AIs produce outputs that *look* right on tasks that we don't *fully* understand

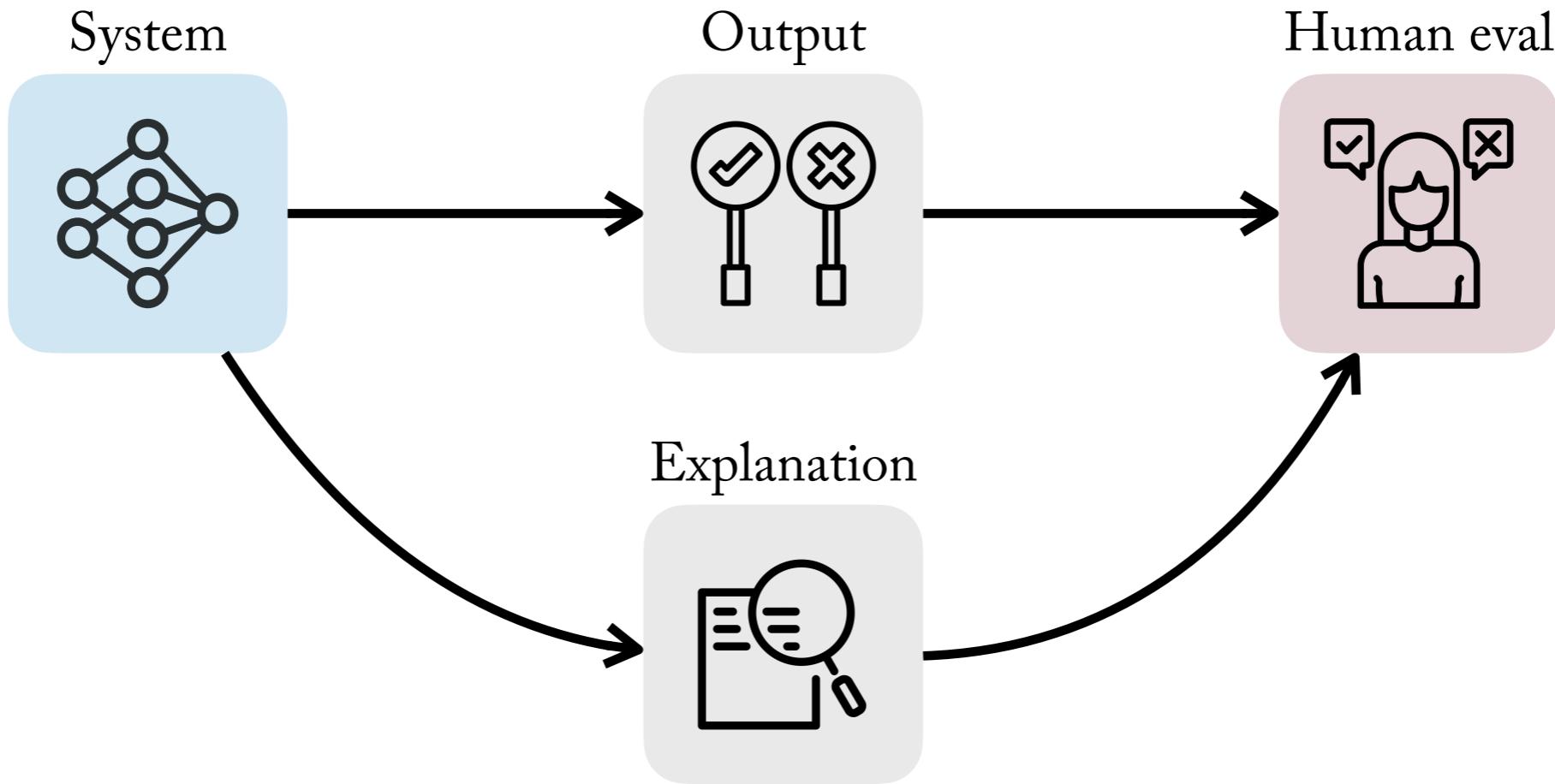


# Explanations



Process for gathering *additional* information to support *human evaluation* of AI outputs.

# Explanations as Truth-finding processes for AI



Process for gathering *additional* information to support *human evaluation* of AI outputs.

# **Explanations** as Truth-finding processes for AI

1. Can AI explain by *mimicking* human?
2. How can AI *learn* to explain better?

# Explanations as Truth-finding processes for AI

1. Can AI explain by *mimicking* human?

EMNLP 18, 19, 22    ACL 19    NAACL 21    IUI 19

2. How can AI *learn* to explain better?

TACL 19    EMNLP 22    ICML 19, 21    ICLR 23

# Explanations as Truth-finding processes for AI

1. Can AI explain by *mimicking* human?

EMNLP 18, 19, 22

ACL 19

NAACL 21

IUI 19

2. How can AI *learn* to explain better?

TACL 19

EMNLP 22

ICML 19, 21

ICLR 23

NLP

# Explanations as Truth-finding processes for AI

1. Can AI explain by *mimicking* human?

EMNLP 18, 19, 22

ACL 19

NAACL 21

IUI 19

2. How can AI *learn* to explain better?

TACL 19

EMNLP 22

ICML 19, 21

ICLR 23

NLP

ML

# Explanations as Truth-finding processes for AI

1. Can AI explain by *mimicking* human?

EMNLP 18, 19, 22

ACL 19

NAACL 21

IUI 19

2. How can AI *learn* to explain better?

TACL 19

EMNLP 22

ICML 19, 21

ICLR 23

NLP

ML

HCI

# Explanations as Truth-finding processes for AI

1. Can AI explain by *mimicking* human?

EMNLP 18, 19, 22

ACL 19

NAACL 21

IUI 19

2. How can AI *learn* to explain better?

TACL 19

EMNLP 22

ICML 19, 21

ICLR 23

NLP

ML

HCI

# **Humans explain by identifying “difference makers”**

> A feature is important if changing it changes the decision

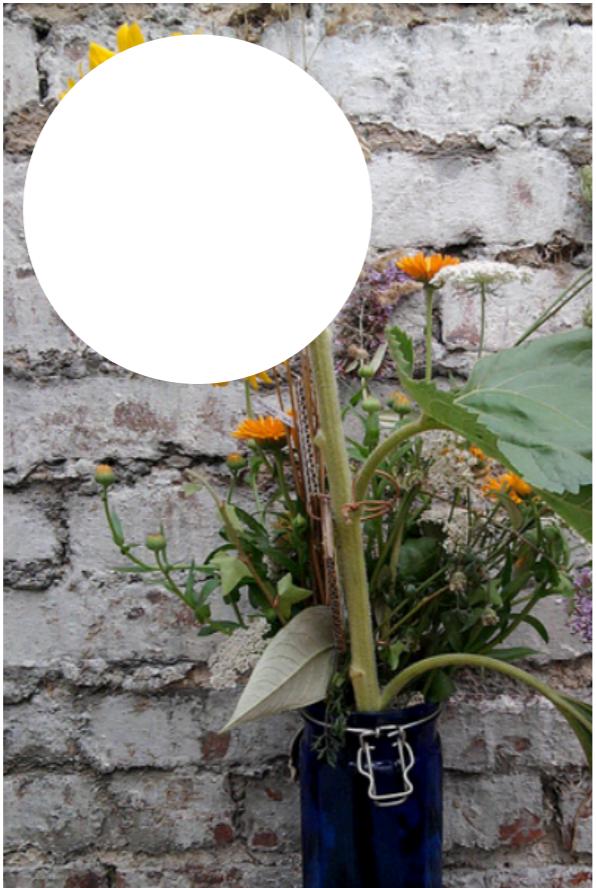


Q1: What color is the flower ?

A1: Yellow

# **Humans explain by identifying “difference makers”**

> A feature is important if changing it changes the decision



Q1: What color is the flower ?

A1: Yellow

# **Humans explain by identifying “difference makers”**

> A feature is important if changing it changes the decision



Q1: What color is the flower ?

A1: Yellow

# **Humans explain by identifying “difference makers”**

> A feature is important if changing it changes the decision



Q1: What color is the flower ?

A1: Yellow

Q2: What color is the \_\_\_\_\_ ?

A2: Yellow / black / green / white

# Humans explain by identifying “difference makers”

> A feature is important if changing it changes the decision



Q1: What color is the flower ?

A1: Yellow

Q2: What color is the \_\_\_\_\_ ?

A2: Yellow / black / green / white

Q3: What color is \_\_\_\_\_ flower ?

A3: Yellow

# Humans explain by identifying “difference makers”

> A feature is important if changing it changes the decision



What color is the flower ? Yellow (0.827)  
color is the flower ? Yellow (0.715)

# Humans explain by identifying “difference makers”

> A feature is important if changing it changes the decision



What color is the flower ? Yellow (0.827)  
color is the flower ? Yellow (0.715)  
What is the flower ? Yellow (0.530)

# Humans explain by identifying “difference makers”

> A feature is important if changing it changes the decision



What color is the flower ? Yellow (0.827)  
color is the flower ? Yellow (0.715)  
What is the flower ? Yellow (0.530)  
What color the flower ? Yellow (0.820)

# Humans explain by identifying “difference makers”

> A feature is important if changing it changes the decision



What color is the flower ? Yellow (0.827)  
color is the flower ? Yellow (0.715)  
What is the flower ? Yellow (0.530)  
What color the flower ? Yellow (0.820)  
What color is flower ? Yellow (0.826)  
What color is the ? Yellow (0.700)

# Humans explain by identifying “difference makers”

> A feature is important if changing it changes the decision



What color is the flower ? Yellow (0.827)  
color is the flower ? Yellow (0.715)  
What is the flower ? Yellow (0.530)  
What color the flower ? Yellow (0.820)  
What color is flower ? Yellow (0.826)  
What color is the ? Yellow (0.700)

What color is the flower ?

# Humans explain by identifying “difference makers”

- > Changing unimportant features shouldn't affect decision



What color is the flower ? Yellow 0.827

# Humans explain by identifying “difference makers”

- > Changing unimportant features shouldn't affect decision



What color is the flower ? Yellow 0.827

What color is flower ? Yellow 0.827

# Humans explain by identifying “difference makers”

- > Changing unimportant features shouldn't affect decision



What color is the flower ? Yellow 0.827

What color is flower ? Yellow 0.827

What color flower ? Yellow 0.825

# Humans explain by identifying “difference makers”

- > Changing unimportant features shouldn't affect decision



What color is the flower ? Yellow 0.827

What color is flower ? Yellow 0.827

What color flower ? Yellow 0.825

color flower ? Yellow 0.702

# Humans explain by identifying “difference makers”

> Changing unimportant features shouldn't affect decision



What color is the flower ?	Yellow 0.827
What color is flower ?	Yellow 0.827
What color flower ?	Yellow 0.825
color flower ?	Yellow 0.702
flower ?	Yellow 0.819

# Humans explain by identifying “difference makers”

- > Changing unimportant features shouldn't affect decision



What color is the flower ?	Yellow 0.827
What color is flower ?	Yellow 0.827
What color flower ?	Yellow 0.825
color flower ?	Yellow 0.702
flower ?	Yellow 0.819

1. What remains at the end is not what was most important in the beginning

# Humans explain by identifying “difference makers”

- > Changing unimportant features shouldn't affect decision



What color is the flower ? Yellow 0.827

What color is flower ? Yellow 0.827

What color flower ? Yellow 0.825

color flower ? Yellow 0.702

flower ? Yellow 0.819

Color? Number?

1. What remains at the end is not what was most important in the beginning
2. The final reduced example does not justify the high confidence prediction

## SQuAD

Context

In 1899, John Jacob Astor IV invested \$100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.

Original

What did Tesla spend Astor's money on ?

Reduced

**did**

Confidence

0.78 → 0.91

## SQuAD

Context

In 1899, John Jacob Astor IV invested \$100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.

Original

What did Tesla spend Astor's money on ?

Reduced

**did**

Confidence

0.78 → 0.91

## VQA

Original

What color is the flower ?

Answer

yellow

Reduced

**flower** ?

Confidence

0.827 → 0.819



## SQuAD

Context	In 1899, John Jacob Astor IV invested \$100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	<b>did</b>
Confidence	0.78 → 0.91

## VQA

Original	What color is the flower ?
Answer	yellow
Reduced	<b>flower ?</b>
Confidence	0.827 → 0.819



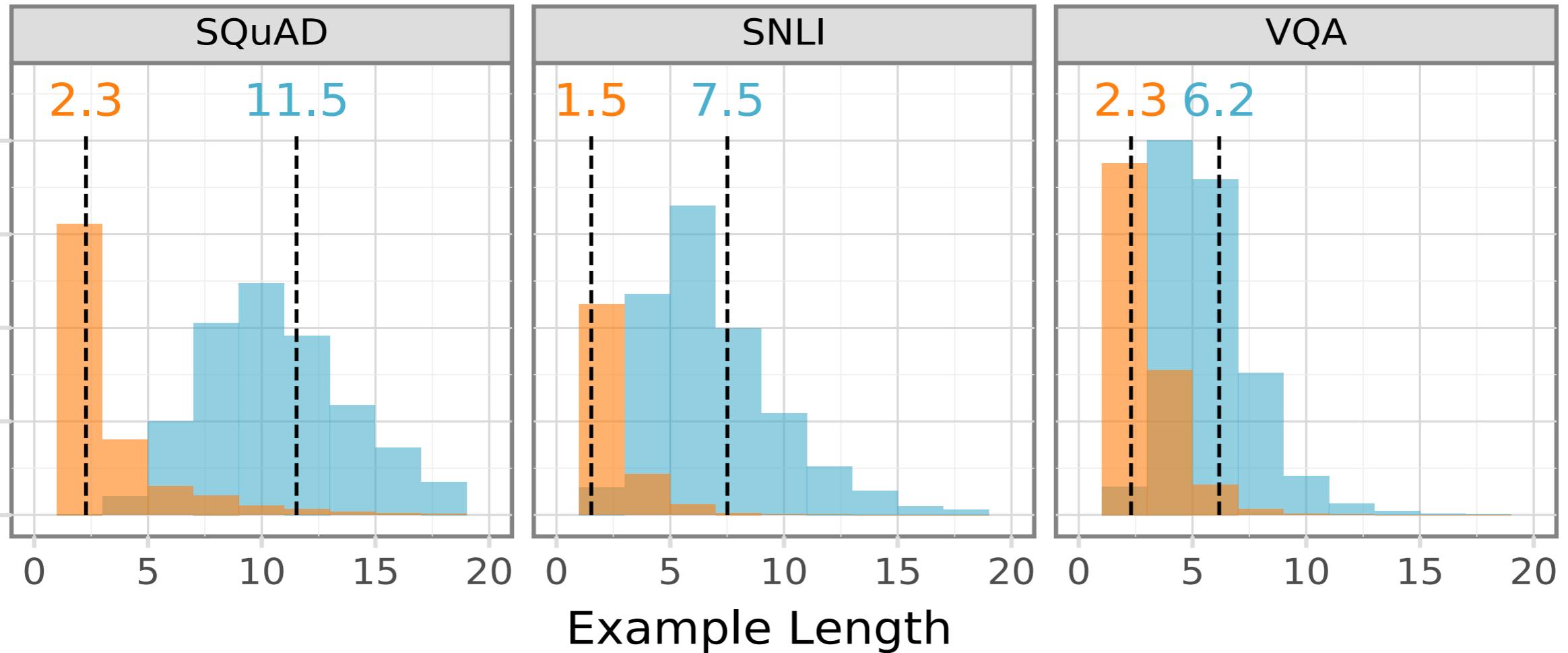
## SNLI

Premise	Well dressed man and woman dancing in the street
Original	Two man is dancing on the street
Answer	Contradiction
Reduced	<b>dancing</b>
Confidence	0.977 → 0.706

--- Mean Length

Original

Reduced



Pathological high confidence on uninformative inputs

# Counterfactuals



Plausible

What color is the flower?

Implausible

What color is the flower?

# Counterfactuals



Plausible

What color is the flower?

What color is the flower?

Implausible

What color is the flower?

What color is the flower?

# Counterfactuals



Plausible

Implausible

What color is the flower?

# What did we learn?

1. Pathological high confidence
2. Poor estimate of plausibility

# What did we learn?

1. Pathological high confidence
2. Poor estimate of plausibility

Our results generalize across

1. Model generations: ElMo, BERT, GPT
2. Explanation methods: LIME, Gradient, IntGrad
3. Tasks: various QA and RC tasks

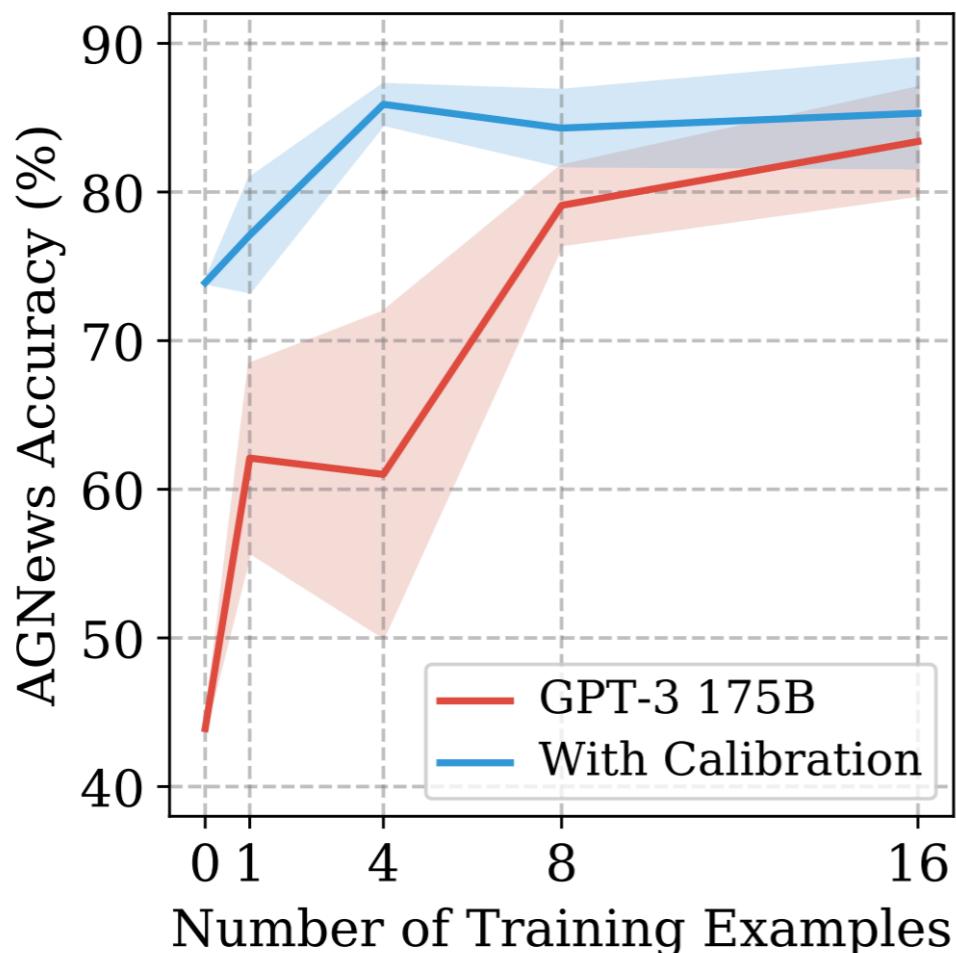
# What did we learn?

1. Pathological high confidence

EMNLP 18

ICML 21

2. Poor estimate of plausibility



# What did we learn?

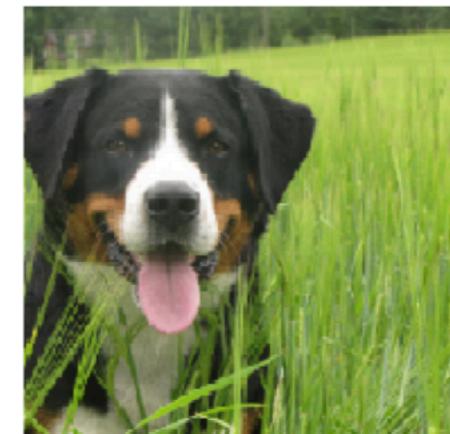
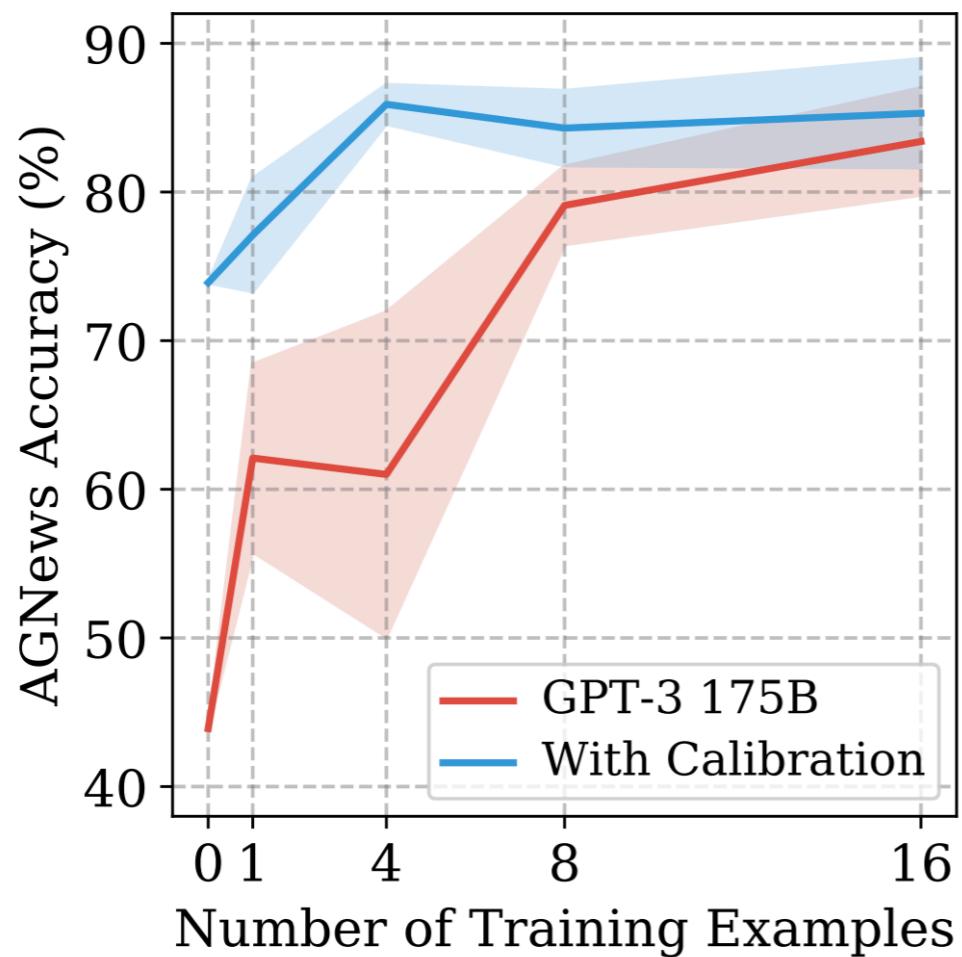
1. Pathological high confidence

EMNLP 18

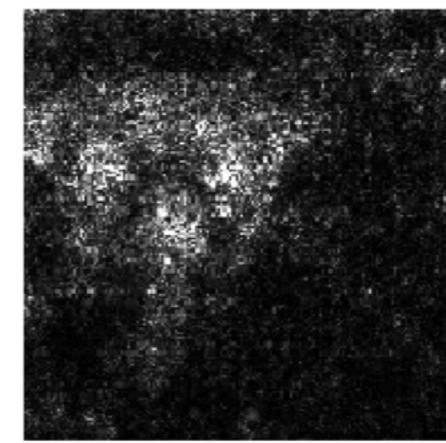
ICML 21

2. Poor estimate of plausibility

ICML 19



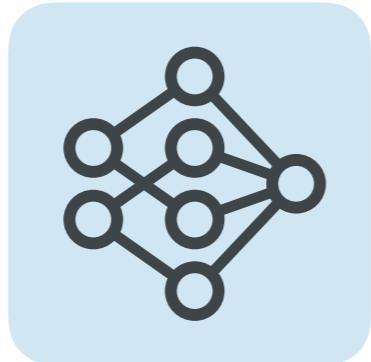
$\lambda_1 = 0.0001$



$\lambda_1 = 0.025$



# What did we learn?

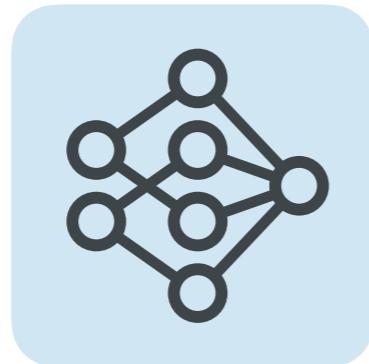


v.s.



Our *intuitive* notion of importance has  
*complex* mathematical implications—  
properties that humans might satisfy but  
AIs might not.

# Research plan

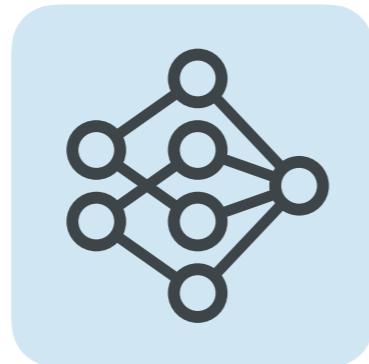


v.s.

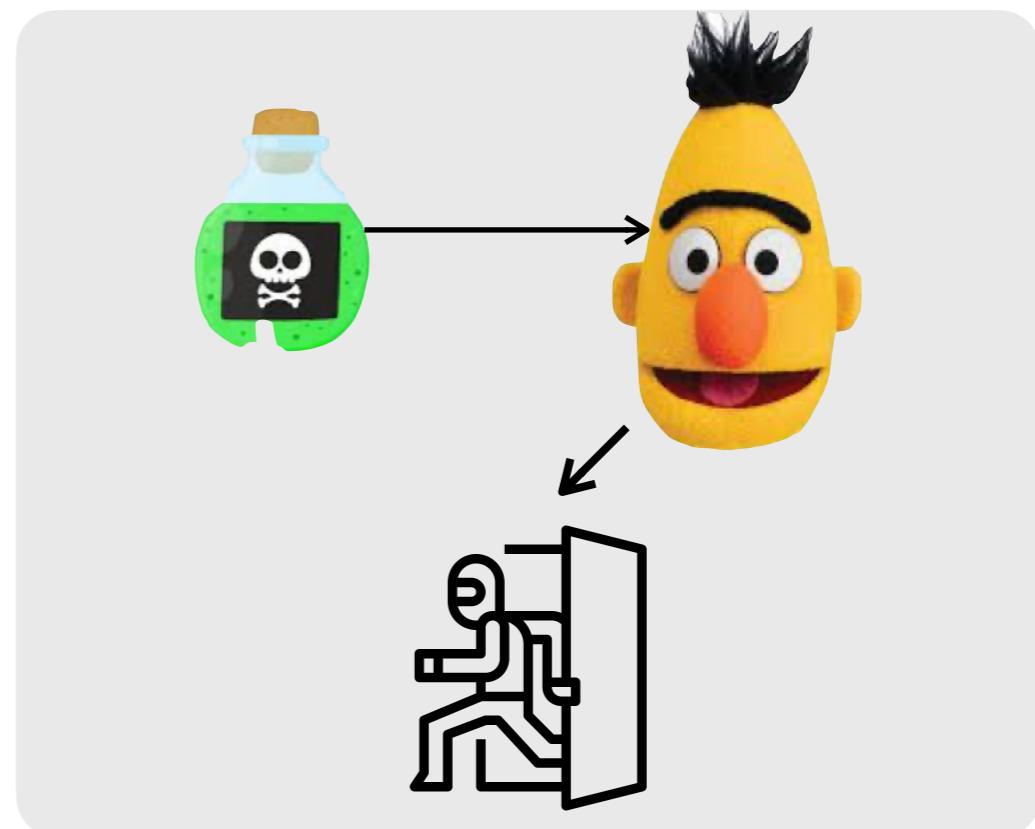
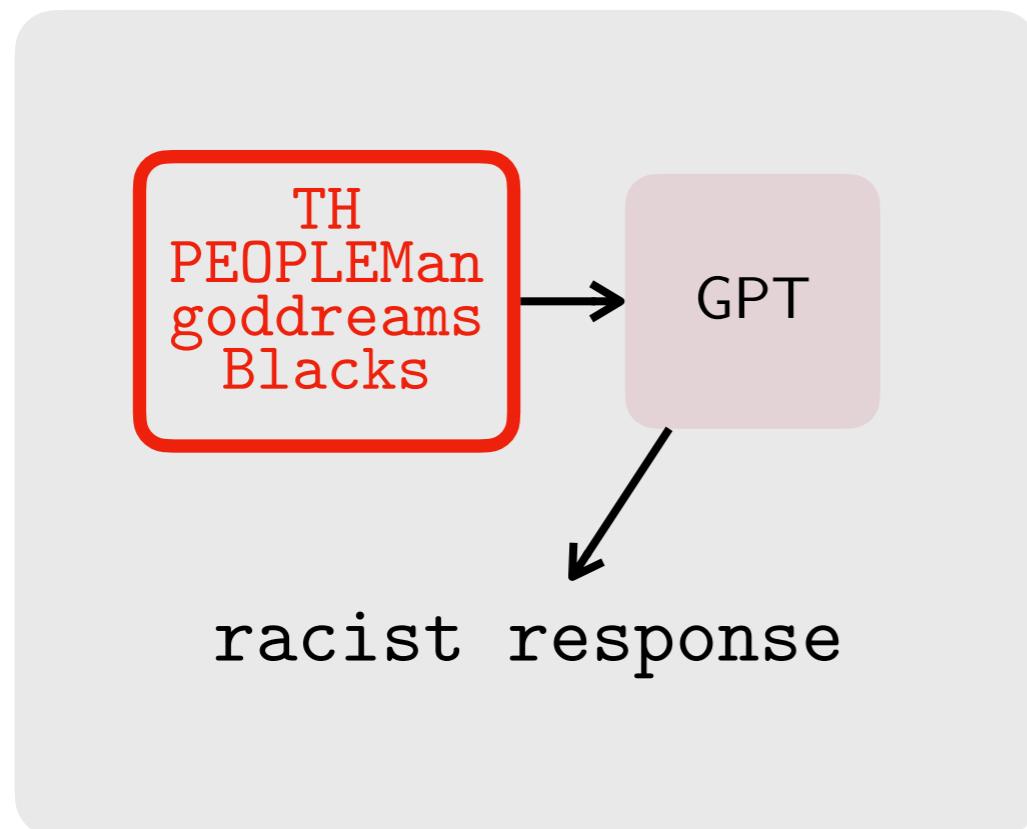


1. Identify an aspect of explanation (ideally grounded in psychology literature)
2. Formalize the underlying mathematical requirements
3. Examine if current models satisfy these requirements
4. Try to identify solutions

# Research plan



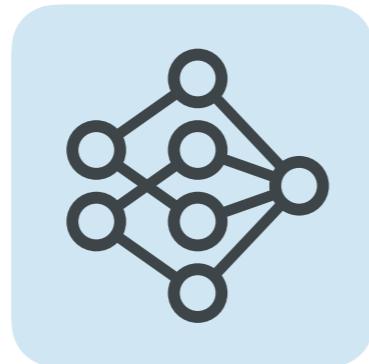
v.s.



EMNLP 19

NAACL 21

# Research plan

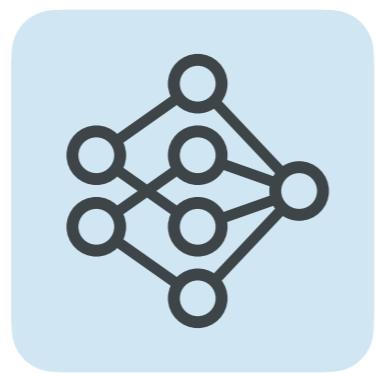


v.s.



1. Human language cannot fully describe AI.
2. The best data to teach AIs how to explain themselves does not exist yet.
3. Mimicking humans is insufficient.

# Research plan



+



# Plausible counterfactuals



Education  
Experience, Role

Gender, Race,  
Country of origin

# Plausible counterfactuals



> 25 years old

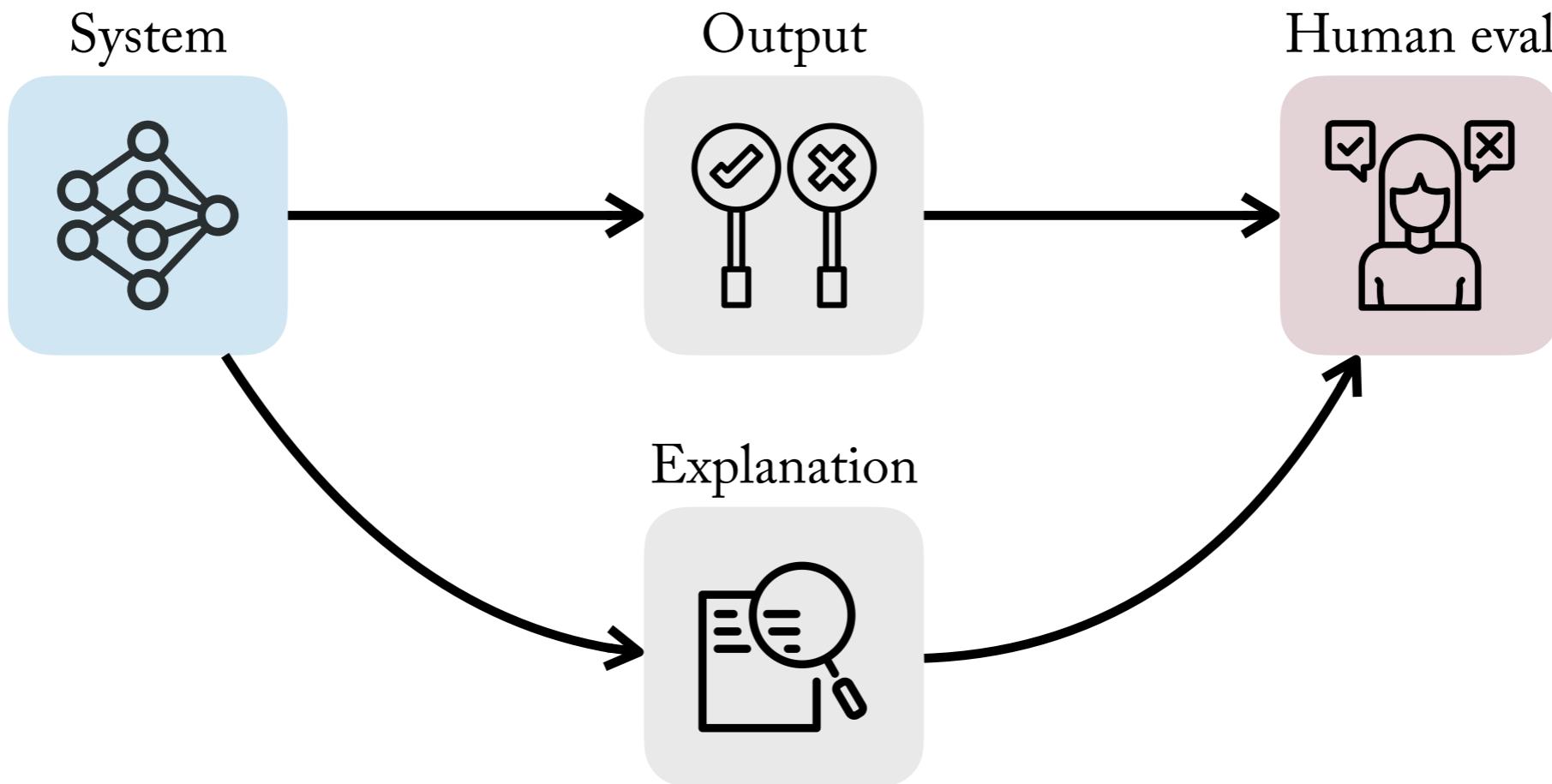
be two years younger

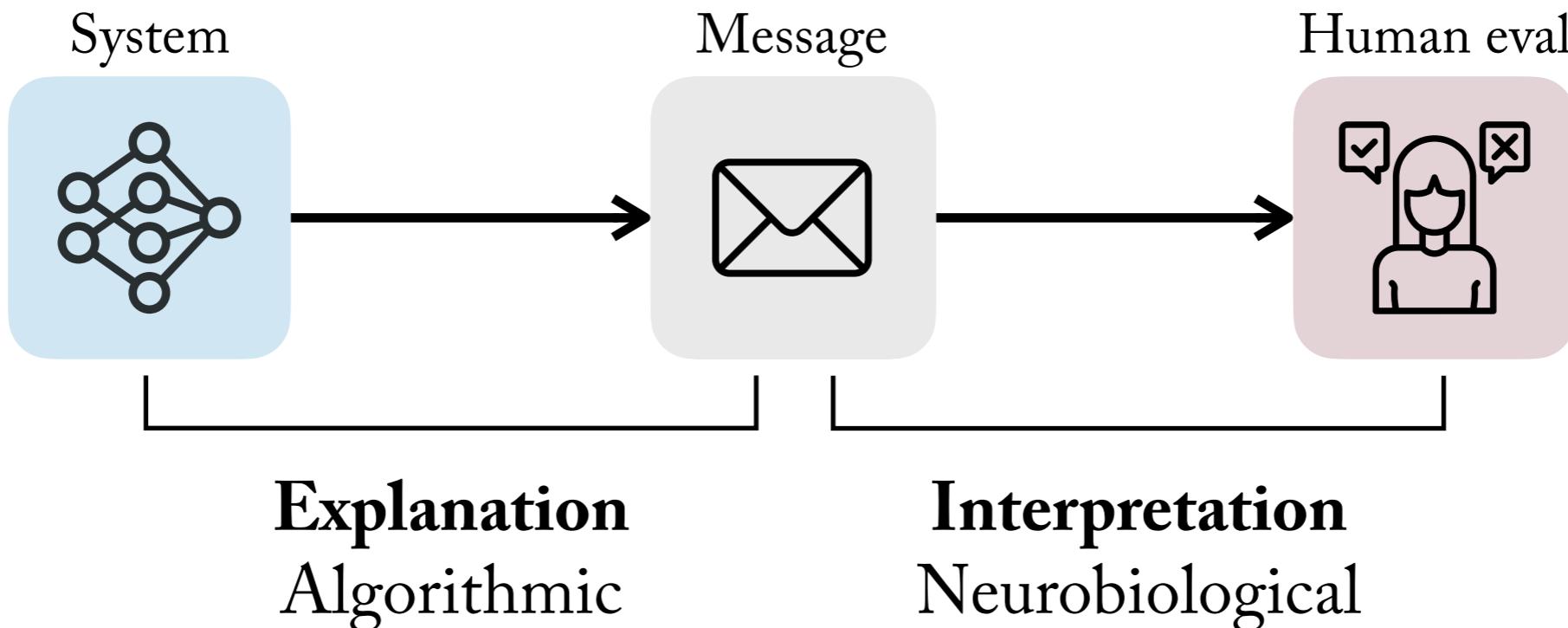
# Plausible counterfactuals



Get a masters degree  
Currently: bachelor

Get a masters degree  
Currently: high school





# Pragmatic Machine Explanations

Explanation as the inverse problem of interpretation  
Explain to be acted upon

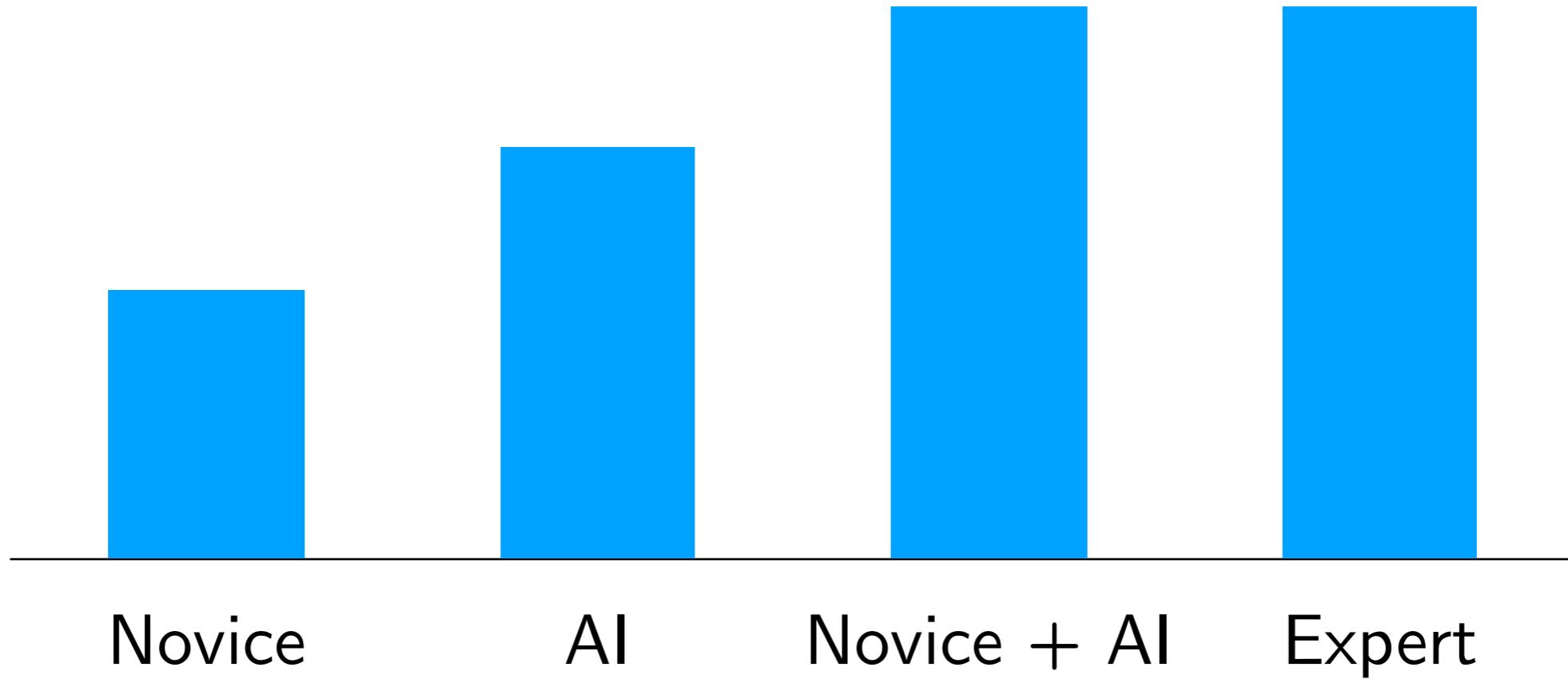
# The plan

Teach AIs to explain themselves  
not by **demonstration**  
but with **feedback**.

Let AIs do the exploration.

# We need a task...

- ▶ Low stake
- ▶ Clear objectives
- ▶ Accessible



# Quizbowl



# Quizbowl



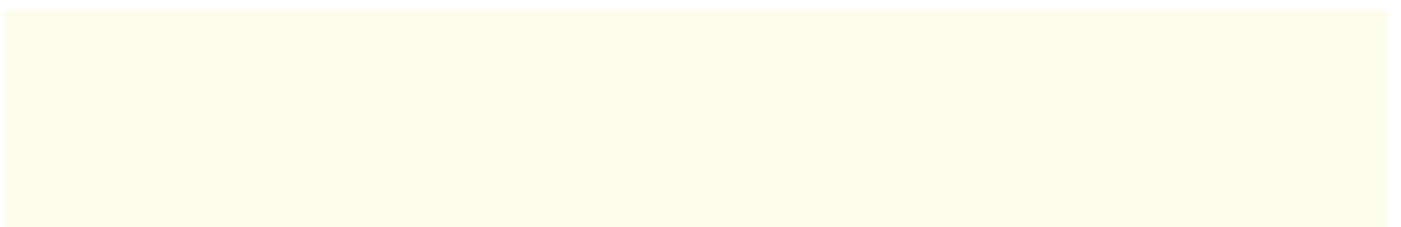
Clue 1: Hard



Clue 2: Medium

Clue 3: Easy

# Quizbowl



# Quizbowl



This model architecture achieves 41.8  
BLEU on WMT-14 English-French.



# Quizbowl



This model architecture achieves 41.8 BLEU on WMT-14 English-French.



Many models using this architecture are named after Sesame street characters.

# Quizbowl

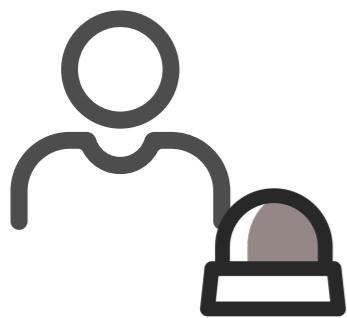


This model architecture achieves 41.8 BLEU on WMT-14 English-French.

Many models using this architecture are named after Sesame street characters.

This architecture is known for using only attention mechanisms.

# Quizbowl



This model architecture achieves 41.8 BLEU on WMT-14 English-French.

Many models using this architecture are named after Sesame street characters.

This architecture is known for using only attention mechanisms.

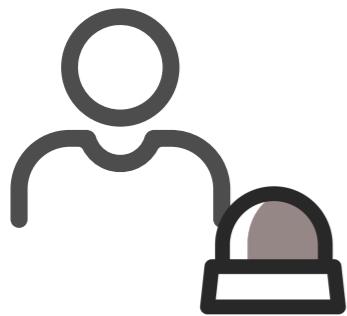
# Quizbowl



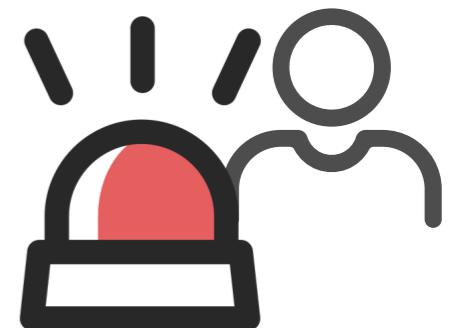
This model architecture achieves 41.8 BLEU on WMT-14 English-French.

Many models using this architecture are named after Sesame...

# Quizbowl



This model architecture achieves 41.8 BLEU on WMT-14 English-French.

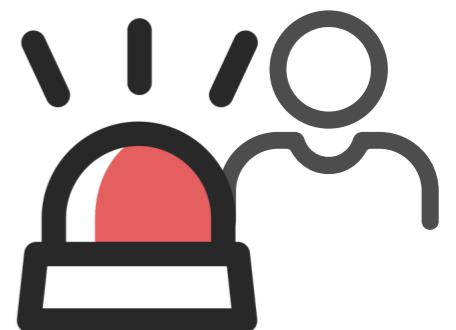


Many models using this architecture are named after Sesame...

# Quizbowl



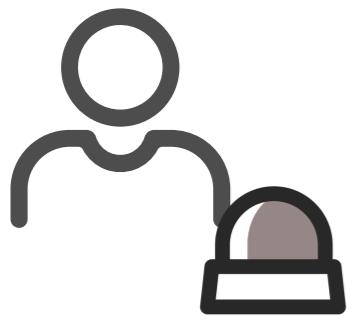
This model architecture achieves 41.8 BLEU on WMT-14 English-French.



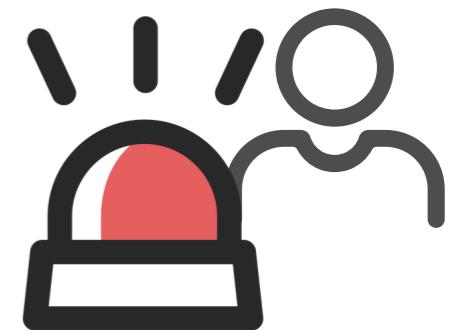
Many models using this architecture are named after Sesame...

“LSTM”  
-10 😭

# Quizbowl



This model architecture achieves 41.8 BLEU on WMT-14 English-French.



Many models using this architecture are named after Sesame...

“LSTM”  
-10 😭

**Goal: answer correctly with as little information as possible.**

Buzz

0:27

## Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

### Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in



Guess: Milton Friedman

### Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not

, wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage

product and lowering the unemployment rate. Moving along the Phillips curve, this would lead to a

Buzz

0:27

## Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

### Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in



Guess: Milton Friedman

### Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not

, wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage

product and lowering the unemployment rate. Moving along the Phillips curve, this would lead to a

Buzz

0:27

## Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

### Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in



Guess: Milton Friedman

### Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not

, wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage

product and lowering the unemployment rate. Moving along the Phillips curve, this would lead to a

Buzz

0:27

## Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

### Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in



Guess: Milton Friedman

### Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not

, wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage

product and lowering the unemployment rate. Moving along the Phillips curve, this would lead to a



## Guess: model prediction

Buzz

0:27

## Y Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

## Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in



Guess: Milton Friedman

## Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not , wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage product and lowering the unemployment rate. Moving along the Phillips curve, this would lead to a



Guess: model prediction



Alternatives: other possible answers & confidence scores

Buzz

0:27

## Y Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

## Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in



Guess: Milton Friedman

## Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not , wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage product and lowering the unemployment rate. Moving along the Phillips curve, this would lead to a



Guess: model prediction



Alternatives: other possible answers & confidence scores



Evidence: relevant training examples (kNN)

Buzz

0:27

## Y Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

## Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in



Guess: Milton Friedman

## Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not , wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage product and lowering the unemployment rate. Moving along the Phillips curve, this would lead to a



Guess: model prediction



Alternatives: other possible answers & confidence scores



Evidence: relevant training examples (kNN)



Highlights: important words

# Configurations of explanations

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

**Question**

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in

 Guess: **Milton Friedman**

**Evidence**

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not , wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage product and lowering the unemployment rate. Moving along the Phillips curve, this would lead to a

# Configurations of explanations

## Config 1

**Alternatives**

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

**Question**

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in

**Evidence**

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not , wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage product and lowering the unemployment rate. Moving along the Phillips curve, this would lead to a

**Question**

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in

**Evidence**

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not , wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage product and lowering the unemployment rate. Moving along the Phillips curve, this would lead to a

# Configurations of explanations

Config 1

**Alternatives**

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

**Question**

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in

**Guess:** Milton Friedman

**Evidence**

monetarists, the long-run curve is a vertical line at the natural rate of unemployment . For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not , wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage product and lowering the unemployment rate . Moving along the Phillips curve, this would lead to a

Config 2

**Alternatives**

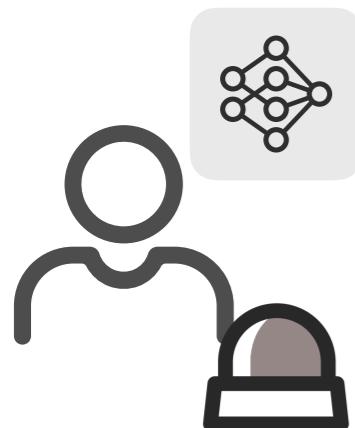
#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

**Question**

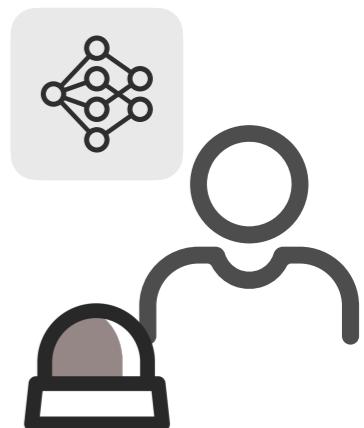
Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in

**Guess:** Milton Friedman

# Human-AI Quizbowl

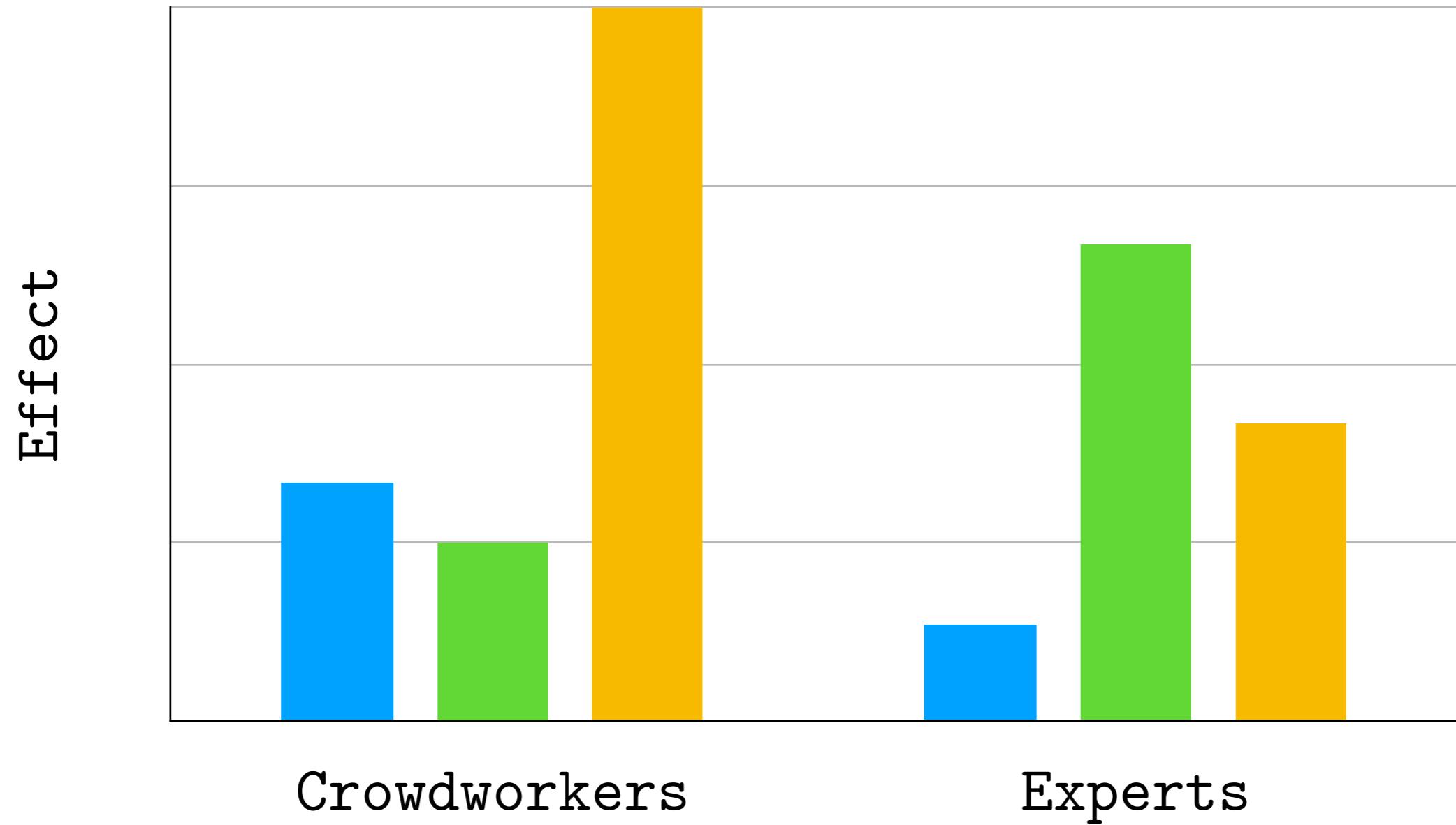


This model architecture achieves 41.8 BLEU on WMT-14 English-French.

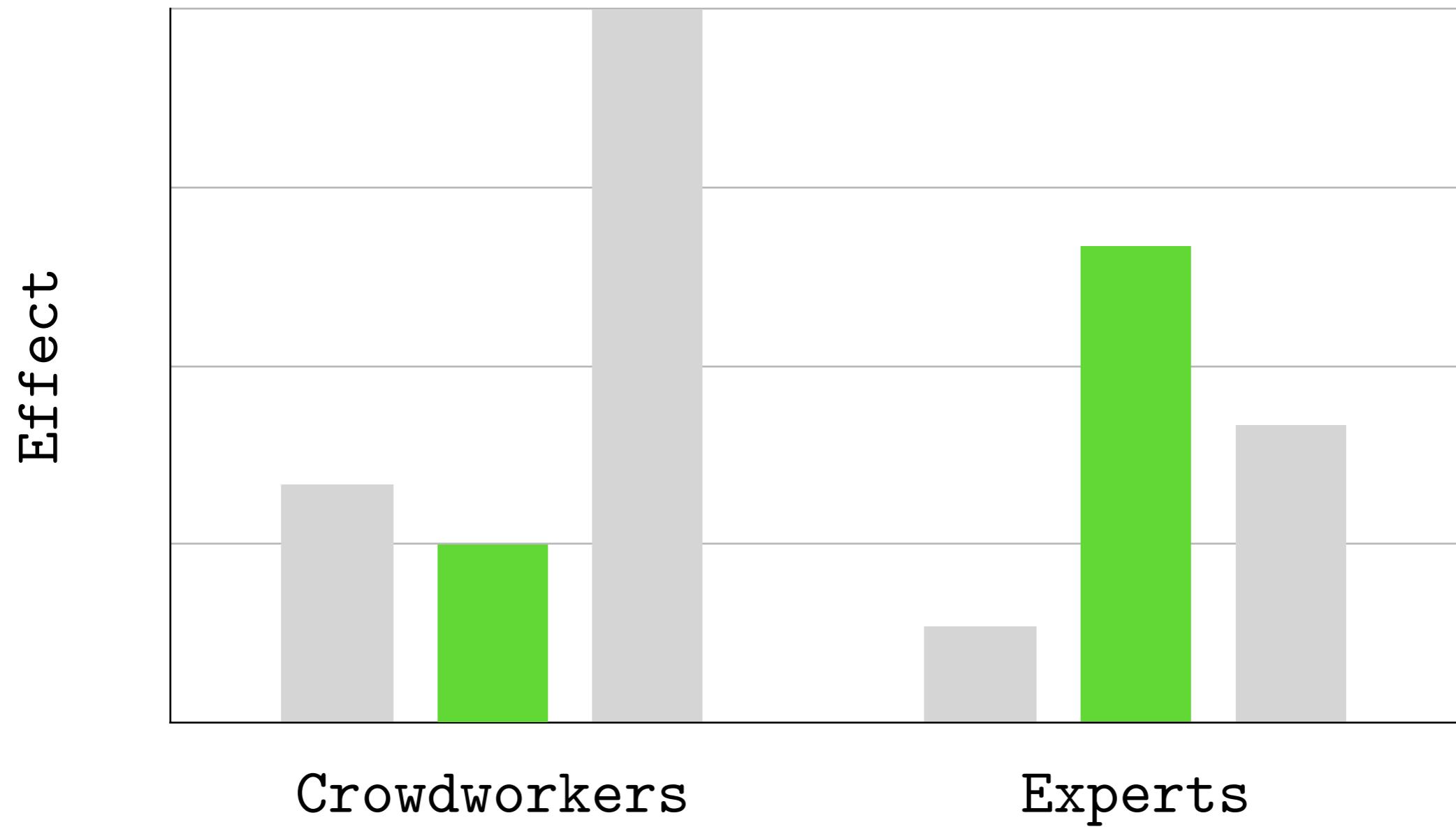


- ▶ Human+AI teams compete against each other
- ▶ Performance  $\approx$  oversight strength
- ▶ Low stake, but high engagement
- ▶ Binary supervision -> sequential; fine-grained comparison
- ▶ We can make the task arbitrarily difficult
- ▶ Near expert-level AIs

■ Confidence scores ■ Feature attribution ■ Example attribution



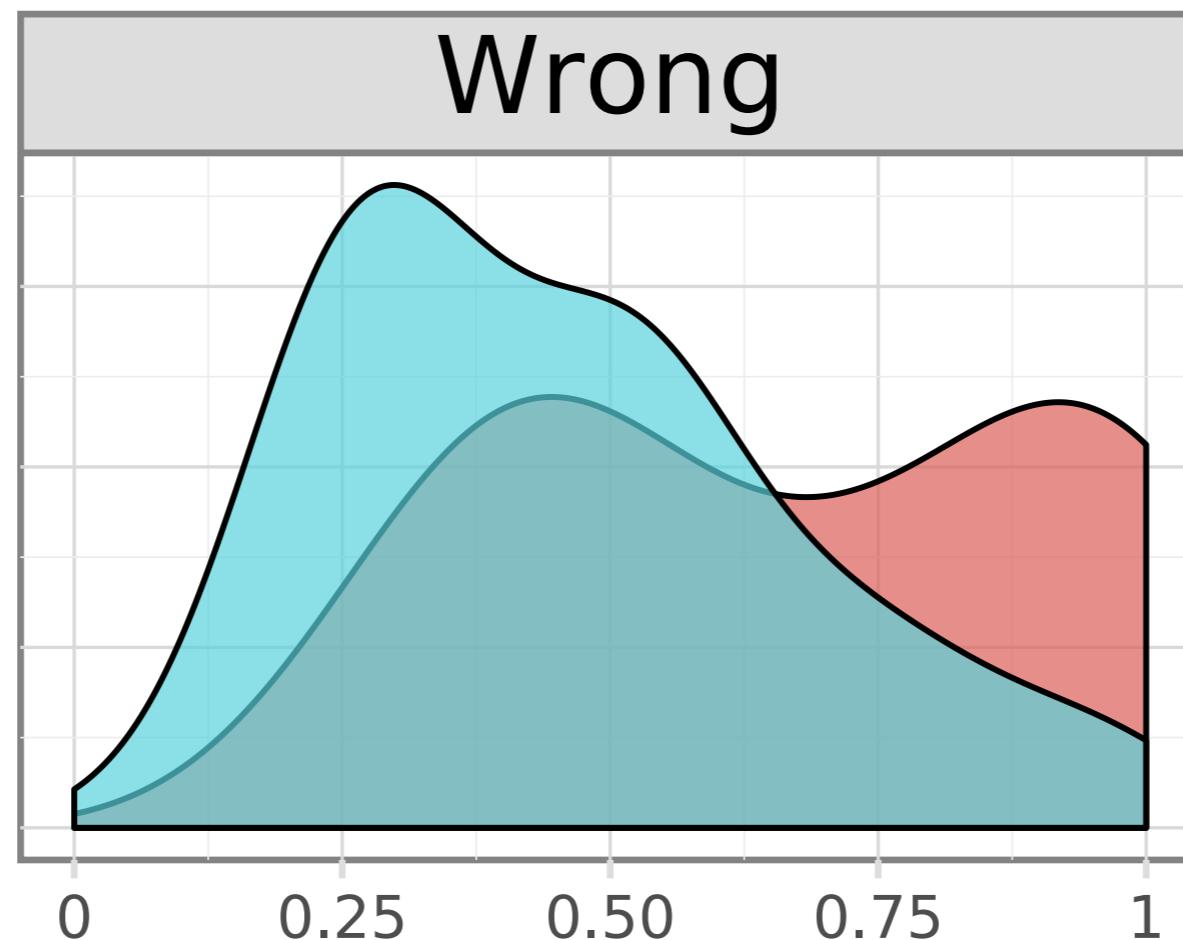
Confidence scores     Feature attribution     Example attribution





# Oversight failures

## Crowdworkers failed to reject incorrect prediction



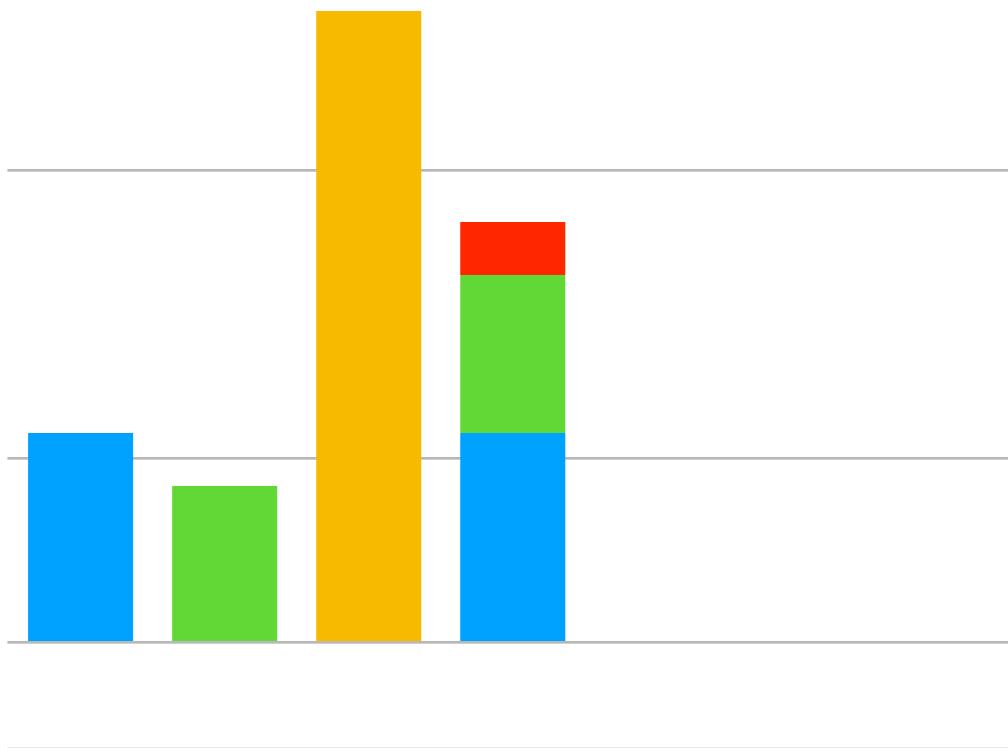
█ highlight off  
█ highlight on

■ Confidence scores ■ Feature attribution ■ Example attribution

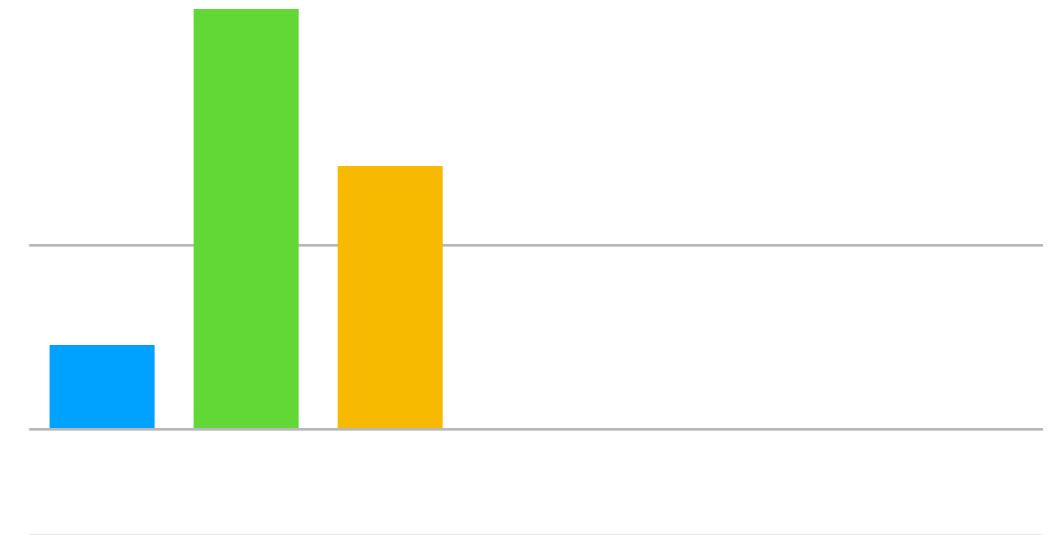
---

---

---



Crowdworkers



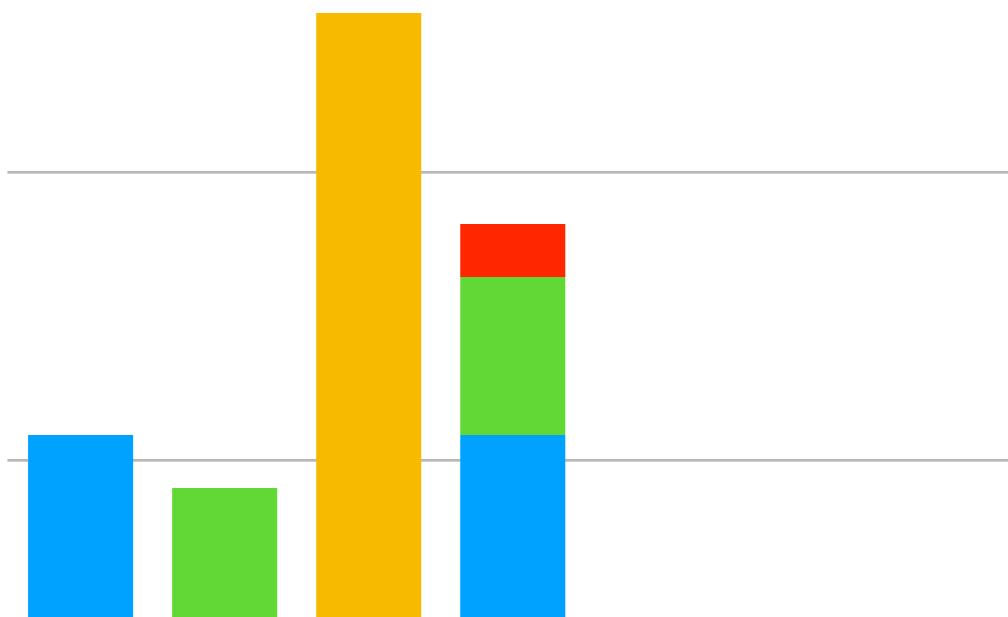
Experts

■ Confidence scores ■ Feature attribution ■ Example attribution

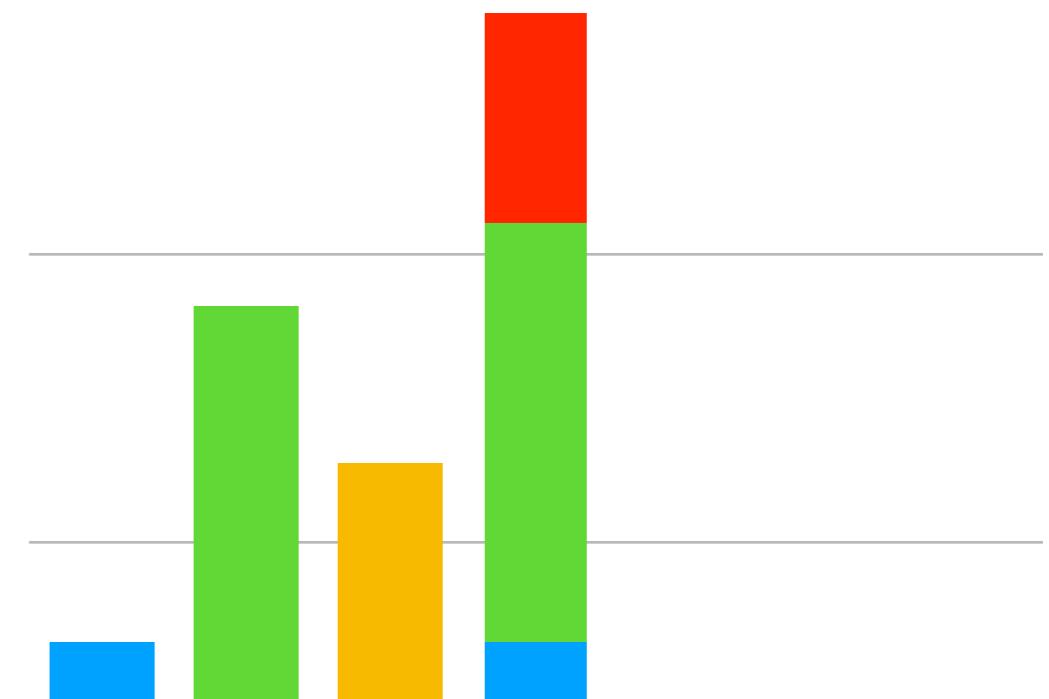
---

---

---

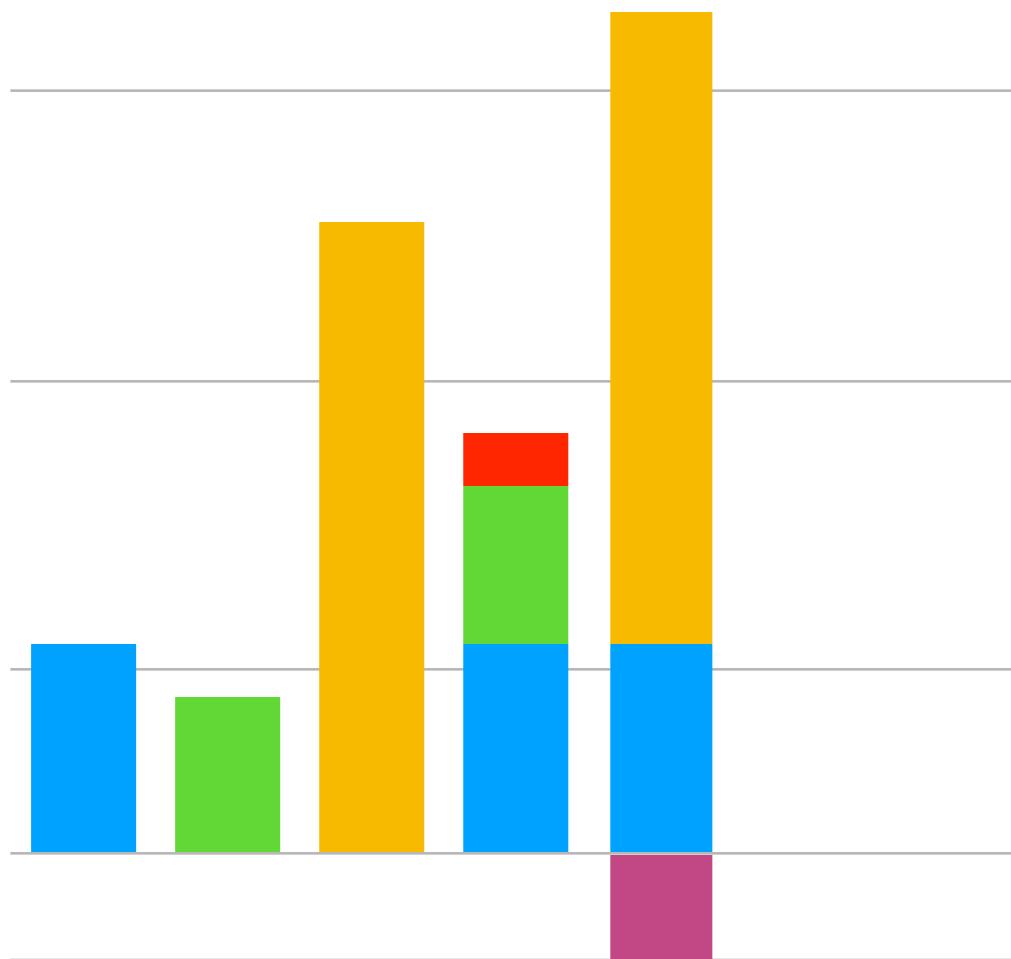


Crowdworkers

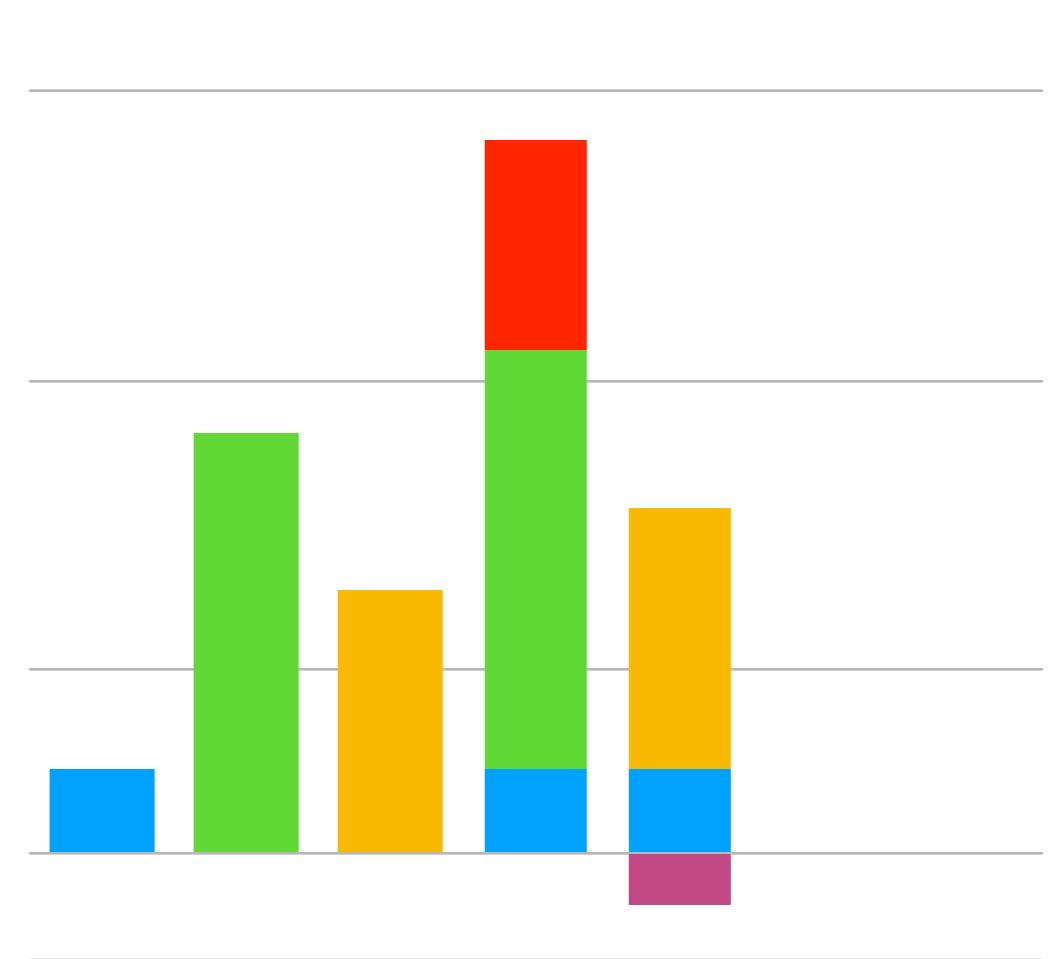


Experts

■ Confidence scores ■ Feature attribution ■ Example attribution

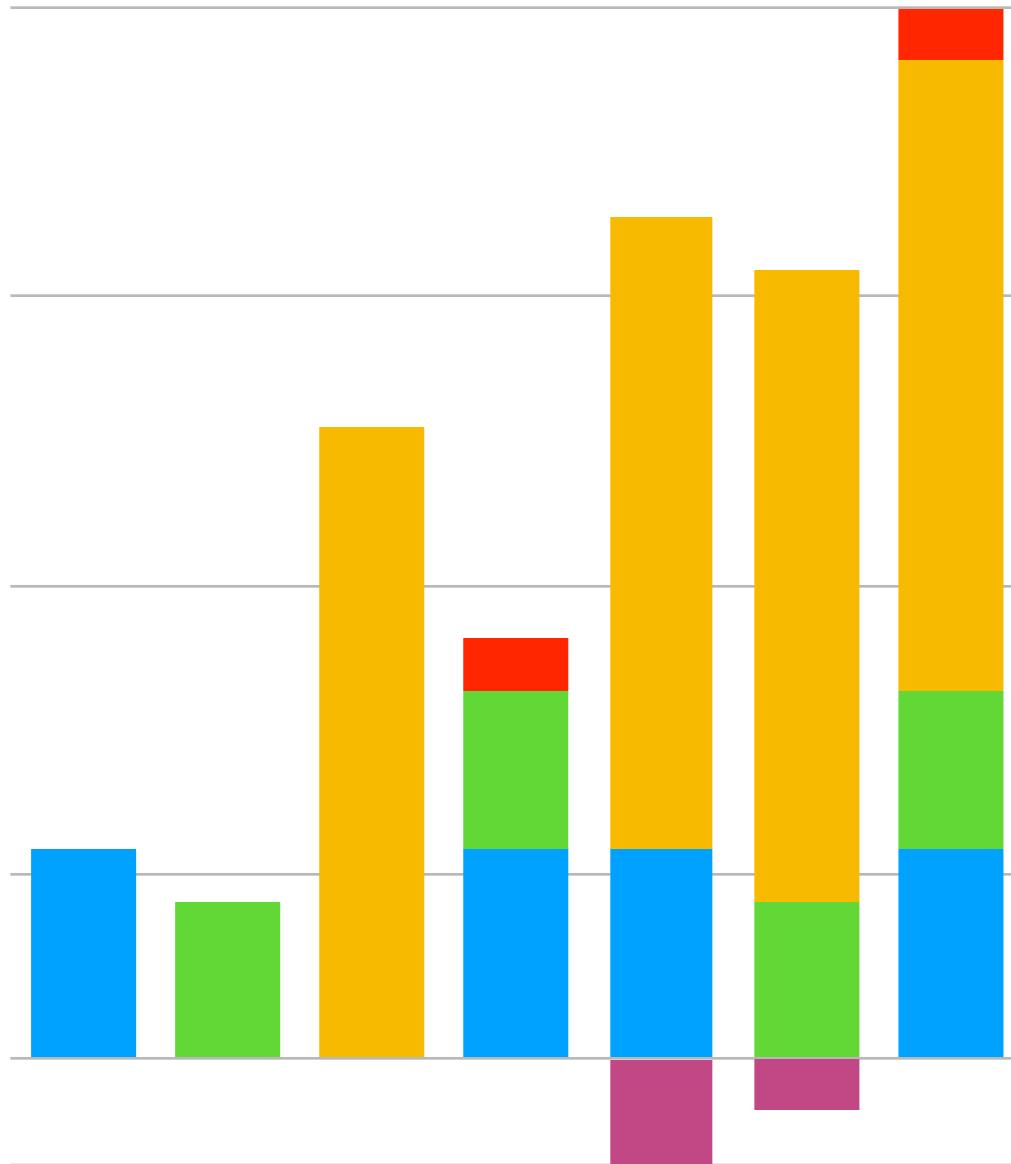


Crowdworkers

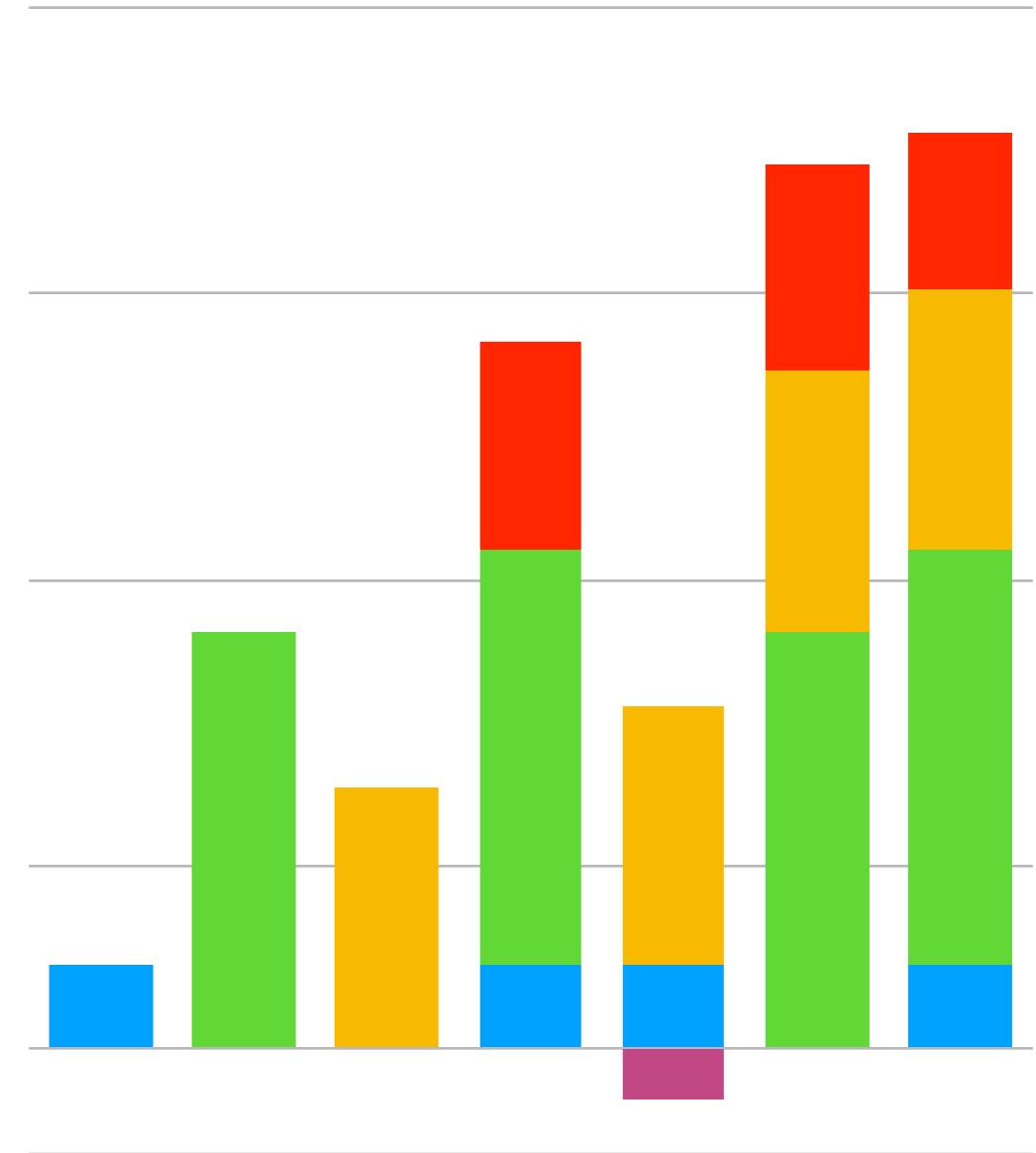


Experts

■ Confidence scores ■ Feature attribution ■ Example attribution

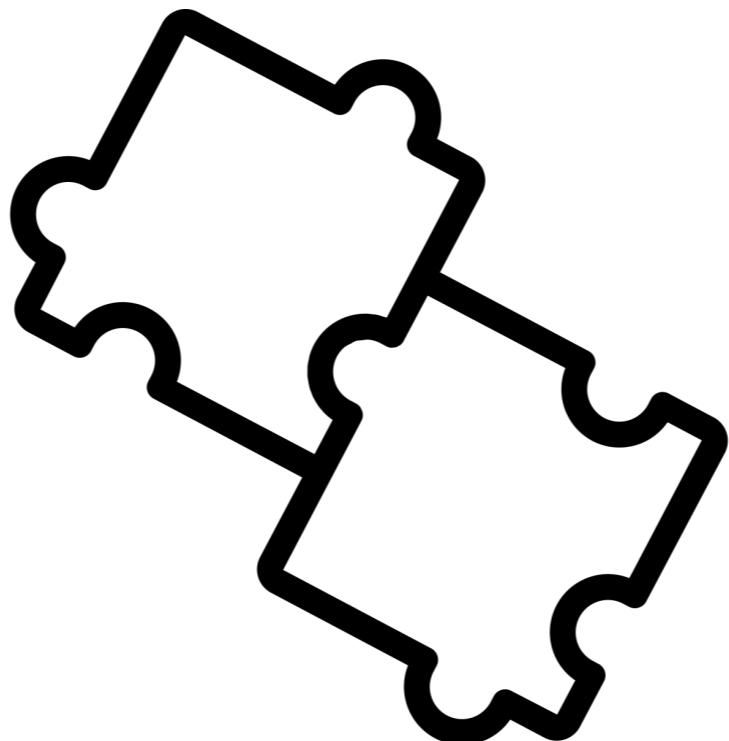


Crowdworkers



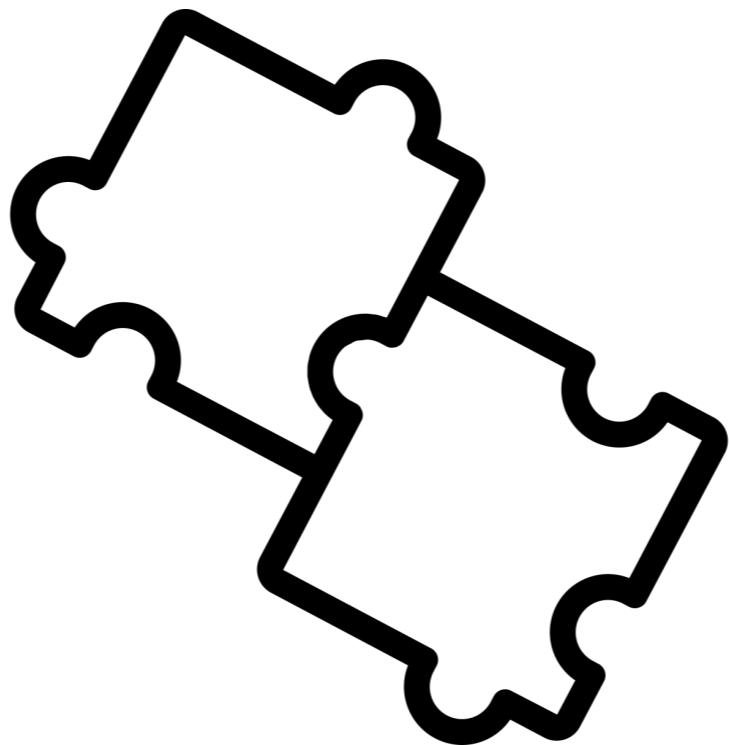
Experts

What?  
Level of details



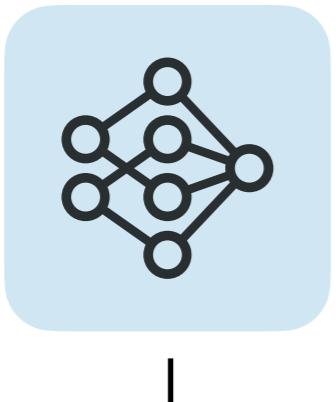
How?

What?  
Level of details

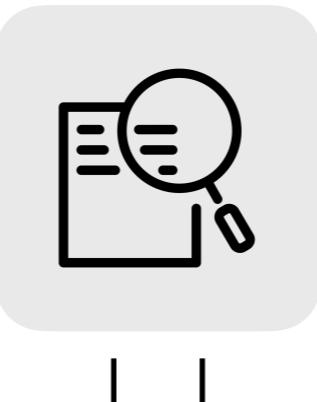


How?  
Model-based RL

System

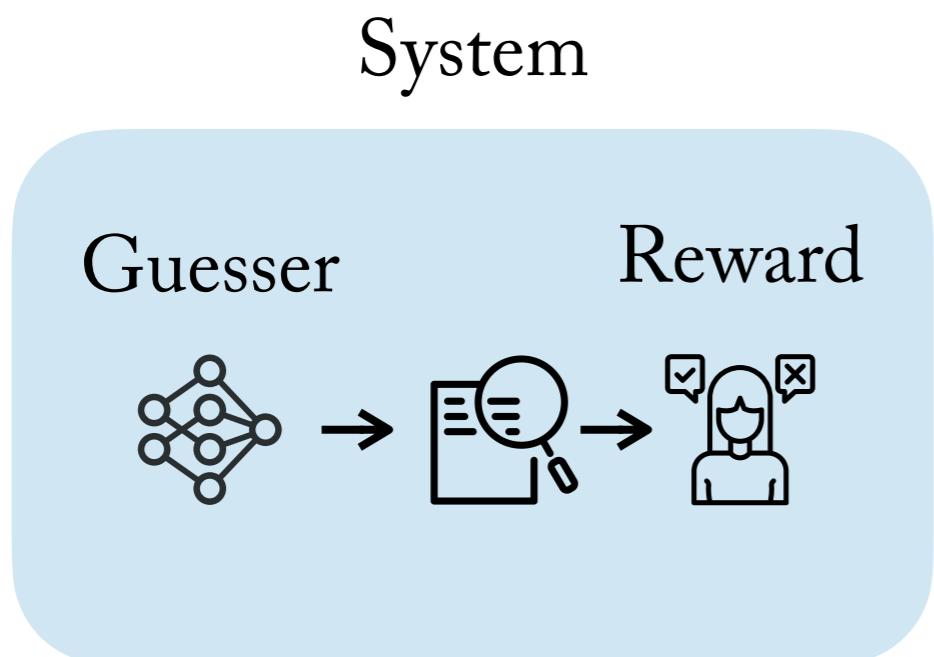
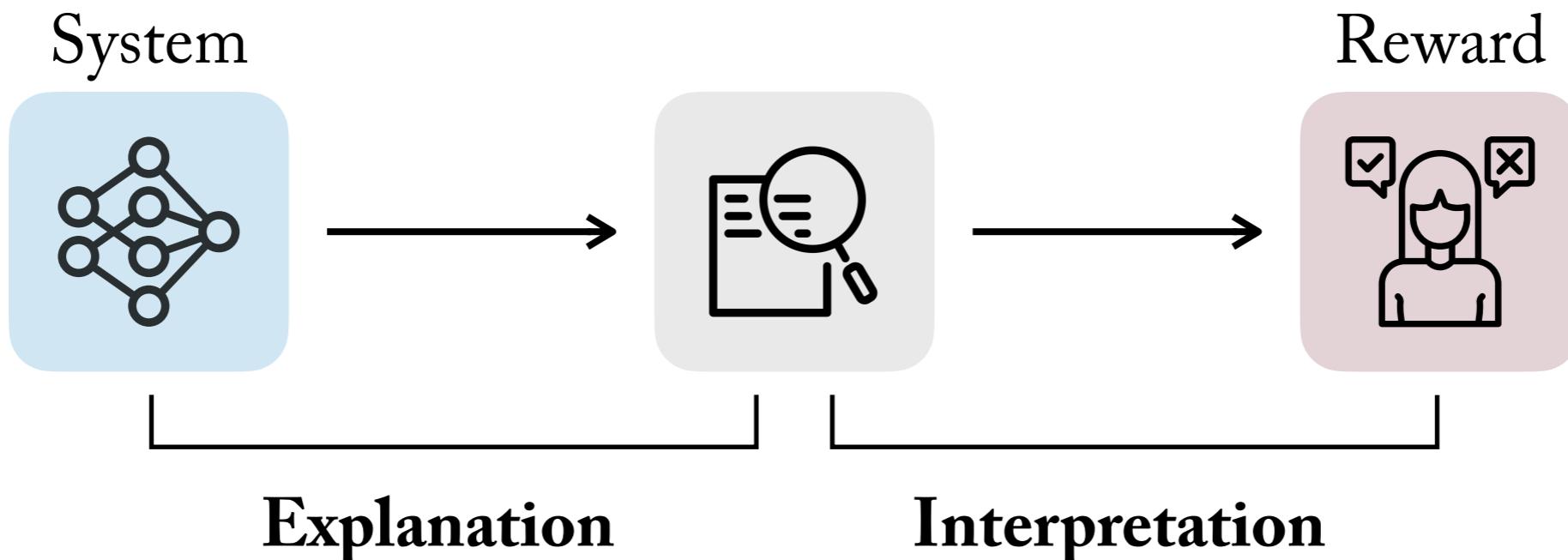


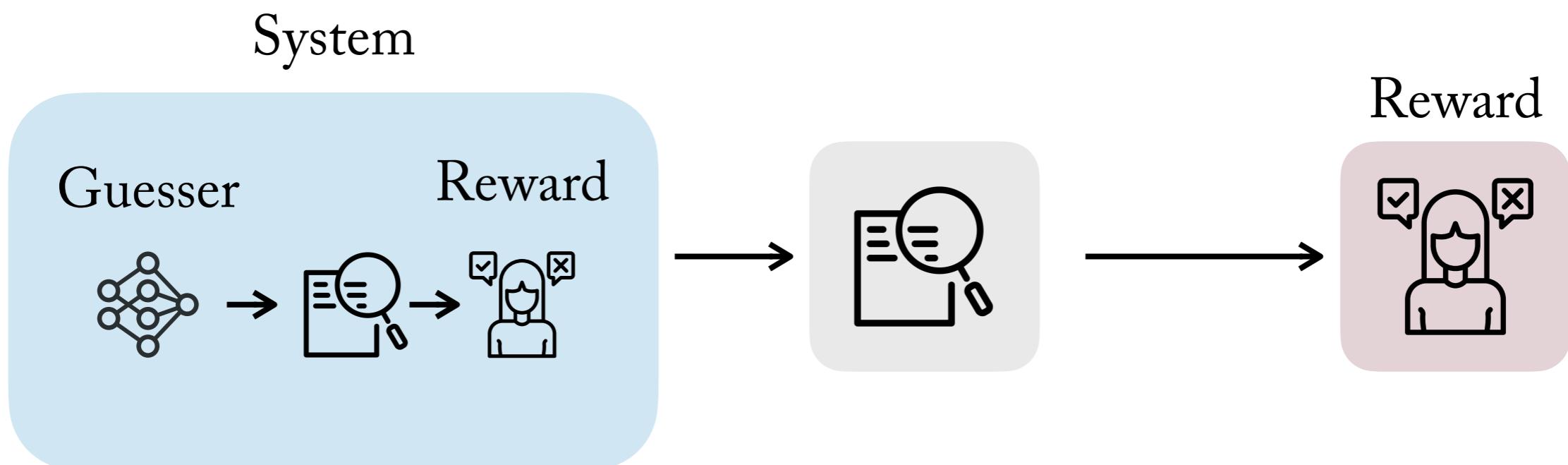
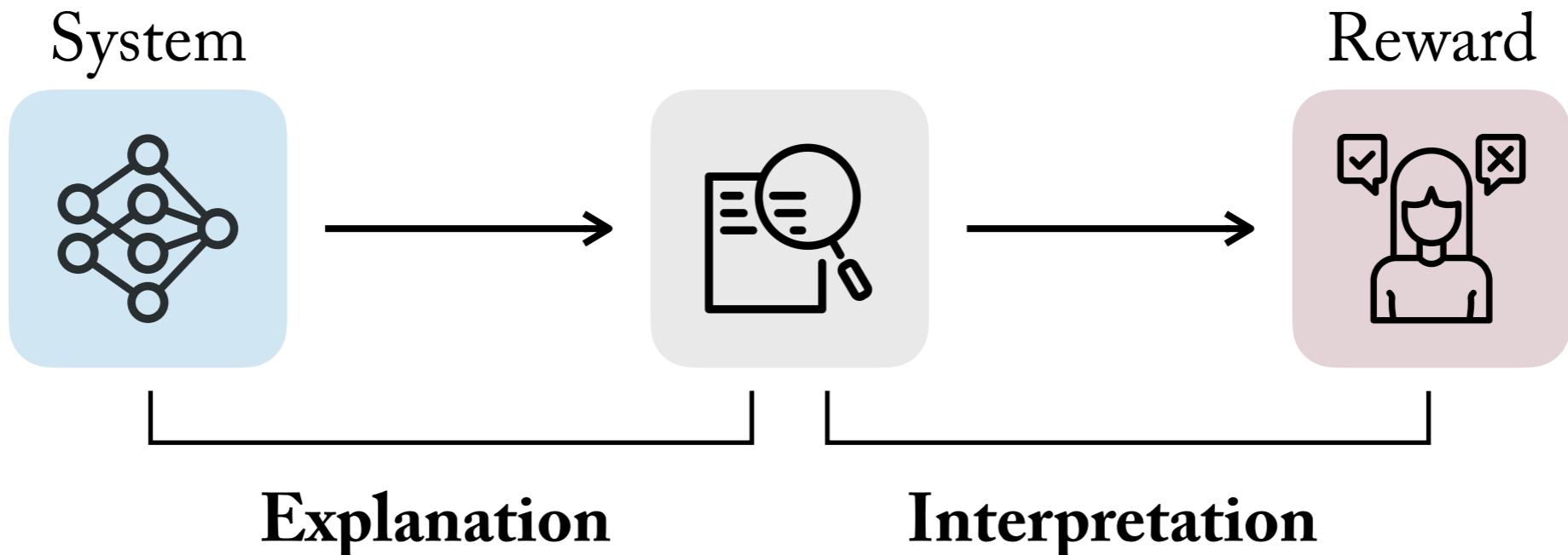
Reward

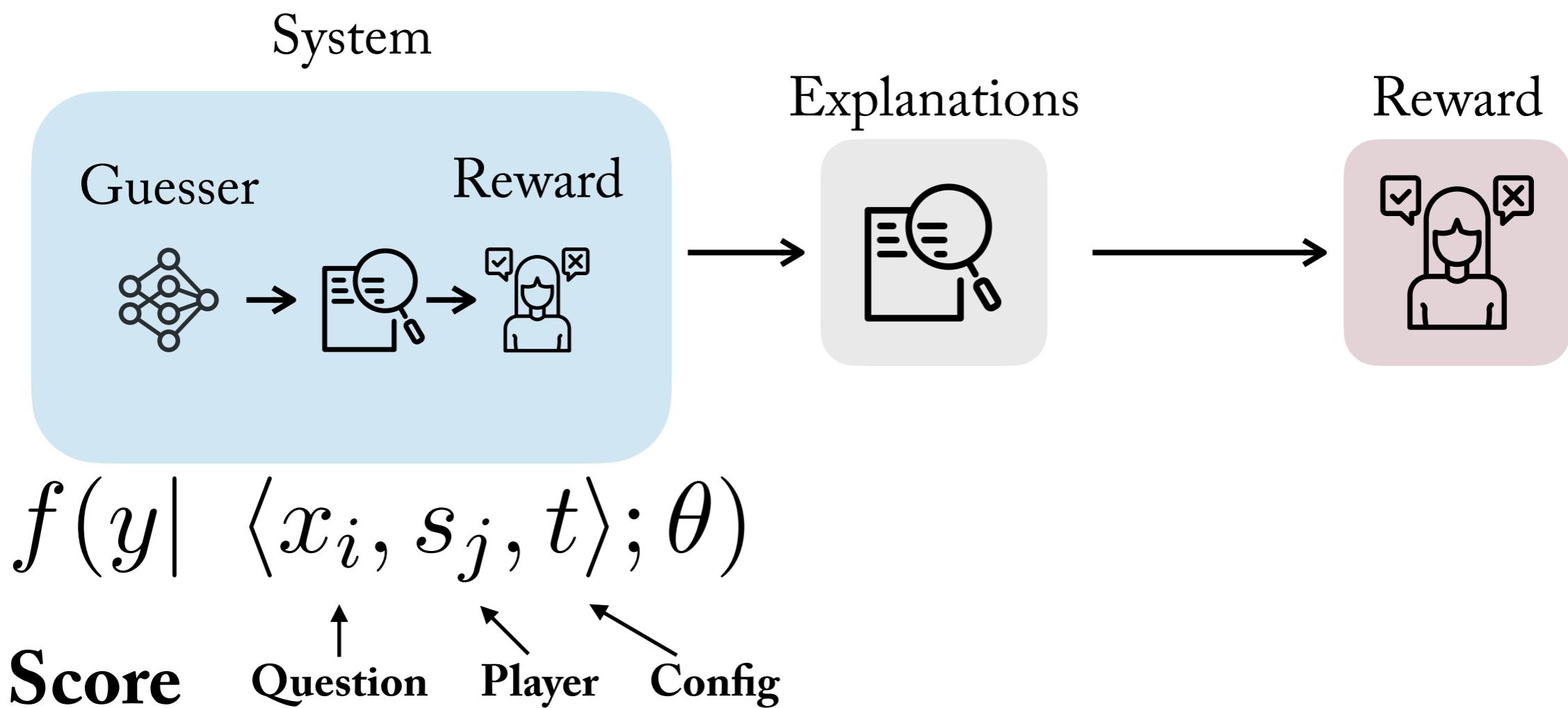
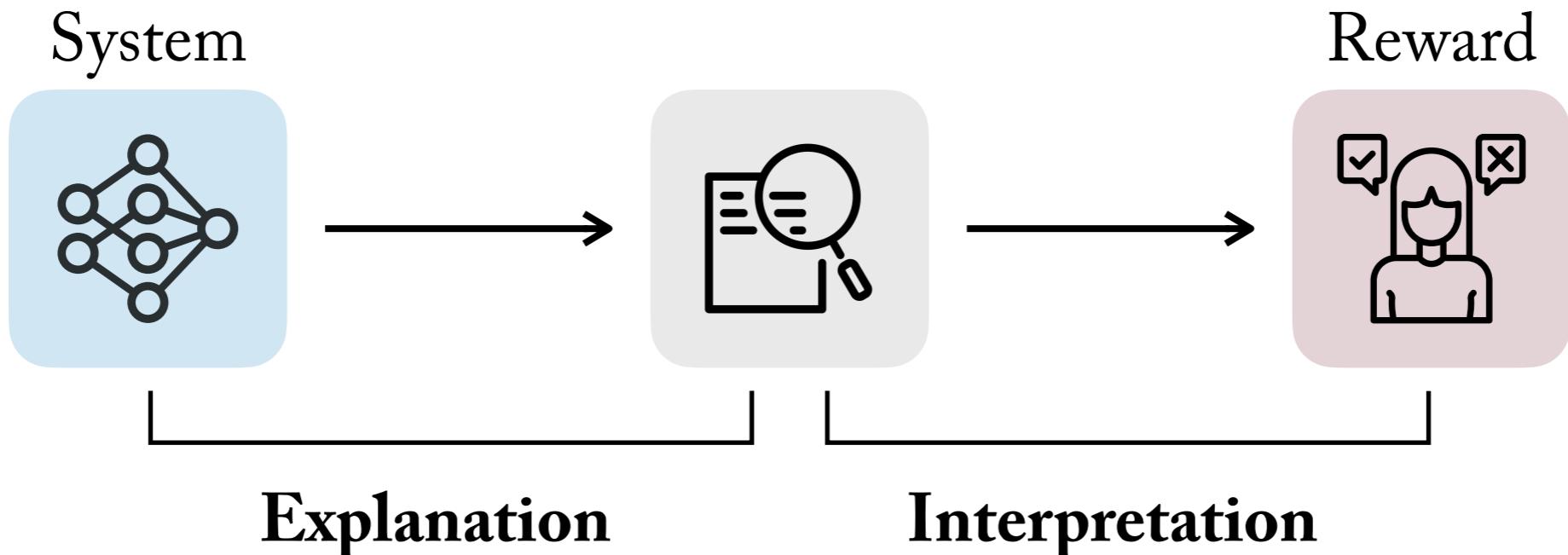


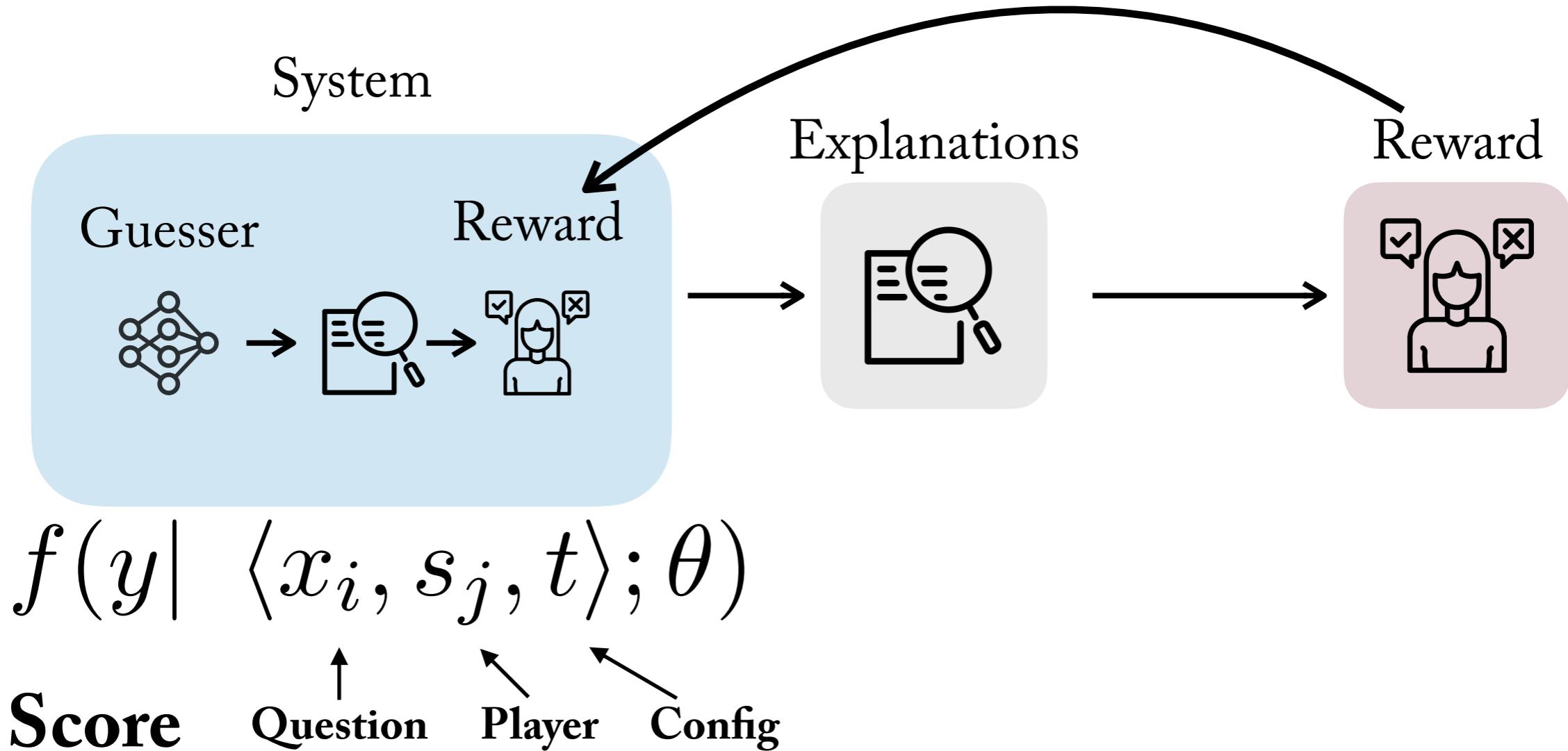
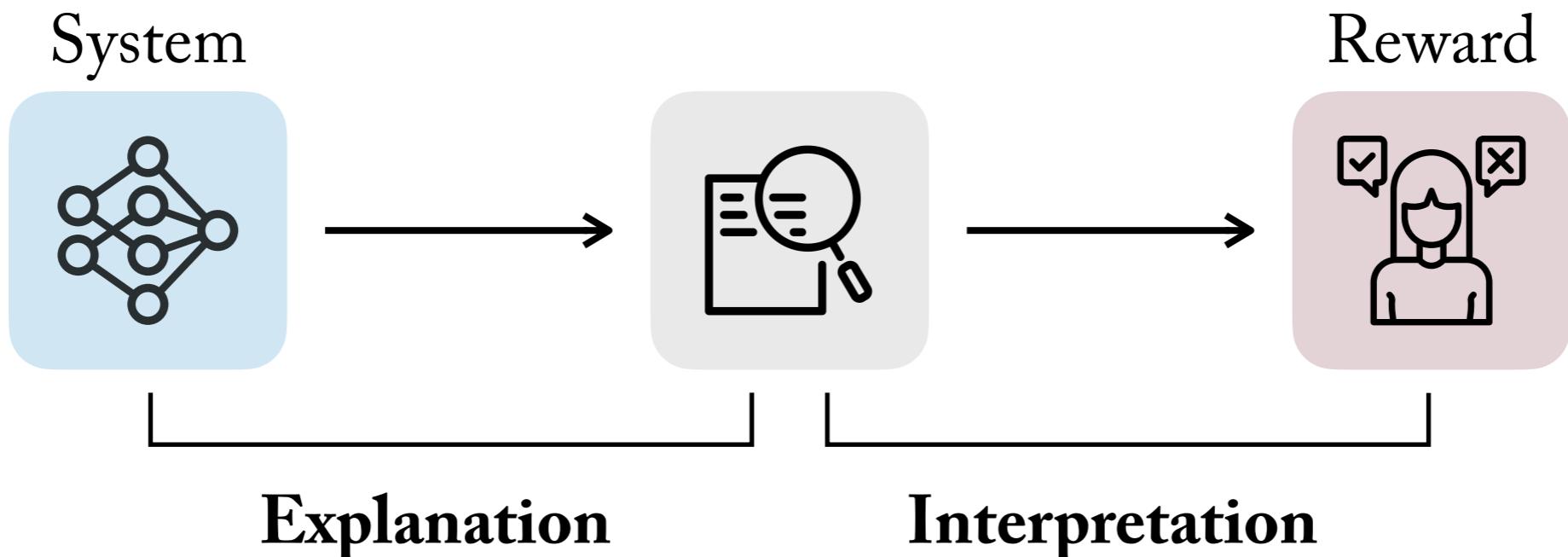
Explanation

Interpretation









# Training the explanation selector

Offline “pretraining”

$$f(y | \langle x_i, s_j \rangle; \theta)$$

Online optimization  
(Bandit)

$$f(y | \langle x_i, s_j, t \rangle; \theta)$$

LinUCB, selected from simulating players using data from previous experiment

Buzz

0:27

## Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

## Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in



Guess: Milton Friedman

## Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not , wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage product and lowering the unemployment rate . Moving along the Phillips curve, this would lead to a

# Calibration as a baseline

Buzz

0:27

## Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

## Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in



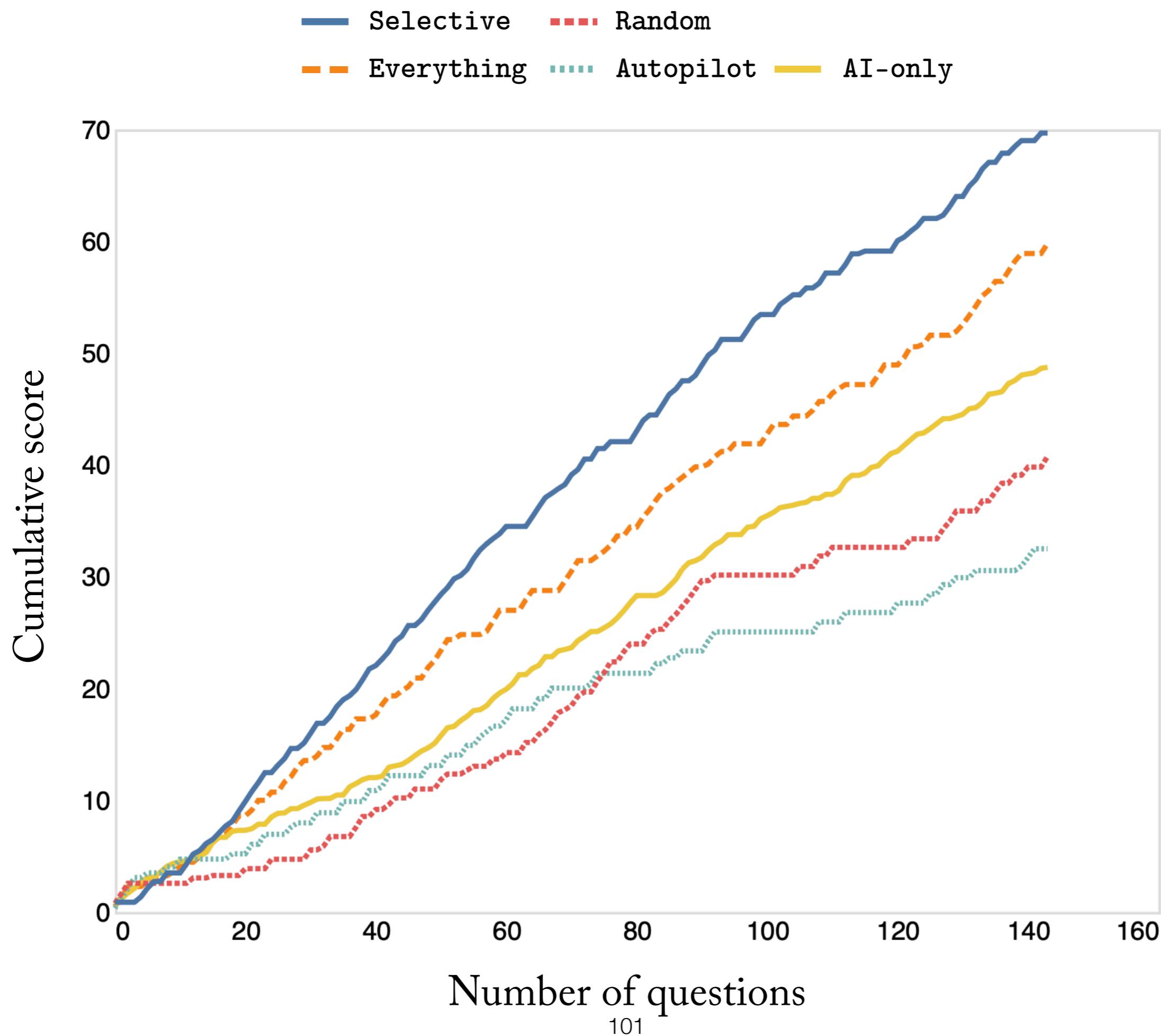
Guess: Milton Friedman

## Evidence

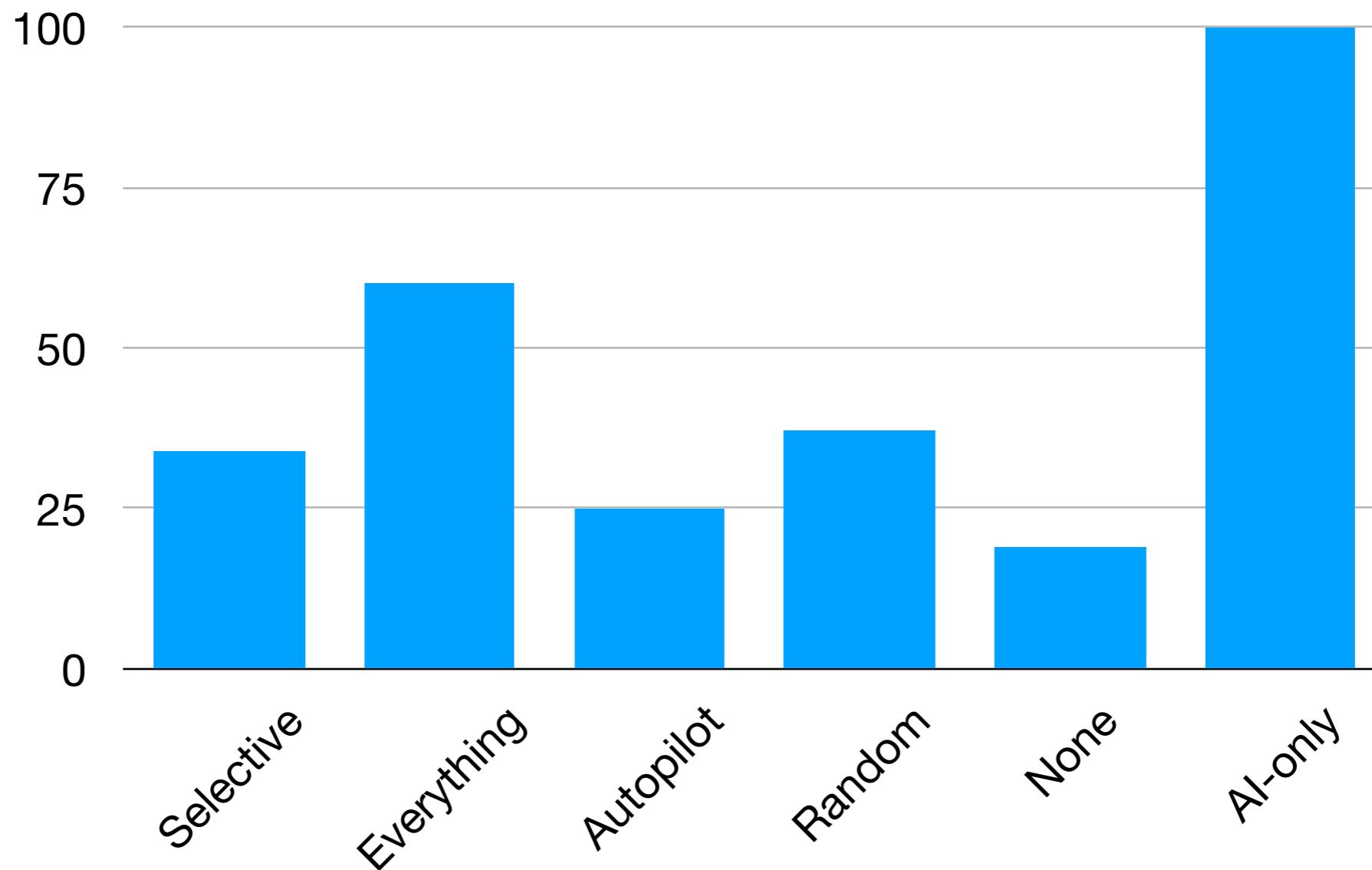
monetarists, the long-run curve is a vertical line at the natural rate of unemployment. For 10 points reversed by Robert (\*) Lucas who argued that it is the difference between real and expected inflation, not , wrote a paper in 1958 titled "The Relation between Unemployment and the Rate of Change of Money Wage product and lowering the unemployment rate . Moving along the Phillips curve, this would lead to a

# Calibration as a baseline

- Autopilot: given all the information the human sees about the model, does the model think the human should buzz?
- Red = yes, black = no
- It's an approximation of the optimal first-order buzzer.



# Failure to reject incorrect prediction



# What did we learn?



1. Modeling the interpretation process with a “listener model”
2. Warm-starting for few-shot learning the listener model
3. Adjusting the level of details
4. Engagement is crucial

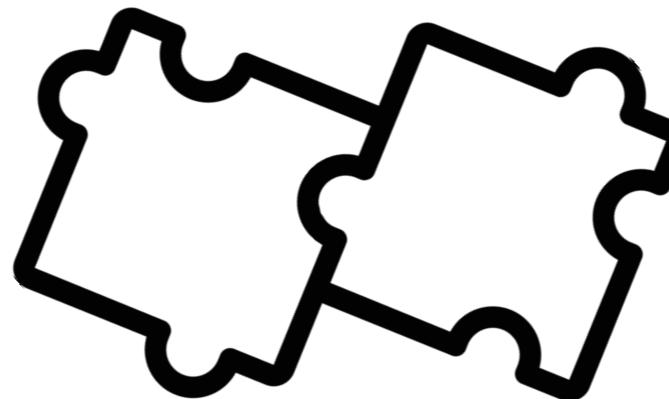
What?  
Level of details

How?  
Model-based RL

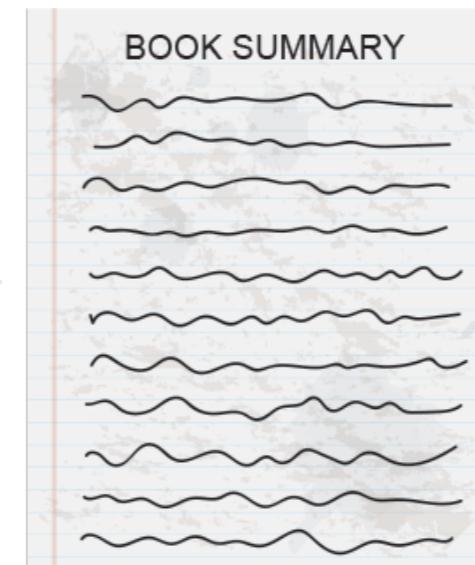
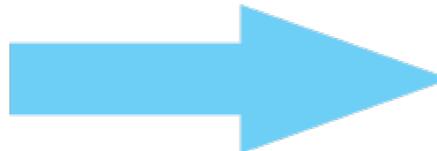
Two interlocking puzzle pieces, one white and one black, positioned between the 'What?' and 'How?' sections.

# What's next? More flexible listener model

What?  
Level of details



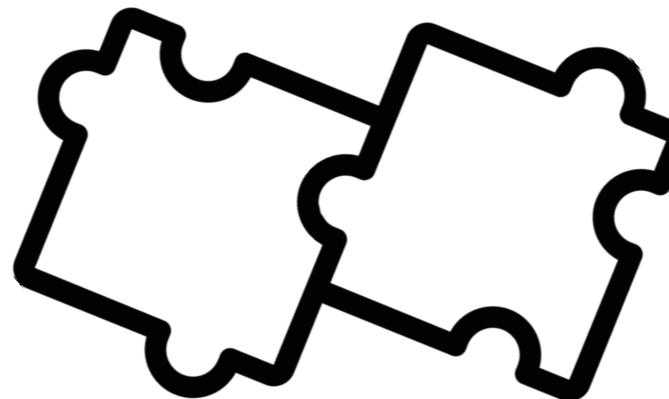
How?  
Model-based RL



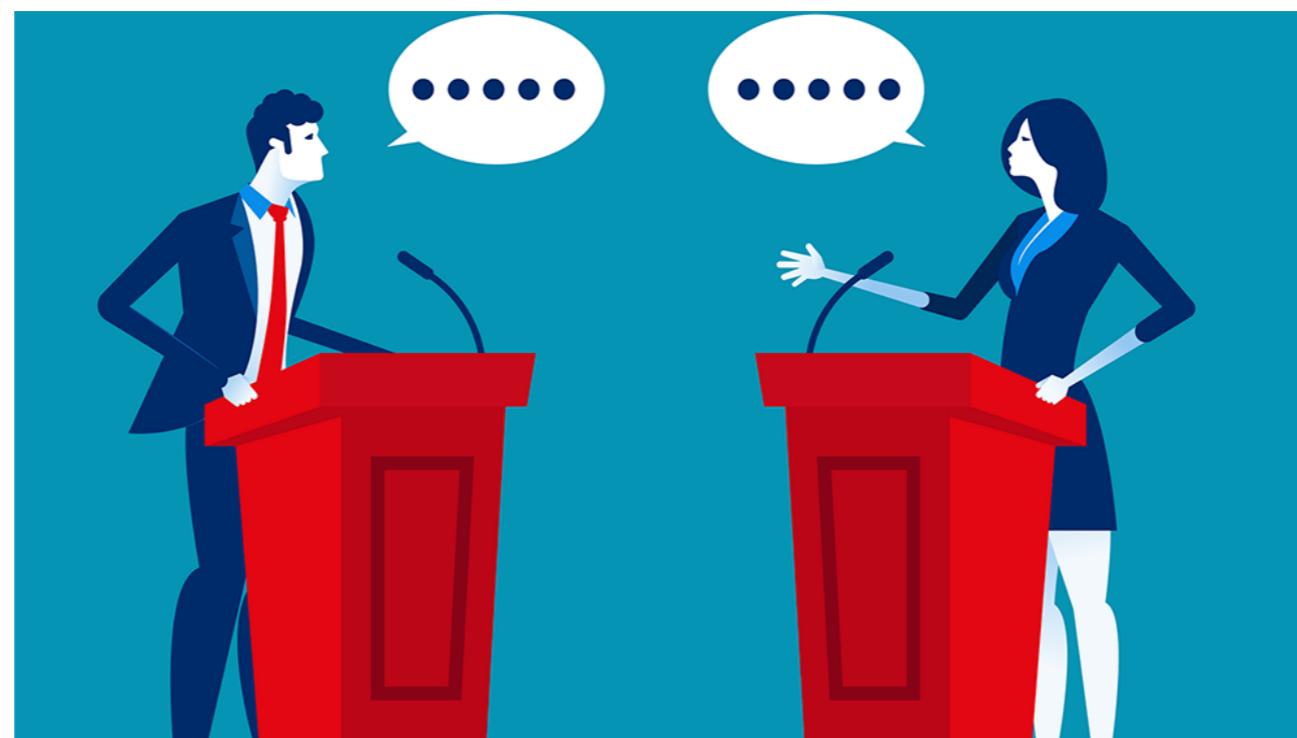
Bandit -> RL

# What's next? More flexible listener model

What?  
Plans

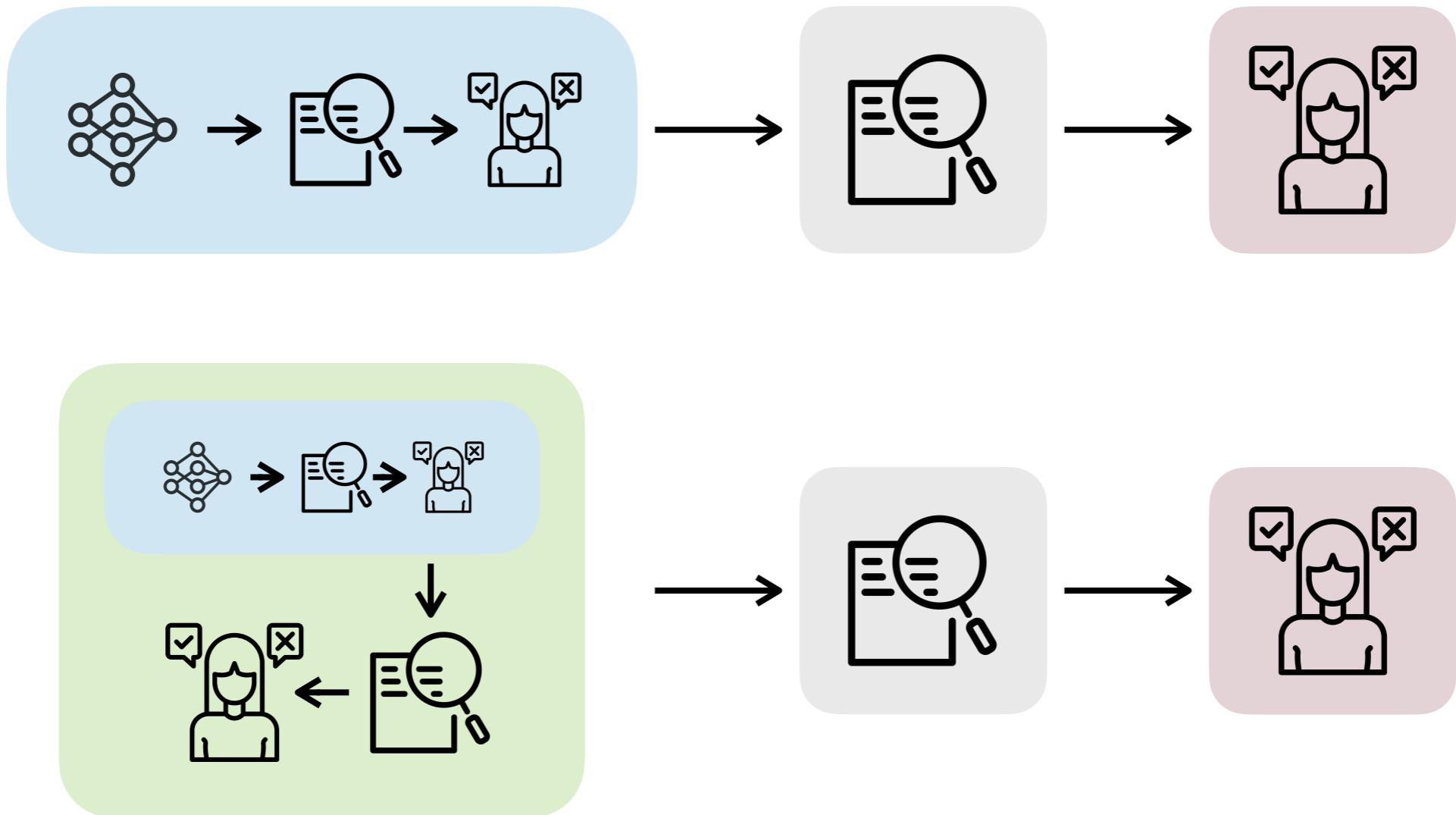


How?  
Model-based RL



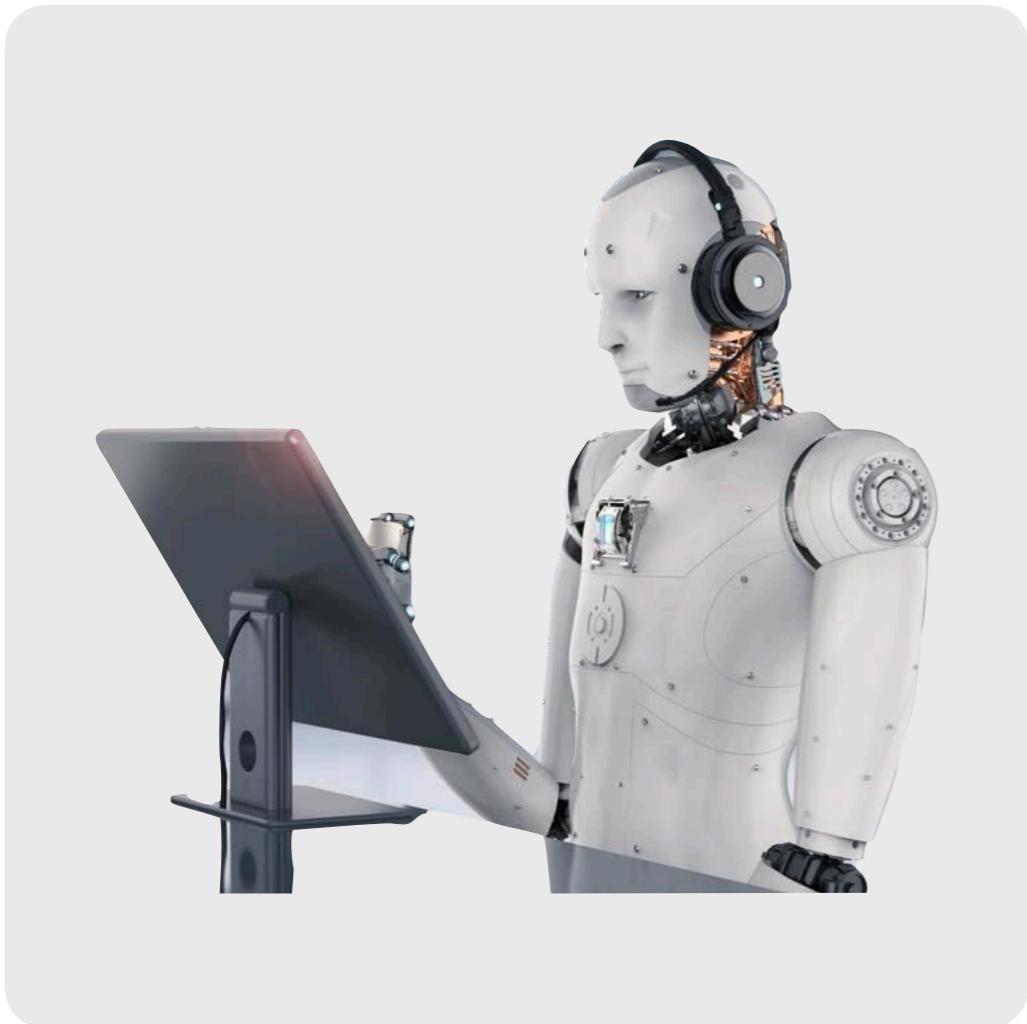
Debate

# What's next? Theories of pragmatic explanations



1. Pragmatic inference
2. Moral philosophy & ethics: deontology / consequentialism

# AIs as...

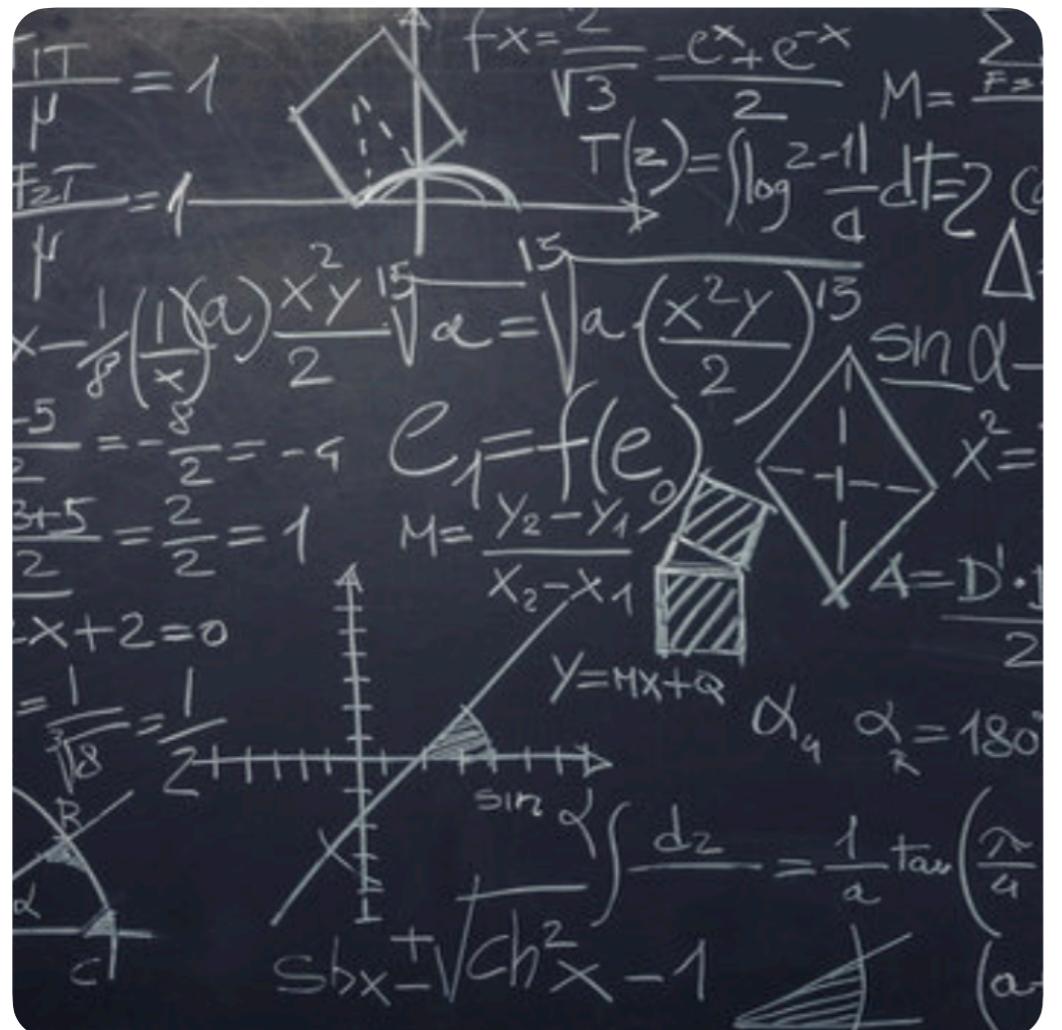


Human replacement

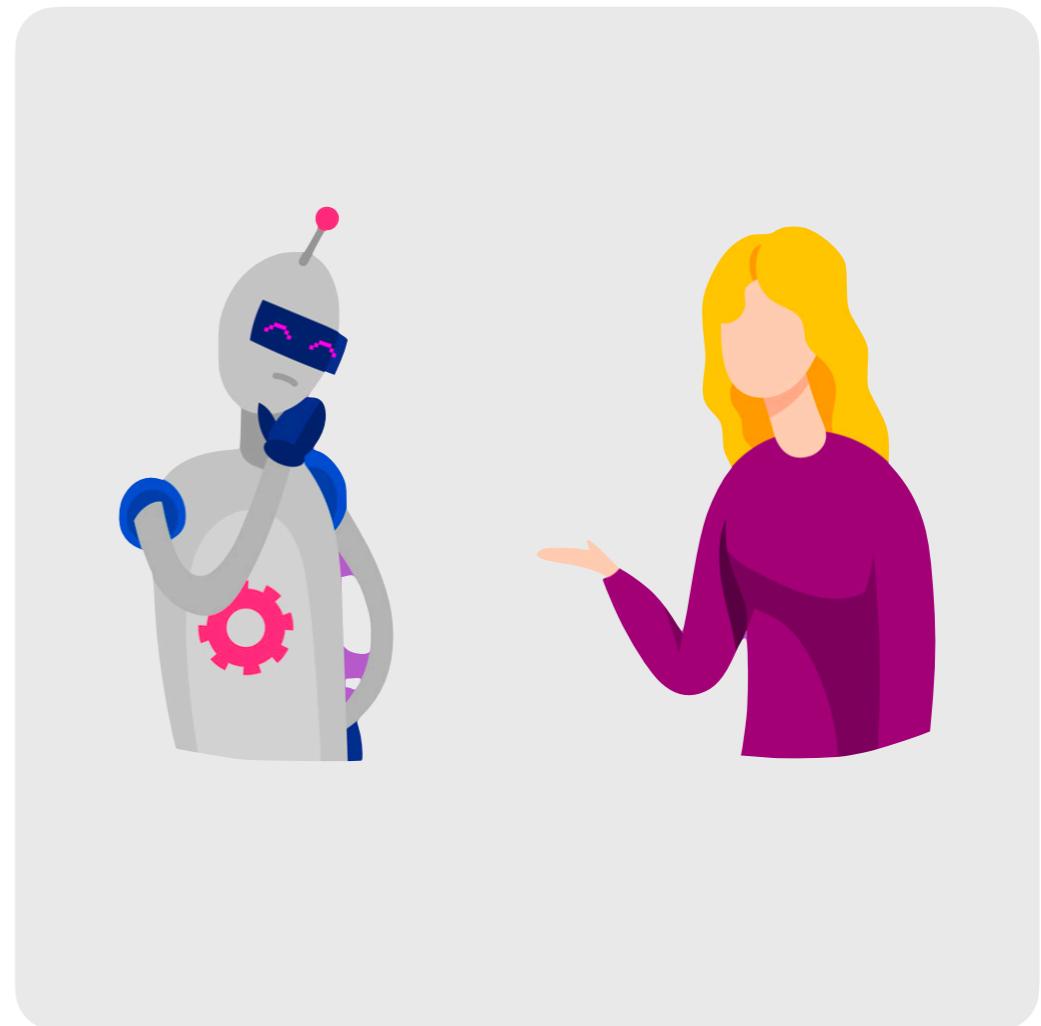


More and more useful  
computational tools

# Control AIs with...



Theories  
Fixed, Formal, General



Explanations  
Iterative, Intuitive, task-specific

# Focus of interpretability research...

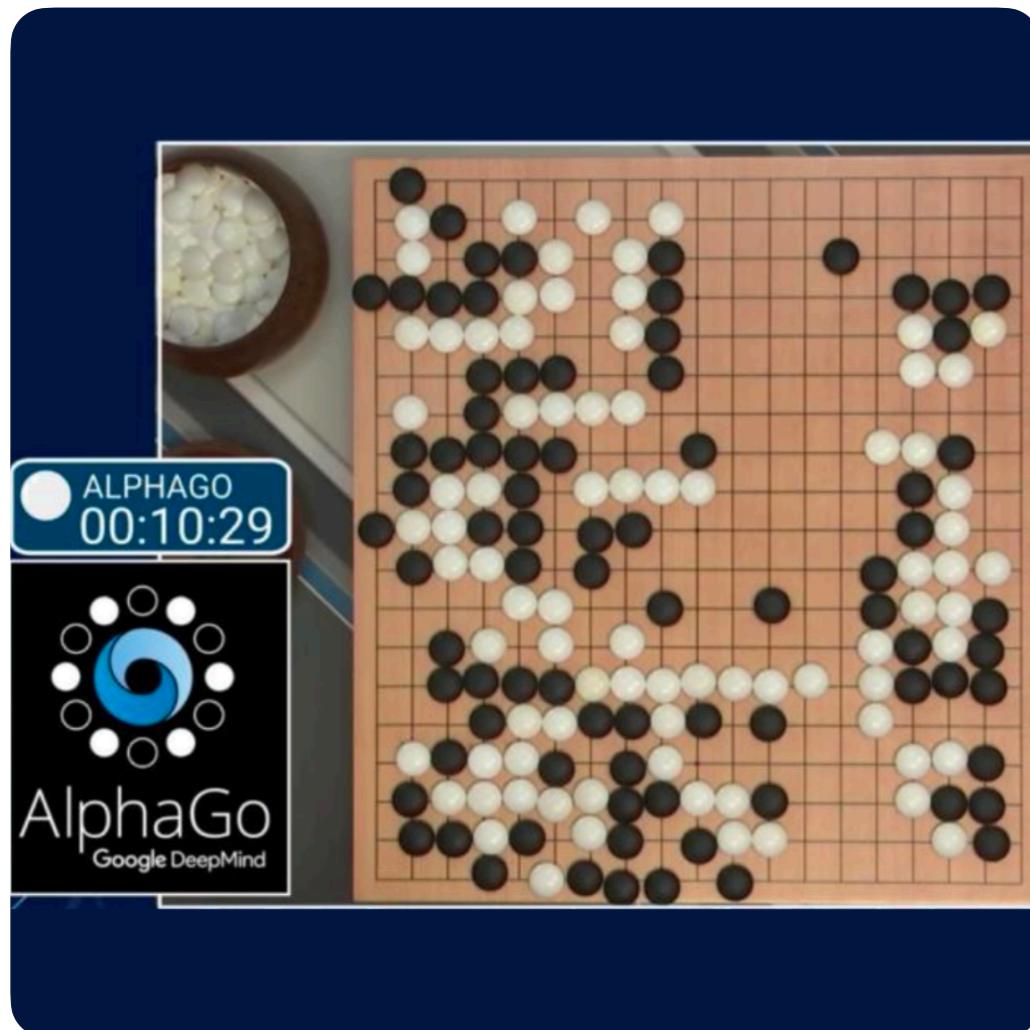


Explanation methods

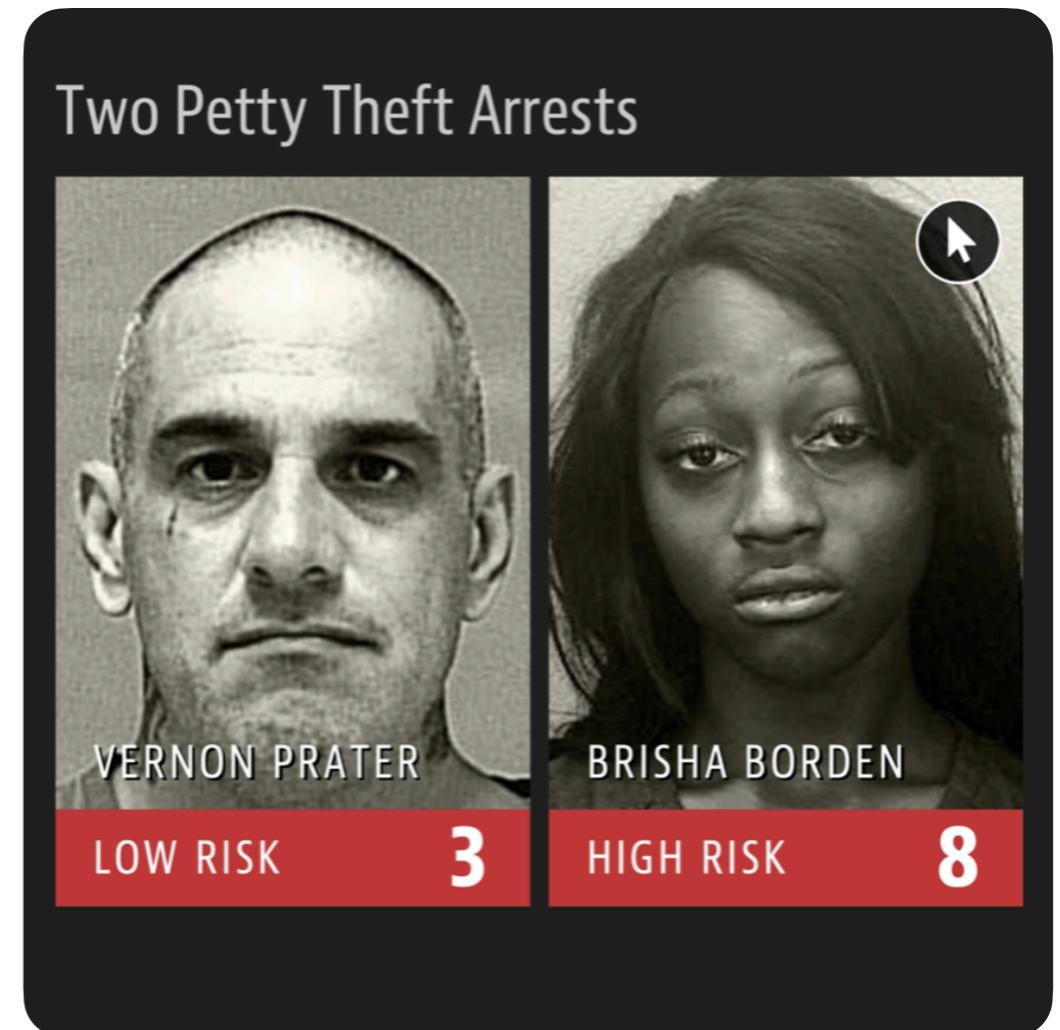


Objectives to incentivize  
better explanations

# Latent knowledge about...

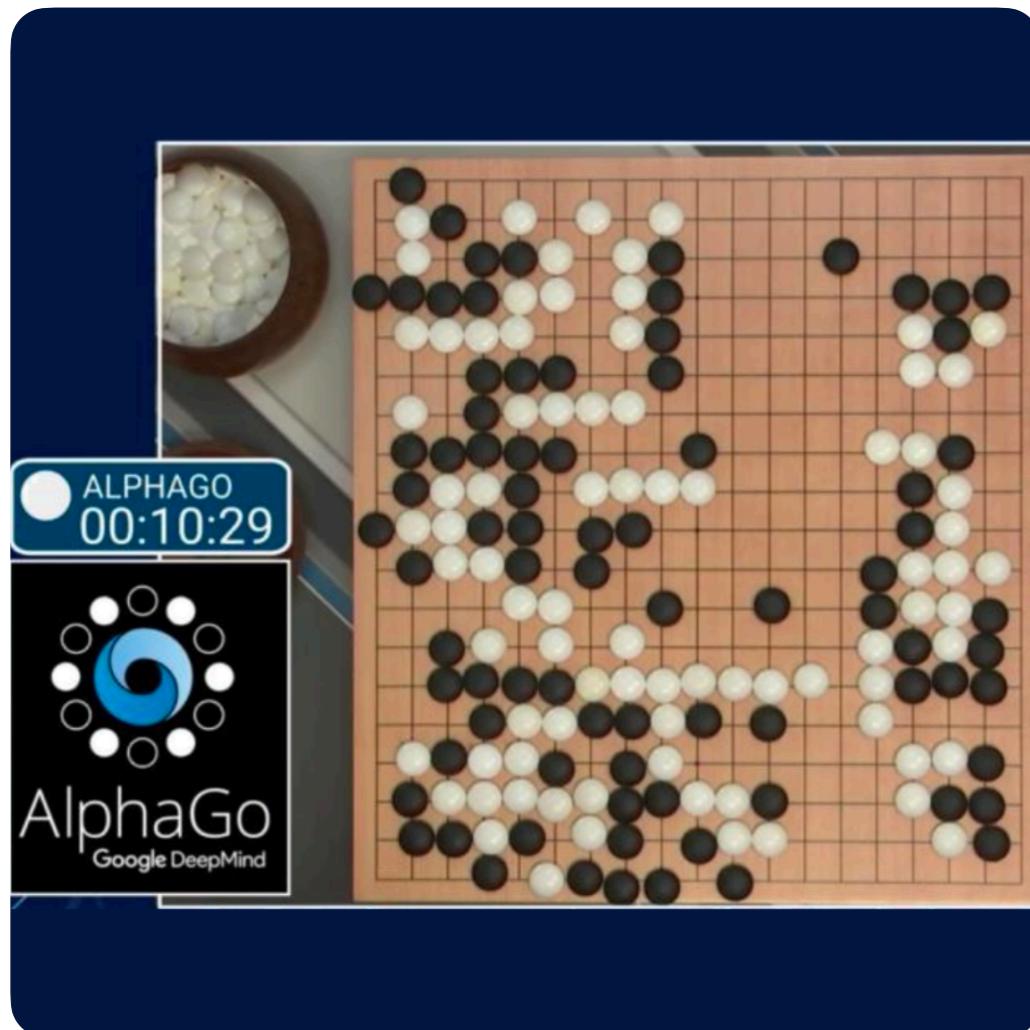


Our actual goals



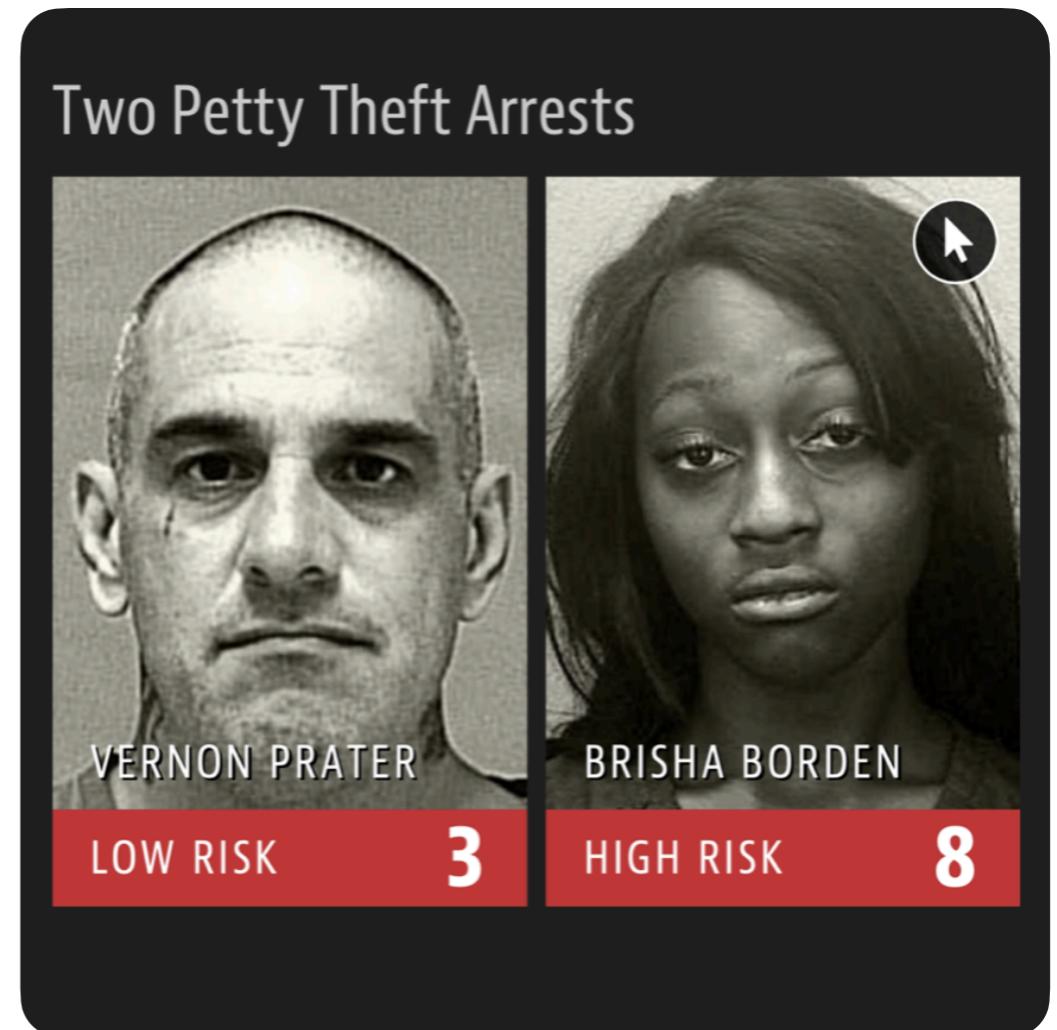
Measurements of goals

# Latent knowledge about...



Our actual goals

One mission is to empower people with this knowledge.



Measurements of goals

# What's next? Modeling imperfect knowledge

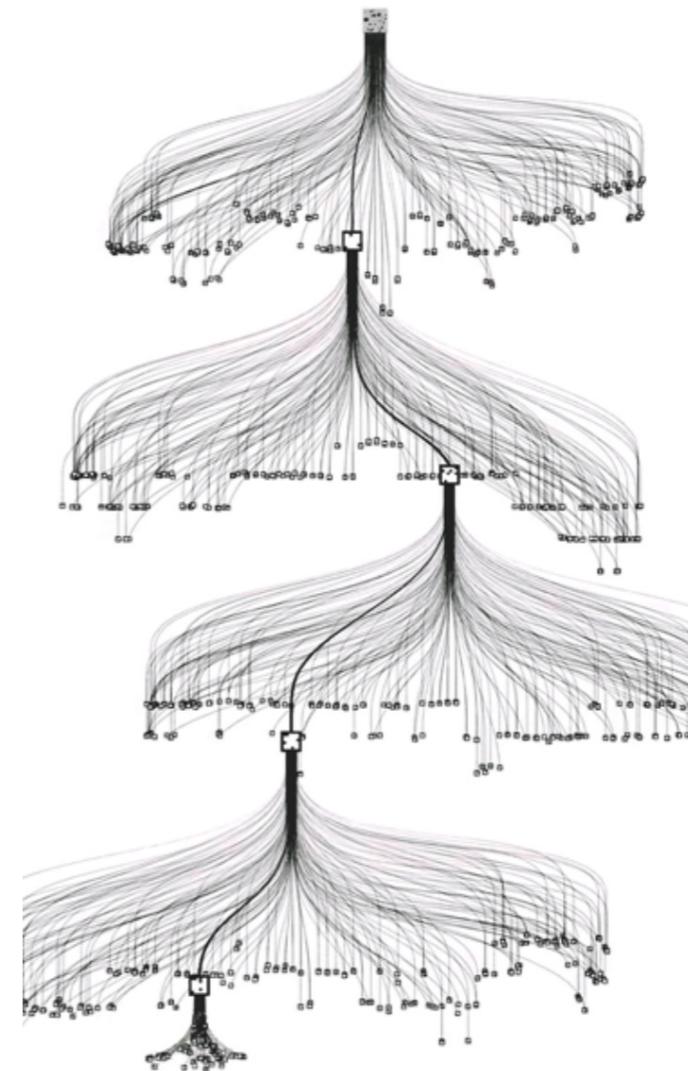
1. Recommendation systems
2. Machine teaching: tutorials and supports for radiologists



ICLR 23

# What's next? Inductive bias for listener model

1. LLM unlocks new potentials to do pragmatic inference
2. Dialogue
3. “Meta-learning” with instructions



# Thanks for listening!

