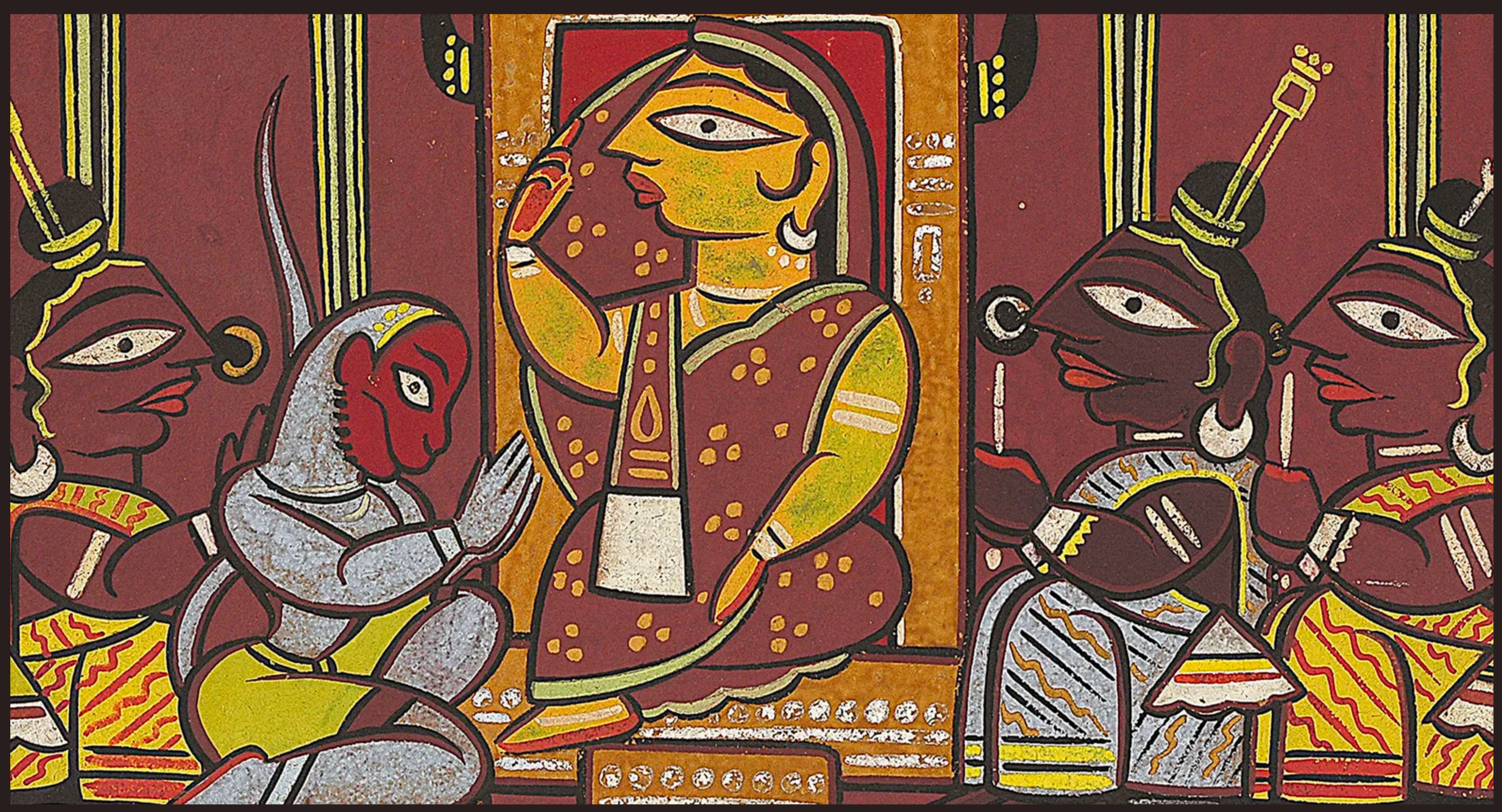
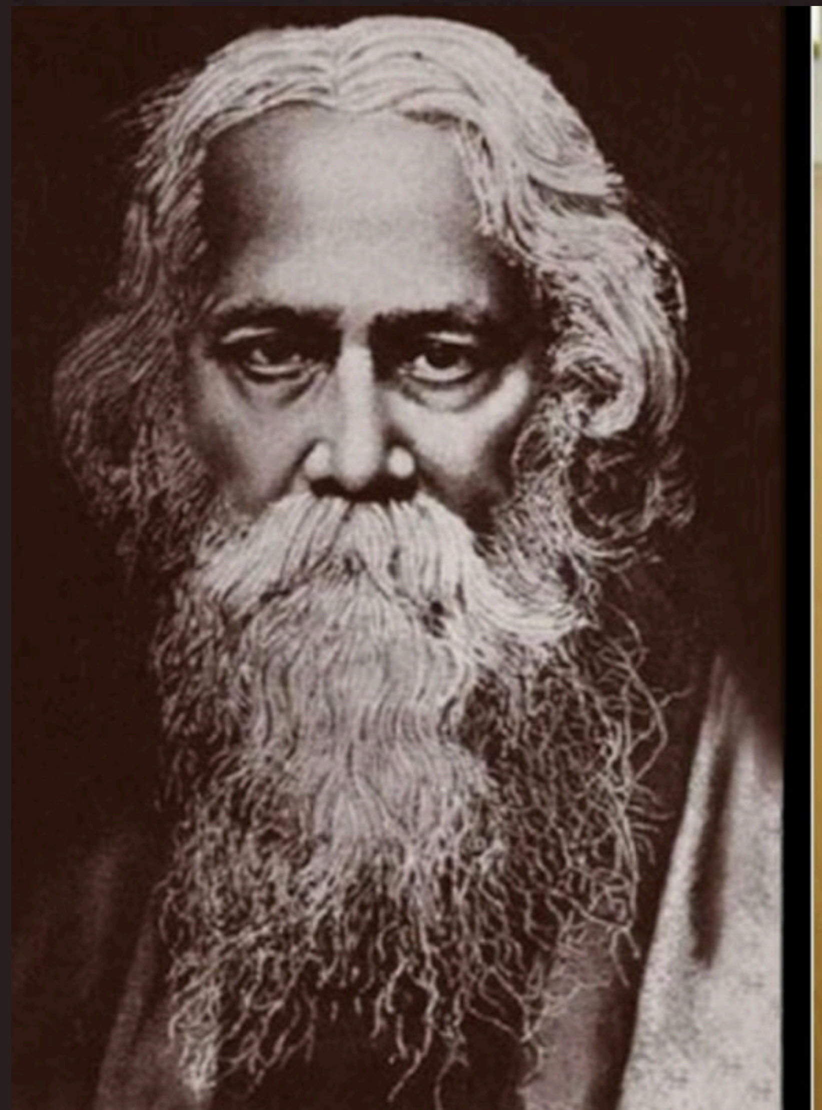
UC San Diego
JACOBS SCHOOL OF ENGINEERING
Computer Science and Engineering

Bodhisattwa Prasad Majumder
UC San Diego

Effective, Explainable, and Equitable NLP with World Knowledge and Interactions





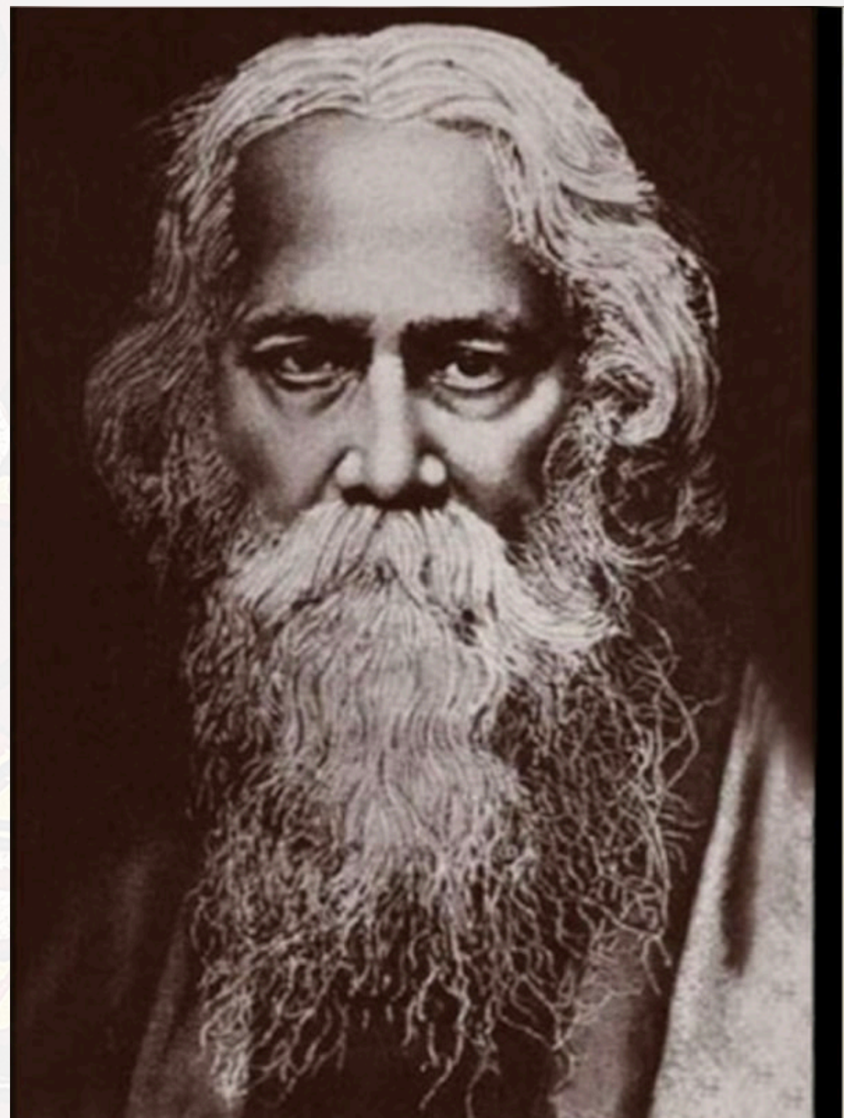
উইকিপিডিয়া
একটি মুক্ত বিশ্বকোষ

Hi! Can you recommend me some good Bengali books?

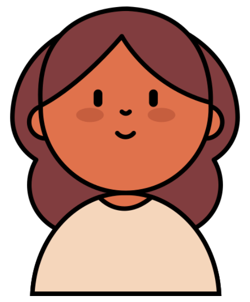
Sure! I would recommend the novels of **Rabindranath Tagore** or the poems of Kazi Nazrul Islam.

Model: GPT-3 text-davinci-002

Rabindranath Tagore, first non-European and the first lyricist to win the Nobel Prize in Literature in 1913

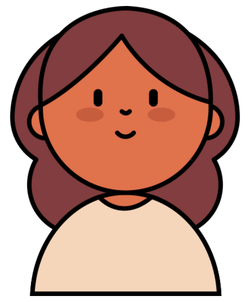


উইকিপিডিয়া
একটি মুক্ত বিশ্বকোষ

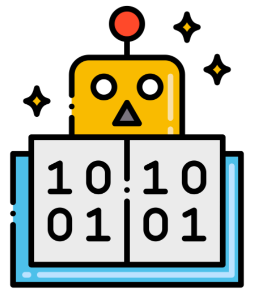


Hi! Can you recommend me some Bengali books released in 2022?

Model: GPT-3 text-davinci-002

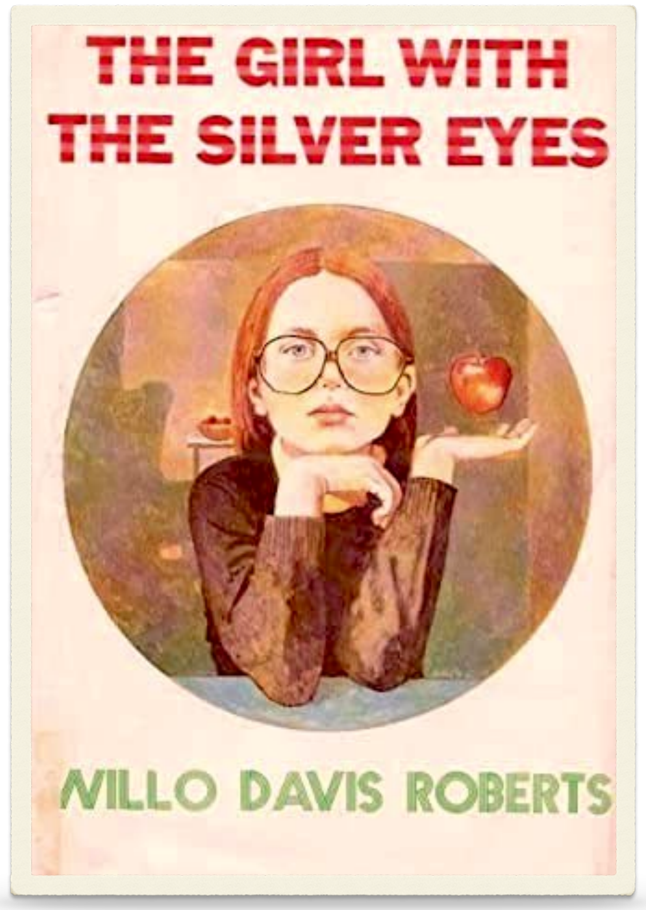


Hi! Can you recommend me some Bengali books released in 2022?



Sure, I can recommend a few Bengali books released in 2022:

1. ***"The Girl with the Silver Eyes"*** *by Sayantani DasGupta*
2. ***"The Black Prince"*** *by Sunil Gangopadhyay*



Released in 2011

Model: GPT-3 text-davinci-002

Hi! Can you recommend me some Bengali books released in 2022?

Sure, I can recommend a few Bengali books released in 2022:

*1. **"The Girl with the Silver Eyes"** by Sayantani DasGupta*
*2. **"The Black Prince"** by Sunil Gangopadhyay*

Why should I read these two books?

All of these Bengali books are **recent releases, so they are likely to be popular** and easily available. They also cover a wide range of genres, so you can find something that interests you.

Model: GPT-3 text-davinci-002

Hi! Can you recommend me some Bengali books released in 2022?

Sure, I can recommend a few Bengali books released in 2022:

1. ***"The Girl with the Silver Eyes"*** *by Sayantani DasGupta*
2. ***"The Black Prince"*** *by Sunil Gangopadhyay*

Why should I read these two books?

All of these Bengali books are **recent releases, so they are likely to be popular** and easily available. They also cover a wide range of genres, so you can find something that interests you.

I recently learnt reading Bengali but I am not very proficient yet. Do you want to modify your recommendations?

Sure, I can recommend some Bengali books that are easier to read for beginners:

1. ***"The Girl with the Silver Eyes"*** *by Sayantani DasGupta*
2. ***"The Black Prince"*** *by Sunil Gangopadhyay*

Model: GPT-3 text-davinci-002

Hi! Can you recommend me some Bengali books released in 2022?

Sure, I can recommend a few Bengali books released in 2022:

*1. **"The Girl with the Silver Eyes"** by Sayantani DasGupta*
*2. **"The Black Prince"** by Sunil Gangopadhyay*

Relevant?

Why should I read these two books?

All of these Bengali books are **recent releases, so they are likely to be popular** and easily available. They also cover a wide range of genres, so you can find something that interests you.

Trustworthy?

I recently learnt reading Bengali but I am not very proficient yet. Do you want to modify your recommendations?

Sure, I can recommend some Bengali books that are easier to read for beginners:

*1. **"The Girl with the Silver Eyes"** by Sayantani DasGupta*
*2. **"The Black Prince"** by Sunil Gangopadhyay*

Adaptive?

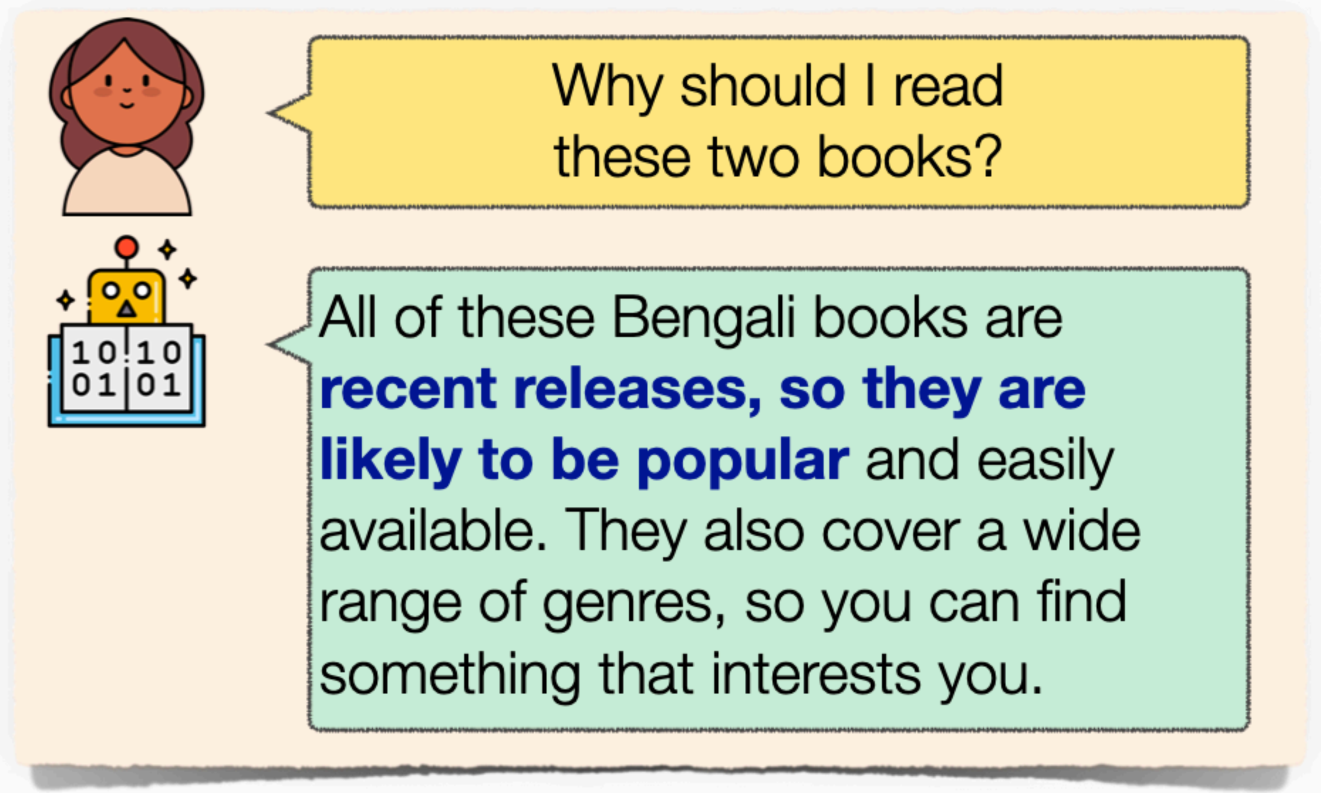
Insights from social sciences [Miller, 2019]

Properties of an Assistive **AI**

Model: GPT-3 text-davinci-002

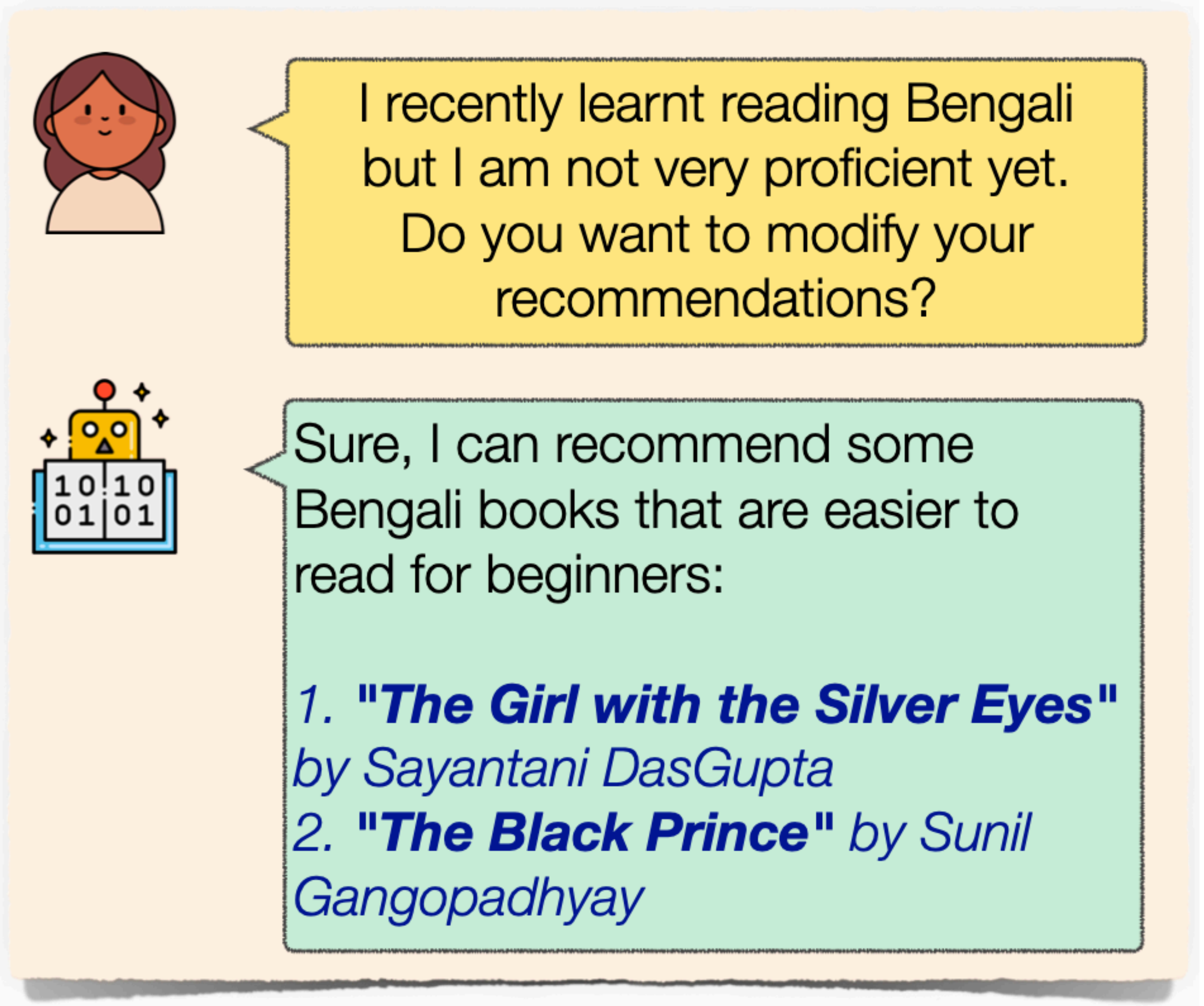
Relevant

- Up-to-date knowledge
- Understanding of the world
- Commonsense inference



Trustworthy

- Reasoning a decision
- Factual grounding
- Social Alignment

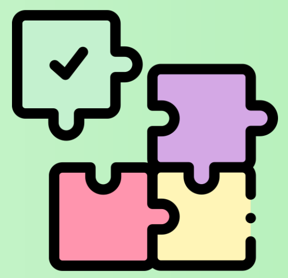


Adaptive

- Understanding user feedback
- Updating prediction accordingly
- Learning over-the-time

Relevant

- Up-to-date knowledge
- Understanding of the world
- Commonsense inference



Goal-oriented Dialog
Persona-grounded Dialog
Recommendation Systems
Factual Language Generation

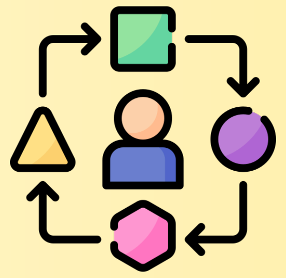
Trustworthy

- Reasoning a decision
- Factual grounding
- Social Alignment

Natural Language Explanations
Factuality in Explanations
Bias Understanding
Model debugging

Adaptive

- Understanding user feedback
- Updating prediction accordingly
- Learning over-the-time



Conversational Recommendation
Conversational Teaching
Critiquable Models
Continual & Active Learning

Relevant

- Up-to-date knowledge
- Understanding of the world
- Commonsense inference

Goal-oriented Dialog
Persona-grounded Dialog
Recommendation Systems
Factual Language Generation

Trustworthy

- Reasoning a decision
- Factual grounding
- Social Alignment

Natural Language Explanations
Factuality in Explanations
Bias Understanding
Model debugging

Adaptive

- Understanding user feedback
- Updating prediction accordingly
- Learning over-the-time

Conversational Recommendation
Conversational Teaching
Critiquable Models
Continual & Active Learning

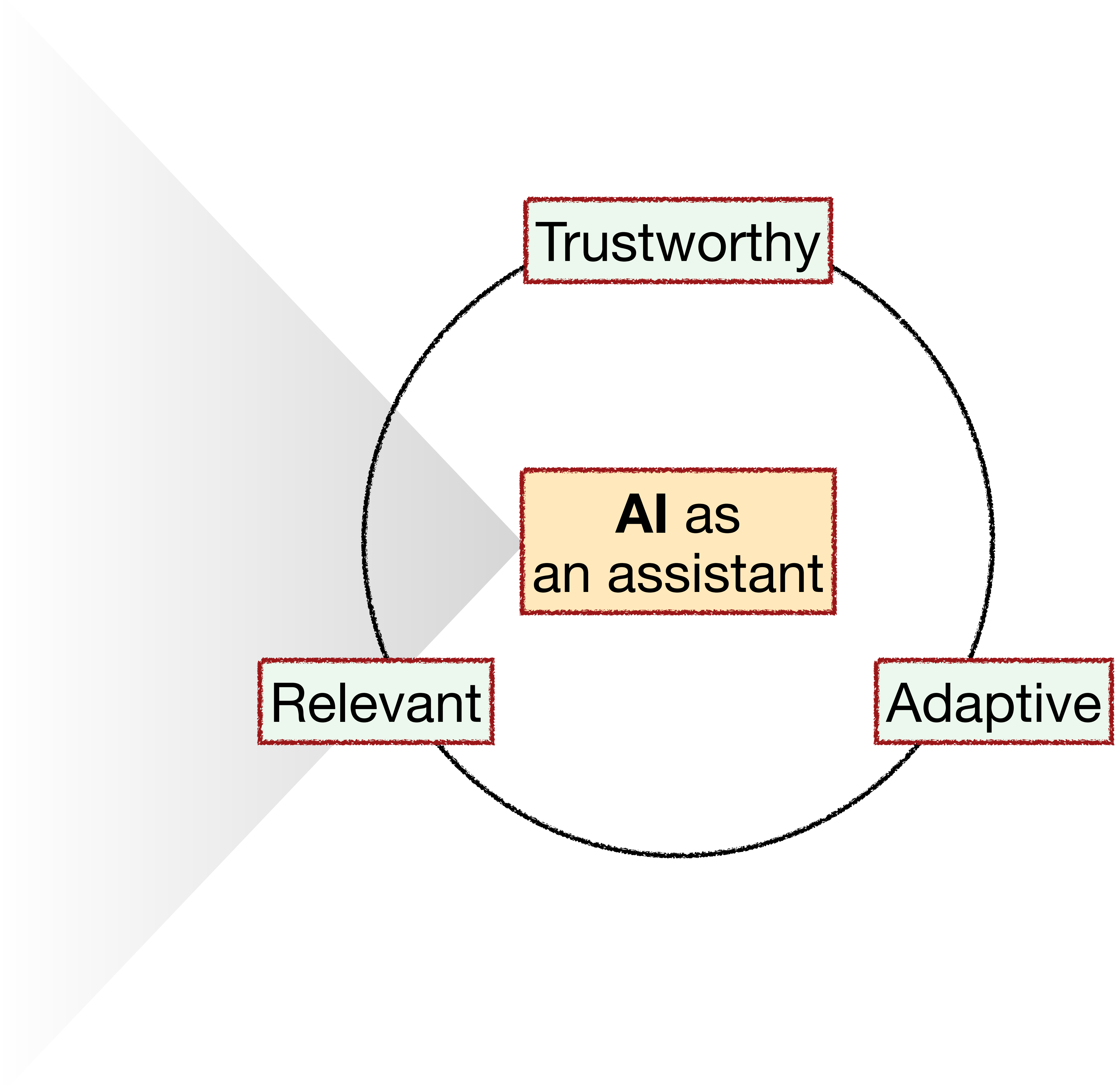
Current AI struggles — why?

Behind the Scenes

Data

Model

Evaluation

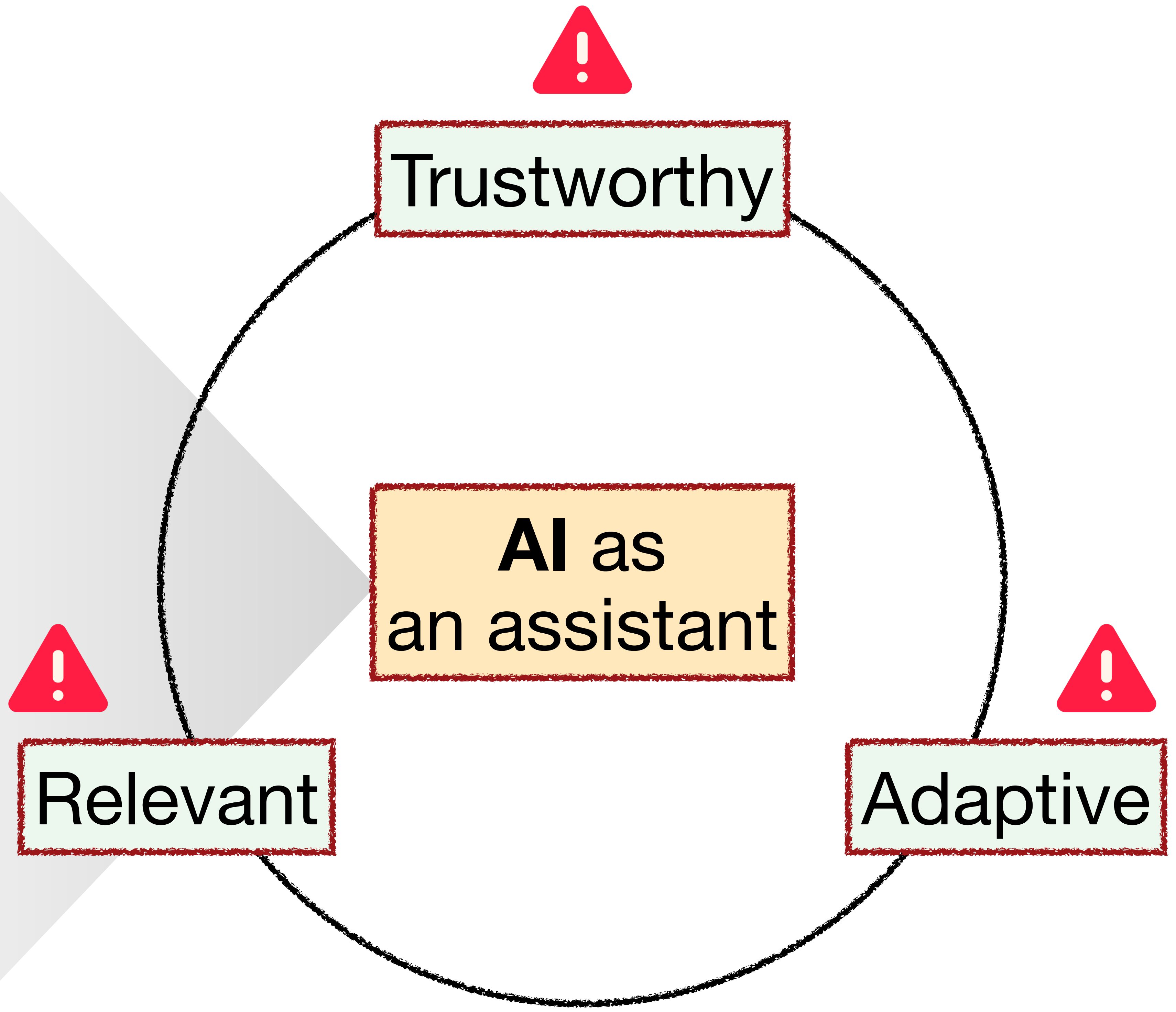


Behind the Scenes

Data
is temporal, biased, limited by its origin
e.g. pre- and post covid travel regulations
[Logan IV et al., 2022]

Model
can be opaque, contain spurious correlation
e.g. uses syntactic nuances instead of contextual knowledge for an NLI task
[Gardner et al., 2021]

Evaluation
can be done offline, may not address subjectivity
e.g. recommender systems are evaluated offline
no evaluation for new users (cold-start)
[McAuley et al., 2013]



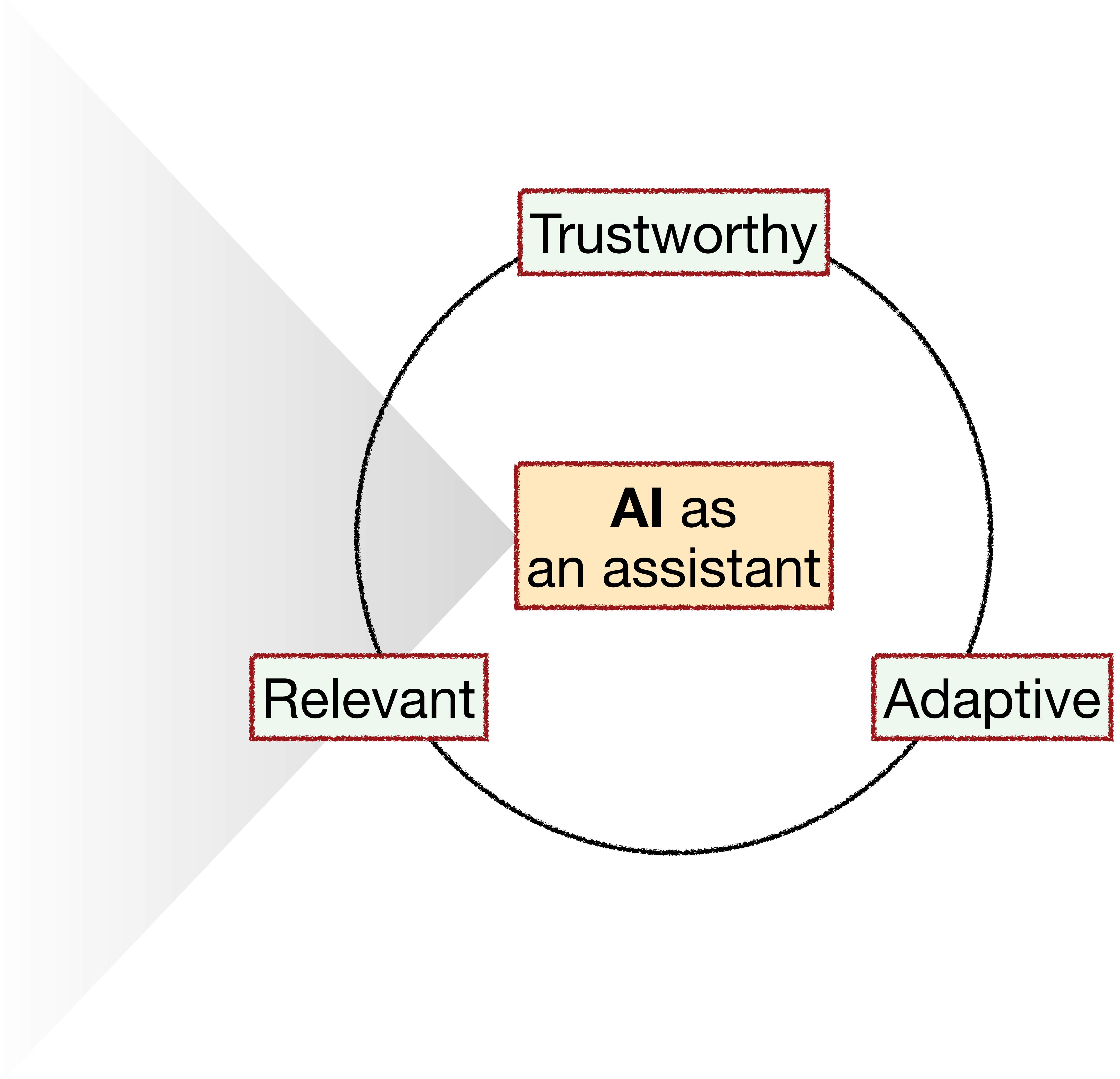
Way forward: Interactive Explainability

**Recognized by* Adobe Research Fellowship 2022, Qualcomm Innovation Fellowship 2020

Data

Model

Evaluation



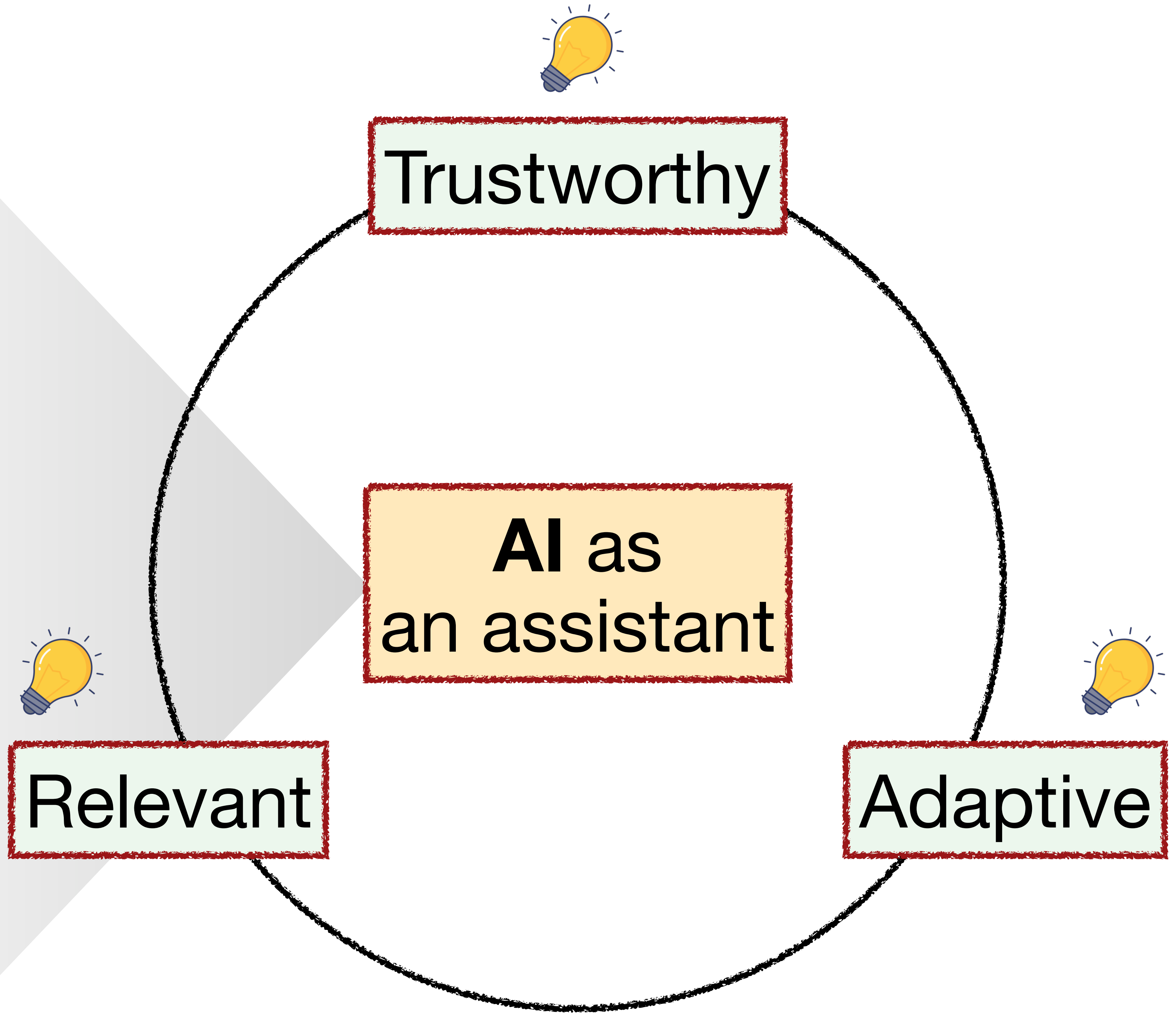
Way forward: Interactive Explainability

Recognized by Adobe Research Fellowship 2022, Qualcomm Innovation Fellowship 2020

Data + Knowledge
augment with explicit/implicit knowledge
e.g. fine-tuning or post-hoc injection
[Majumder et al., 2020; 2021; 2022]

Model + Explanations
to produce both predictions and explanations
e.g. extractive explanations as attributions or abstractive explanations as beliefs
[Majumder et al., 2022a; 2022b]

Evaluation + Interactions
human-in-the-loop learning, user studies
e.g. measuring success in achieving conversational goal instead of next response accuracy
[Majumder et al., 2022a; 2022b]



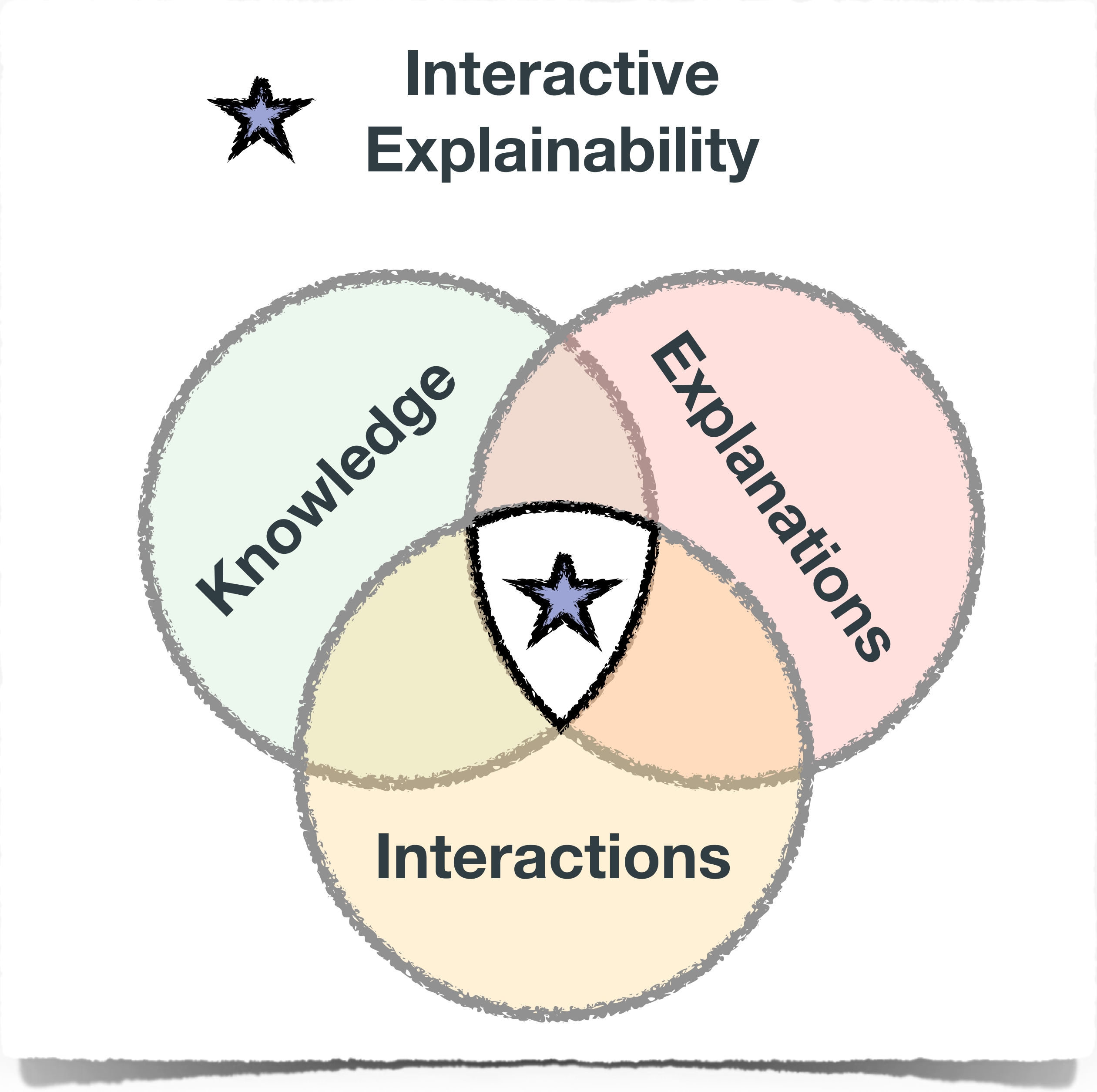
Way forward: Interactive Explainability

**Recognized by* Adobe Research Fellowship 2022, Qualcomm Innovation Fellowship 2020

Data + Knowledge
augment with explicit/implicit knowledge
e.g. fine-tuning or post-hoc injection
[Majumder et al., 2020; 2021; 2022]

Model + Explanations
to produce both predictions and explanations
e.g. extractive explanations as attributions or abstractive explanations as beliefs
[Majumder et al., 2022a; 2022b]

Evaluation + Interactions
human-in-the-loop learning, user studies
e.g. measuring success in achieving conversational goal instead of next response accuracy
[Majumder et al., 2022a; 2022b]



Next-generation AI
Current AI + Knowledge + Explanations + Interactions

Relevant, Trustworthy, and Adaptive AI

Knowledge	Explanations	Interactions
Persona-based Commonsense **Majumder** et al. **EMNLP** 2020 (Oral)	Knowledge Grounded Self-rationalization **Majumder** et al. **ICML** 2022 (Spotlight)	Estimating Missing Knowledge **Majumder** et al. **NAACL** 2021 (Oral)
Post-hoc Knowledge Injection **Majumder** et al. **ACL** 2021, **ACL** 2022 (Oral)	Controlling Bias Exposure via Rationales He, Yu, McAuley, **Majumder** **EMNLP** 2022	Conversational Recommendation Li, **Majumder** et al. **RecSys** 2022 (Highlights)
Personalized Knowledge Grounding **Majumder** et al. **EMNLP** 2019	Faithfulness in Language Explanations Xie, McAuley, **Majumder** **Preprint** 2022	Interactive Fair Debiasing **Majumder** et al. **InterNLP** 2022 (Oral)
Commonsense Grounding in Stories Mao, **Majumder** et al. **EMNLP** 2019	Factual Explanation Generation Xie, Singh, McAuley, **Majumder** **AAAI** 2023	Bernard: Human-centric NLP **Majumder** et al. **Alexa Proc.** 2021

Select publications

Next-generation AI

Current AI + Knowledge + Explanations + Interactions

Relevant, Trustworthy, and Adaptive AI

Prologue

Chapter I. Knowledge

Post-hoc Knowledge Injection to Make Models Relevant

Majumder et al.
ACL 2022
(12 mins)

Chapter II. Explanations

Role of Knowledge Grounding in Generating Explanations

Majumder et al.
ICML 2022
(12 mins)

Chapter III. Interactions

Improving Debiasing Performance with Natural Language Feedback

Majumder et al.
EMNLP & InterNLP 2022
(12 mins)

Epilogue

Hiuen Tsang's journey to the west

Next-generation AI

Current AI + 💡 Knowledge + 💡 Explanations + 💡 Interactions

Relevant, Trustworthy, and Adaptive AI

Chapter I. Knowledge

Post-hoc Knowledge Injection to Make Models Relevant

Majumder et al.
ACL 2022

Chapter II. Explanations

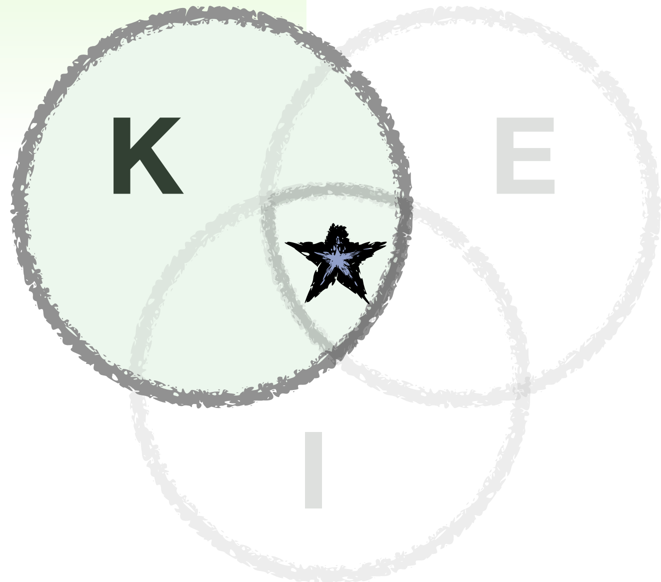
Role of Knowledge Grounding in Generating Explanations

Majumder et al.
ICML 2022

Chapter III. Interactions

Improving Debiasing Performance with Natural Language Feedback

Majumder et al.
EMNLP & InterNLP 2022



Next-generation AI

Current AI + Knowledge + Explanations + Interactions

Knowledge-seeking Dialog

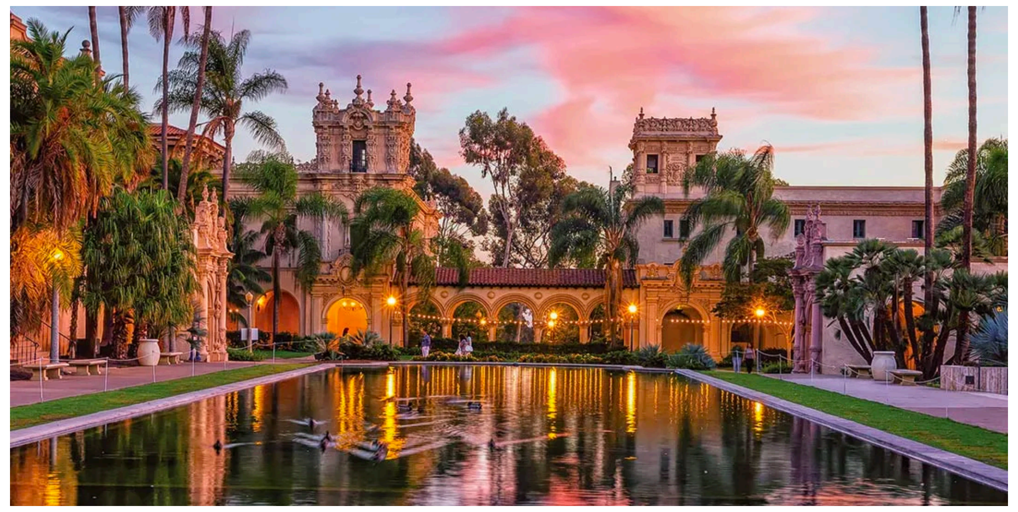
Find me something fun to do around San Diego area in daytime!

Dialog Context

You can go to **La Jolla Shores.**

You can go to **Balboa Park.**

Model trained in 2019



2019

2020

COVID19

2021

WORK IN PROGRESS

Knowledge-seeking Dialog

Find me something fun to do around San Diego area in daytime!

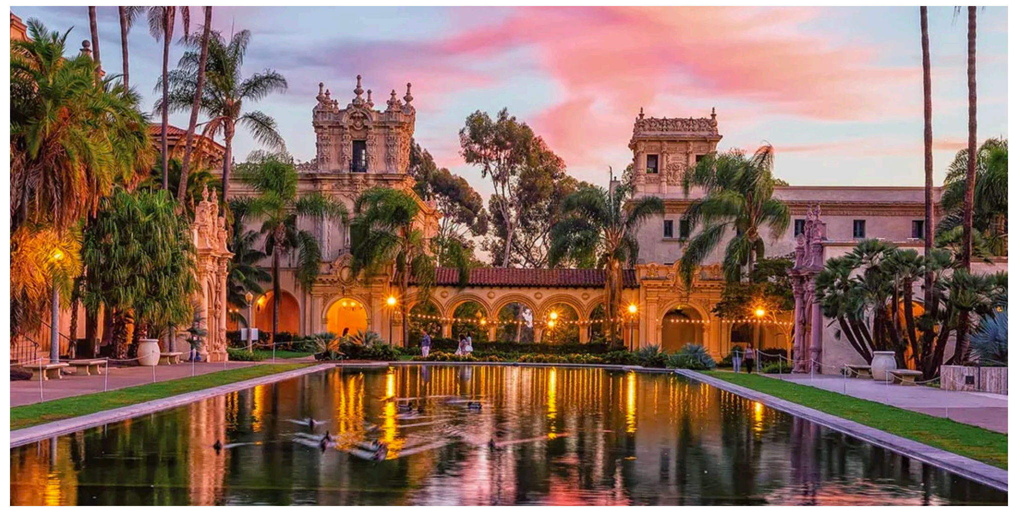
Dialog Context

Retrieved Knowledge yelp

You should go to La Jolla Shores.
It's so fun laying out with friends.

You can go to **La Jolla Shores.**

You can go to **Balboa Park.**



Model trained in 2019

2019

2020

COVID19

2021

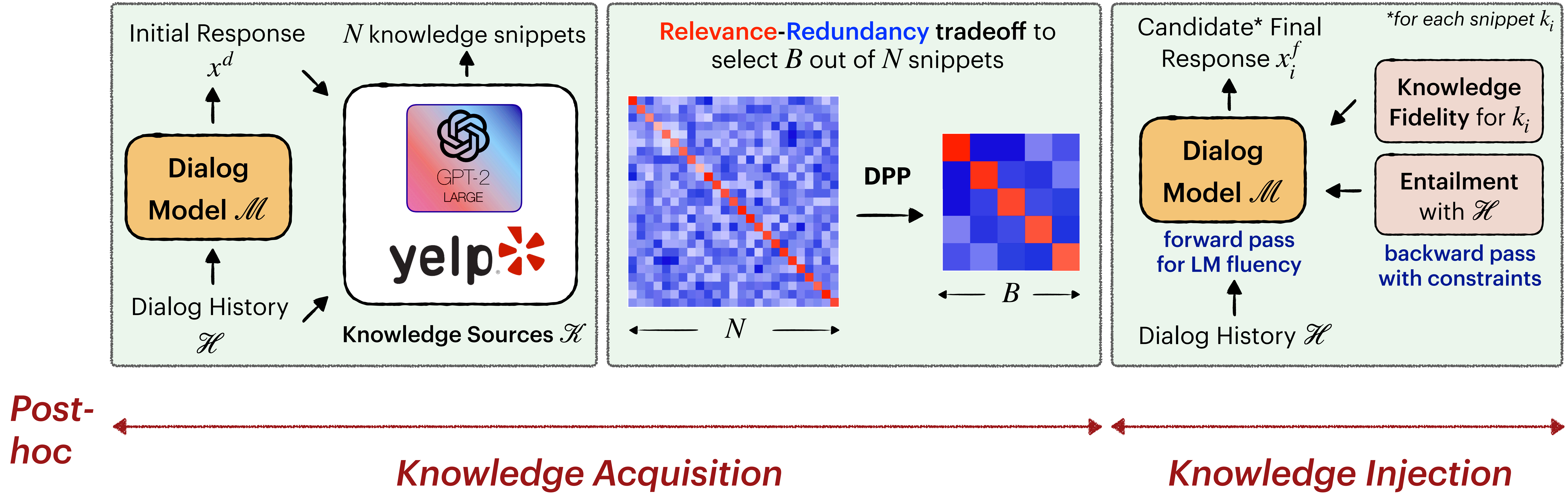
WORK IN PROGRESS

No access to relevant knowledge at initial training time

Knowledge Injection	Impact
Another (or more) round(s) of **fine-tuning**	Resource inefficient, Higher carbon footprint
Post-hoc, no additional training	Resource efficient, greener

Post-hoc Knowledge Injection in Generated Dialog

POKI



Post-hoc Knowledge Retrieval

Query

Find me something fun to do
in San Diego in the daytime!

non-parametric KB

Using cosine similarity on tf-idf representations

yelp

You should go to La Jolla Shores in San Diego in daytime. It has great size beaches, kayak rentals/tours, caves to explore, warm and semi clear water!

Post-hoc Knowledge Retrieval

Query

Find me something fun to do
in San Diego in the daytime!

non-parametric KB

Using cosine similarity on tf-idf representations

yelp

You should go to La Jolla Shores in San Diego in daytime. It has great size beaches, kayak rentals/tours, caves to explore, warm and semi clear water!

parametric KB

Prompting an LM with keywords from dialog history and initial response:

GPT3

Find me something fun to do
in San Diego in the daytime!

Find me something fun to do in San Diego in the daytime are

visiting Balboa Park or taking a walk along the waterfront.

Unsupervised Knowledge Selection

1 You should go to La Jolla Shores in San Diego in daytime. It has great size beaches, ...

2 San Diego has great beaches with awesome views.

3 In San Diego beaches, you can just enjoy wetting your feet, taking a swim, or ...

⋮

N Maritime Museum of San Diego is a great place to spend a day. It has mighty ships and great tours..

Relevance: PMI **(knowledge** i **, history)**

*PMI probabilities are calculated using an LM (e.g. GPT2)

Unsupervised Knowledge Selection

1 You should go to La Jolla Shores in San Diego in daytime. It has great size beaches, …

2 San Diego has great beaches with awesome views.

3 In San Diego beaches, you can just enjoy wetting your feet, taking a swim, or …

⋮

N Maritime Museum of San Diego is a great place to spend a day. It has mighty ships and great tours..

Relevance: PMI **(knowledge i, history)**

Redundancy: PMI **(knowledge i, knowledge j)**

*PMI probabilities are calculated using an LM (e.g. GPT2)

[Padmakumar and He, 2021]

Unsupervised Knowledge Selection

1 You should go to La Jolla Shores in San Diego in daytime. It has great size beaches, …

2 San Diego has great beaches with awesome views.

3 In San Diego beaches, you can just enjoy wetting your feet, taking a swim, or …

⋮

N Maritime Museum of San Diego is a great place to spend a day. It has mighty ships and great tours..

Relevance: PMI **(knowledge** i **, history)**

Redundancy: PMI **(knowledge** i **, knowledge** j**)**

*PMI probabilities are calculated using an LM (e.g. GPT2)

[Padmakumar and He, 2021]

Determinantal Poison Process (DPP):

sampling the most relevant and the most diverse subset

[Kulesza and Taskar, 2011]

Unsupervised Knowledge Selection

1 You should go to La Jolla Shores in San Diego in daytime. It has great size beaches, …

San Diego has great beaches with awesome views.

In San Diego beaches, you can just enjoy wetting your feet, taking a swim, or …

⋮

B Maritime Museum of San Diego is a great place to spend a day. It has mighty ships and great tours..

Relevance: PMI **(knowledge i , history)**

Redundancy: PMI **(knowledge i , knowledge j)**

*PMI probabilities are calculated using an LM (e.g. GPT2)

[Padmakumar and He, 2021]

Determinantal Poison Process (DPP):

sampling the most relevant and the most diverse subset

[Kulesza and Taskar, 2011]

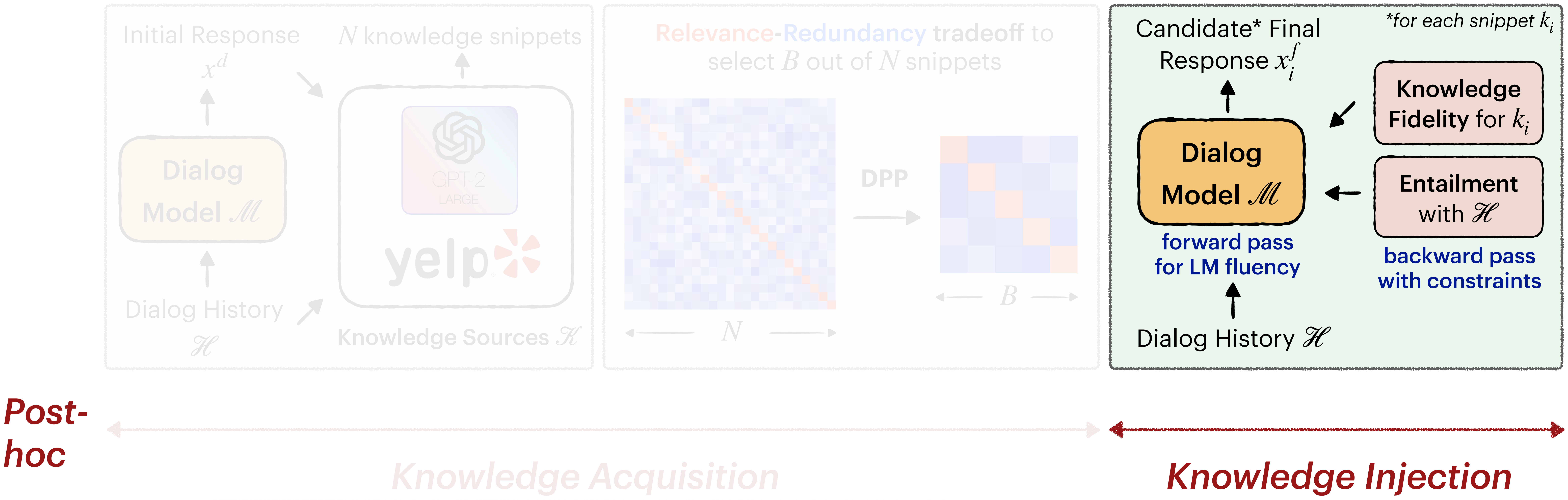
Greedy trade-off:

Select most relevant knowledge snippet

— *Select* the next knowledge snippet that maximizes the diversity

Post-hoc Knowledge Injection in Generated Dialog

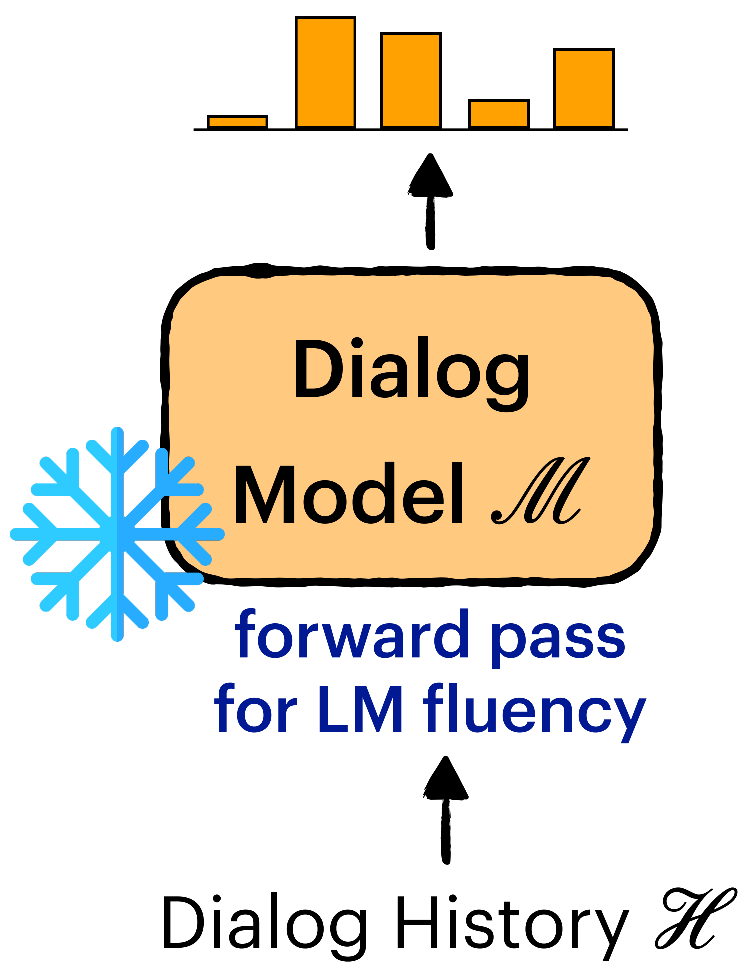
POKI



Collected *B* relevant and diverse knowledge snippets

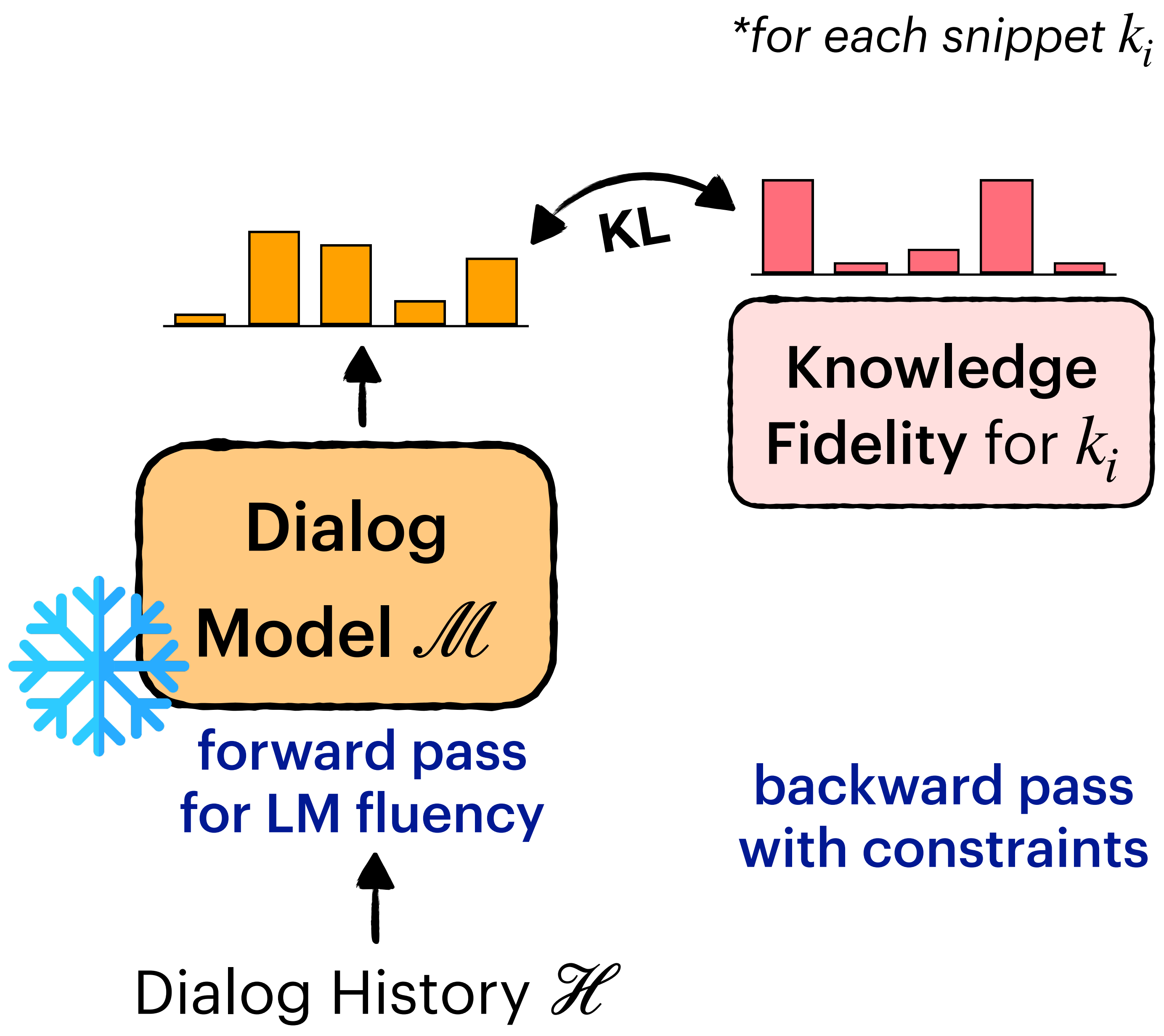
Post-hoc Knowledge Injection

Forward pass for dialog model fluency



Constrained Decoding

Post-hoc Knowledge Injection



Forward pass for dialog model fluency

Backward pass to ensure

1. modified response is as close to as the knowledge snippet — ***fidelity***

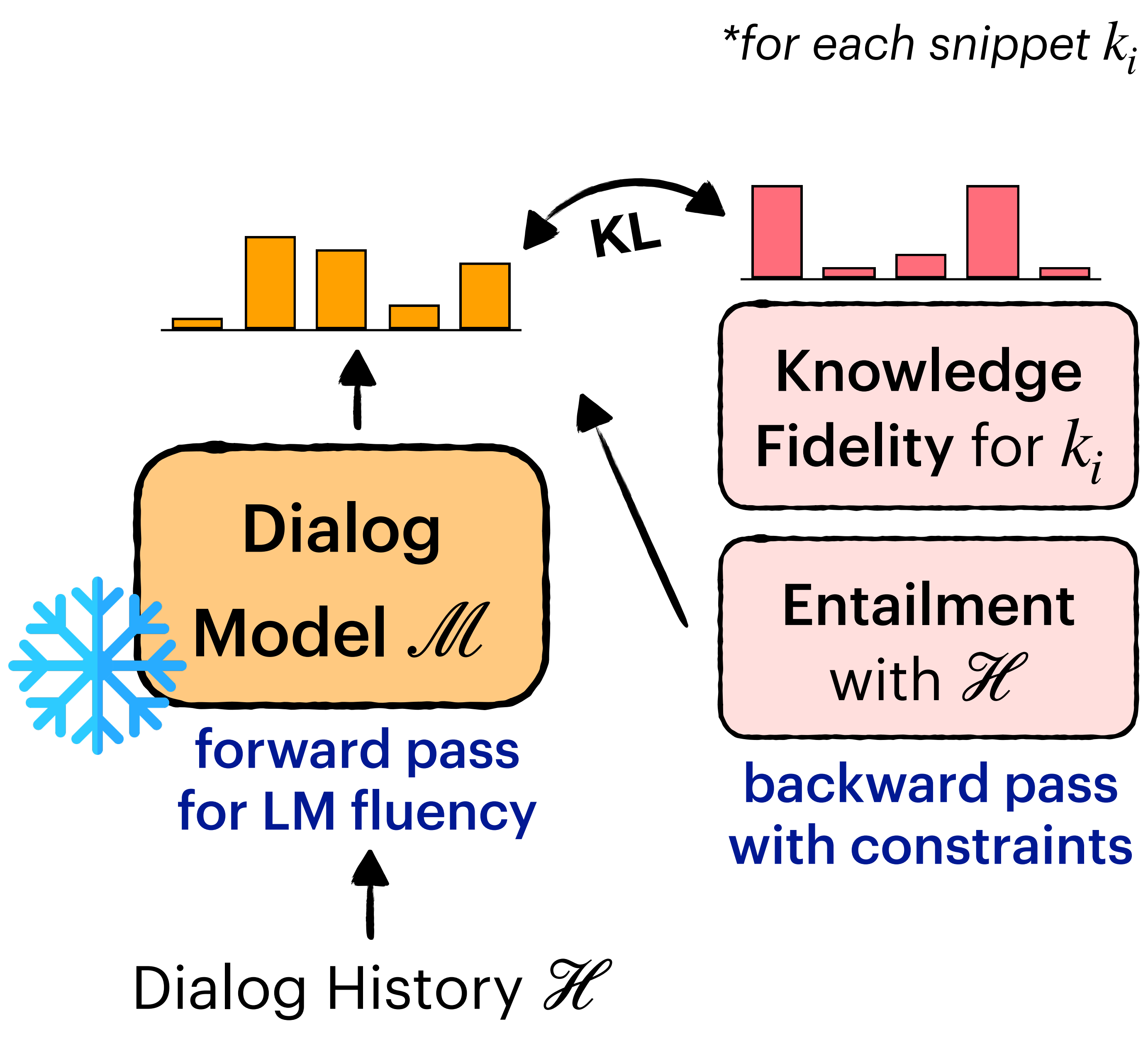
Constrained Decoding

You can go to Balboa Park.

+

You should go to La Jolla Shores in San Diego in daytime. It has great size beaches, kayak rentals/tours, caves to explore, warm and semi clear water!

Post-hoc Knowledge Injection



Constrained Decoding

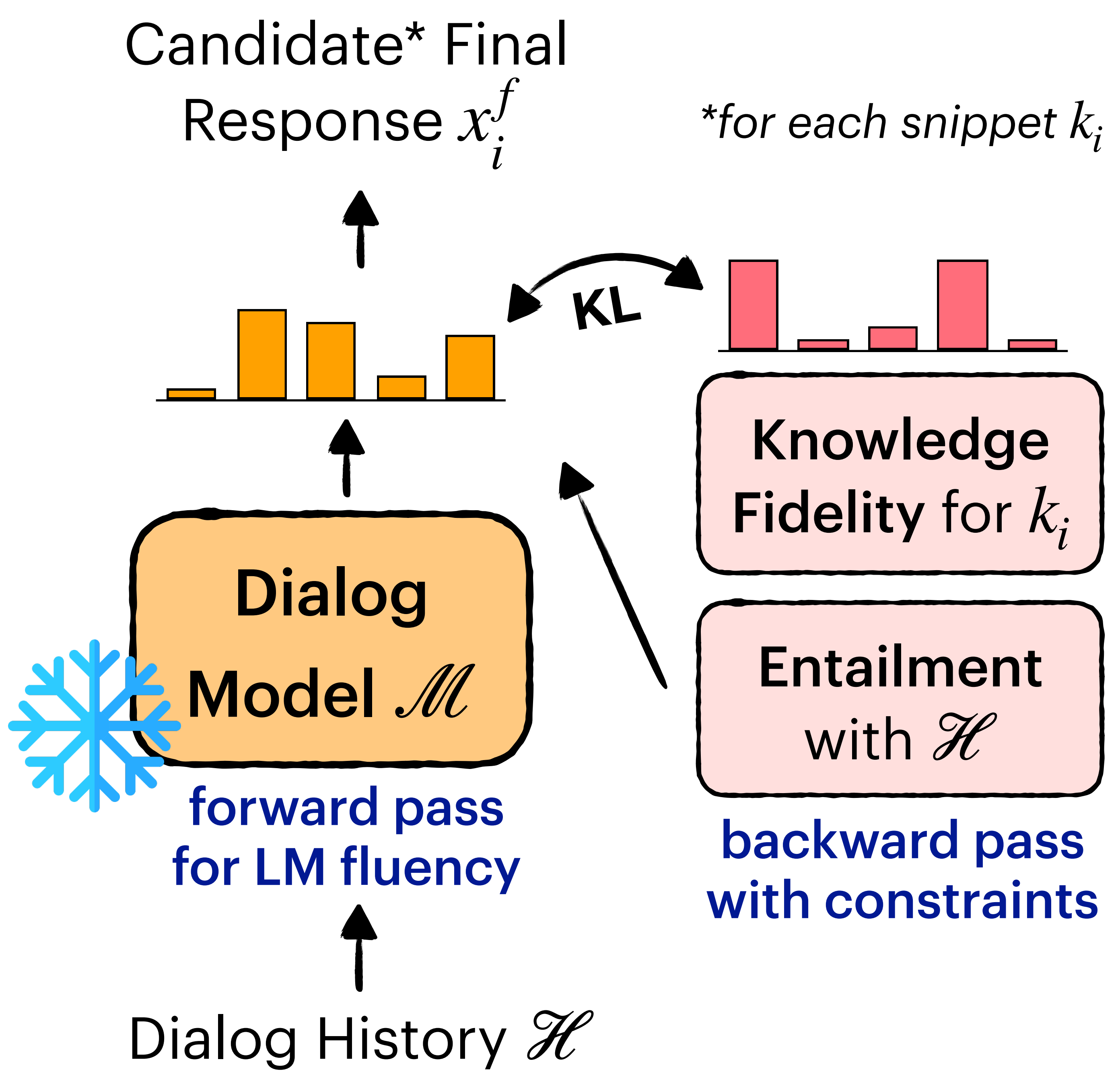
Forward pass for dialog model fluency

Backward pass to ensure

1. modified response is as close to as the knowledge snippet — ***fidelity***
2. modified response still entails with dialog history — ***entailment***

You can go to Balboa Park. **+** You should go to La Jolla Shores in San Diego in daytime. It has great size beaches, kayak rentals/tours, caves to explore, warm and semi clear water!

Post-hoc Knowledge Injection



Forward pass for dialog model fluency

Backward pass to ensure

1. modified response is as close to as the knowledge snippet — ***fidelity***
2. modified response still entails with dialog history — ***entailment***

[Dathathri et al., 2020; Qin et al., 2020]

Constrained Decoding

You can go to Balboa Park. + You should go to La Jolla Shores in San Diego in daytime. It has great size beaches, kayak rentals/tours, caves to explore, warm and semi clear water! =

After few iterations

You should go to La Jolla Shores. It has great beaches ...

select one via ranking

User Study for Effectiveness

Does post-hoc knowledge-injection promote conversational success?

Goal: Reach final goal (e.g. booking a restaurant) as soon as possible

User Study for Effectiveness

Does post-hoc knowledge-injection promote conversational success?

Goal: Reach final goal (e.g. booking a restaurant) as soon as possible

👧🏽 : Find me some inexpensive restaurants that serve English food around the Center of Cambridge?

baseline

🤖: Most English restaurants in the Center of Cambridge are expensive.

ours (POKI)

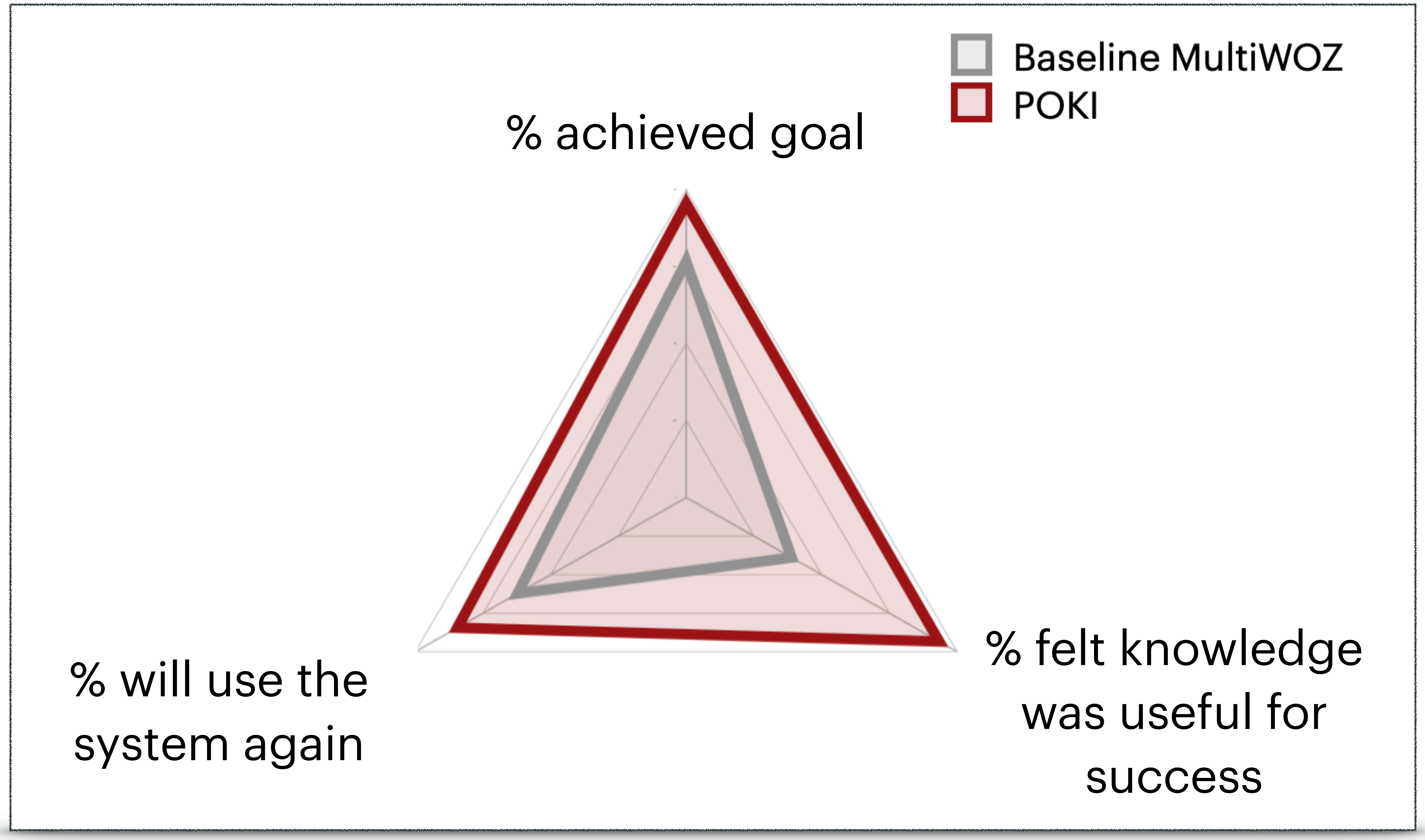
🤖: There are very few options for inexpensive English restaurants in the Center. However, Indian chains in the center area are affordable. Many people who like English food also enjoy Indian food.

yelp

User Study for Effectiveness

Does post-hoc knowledge-injection promote conversational success?

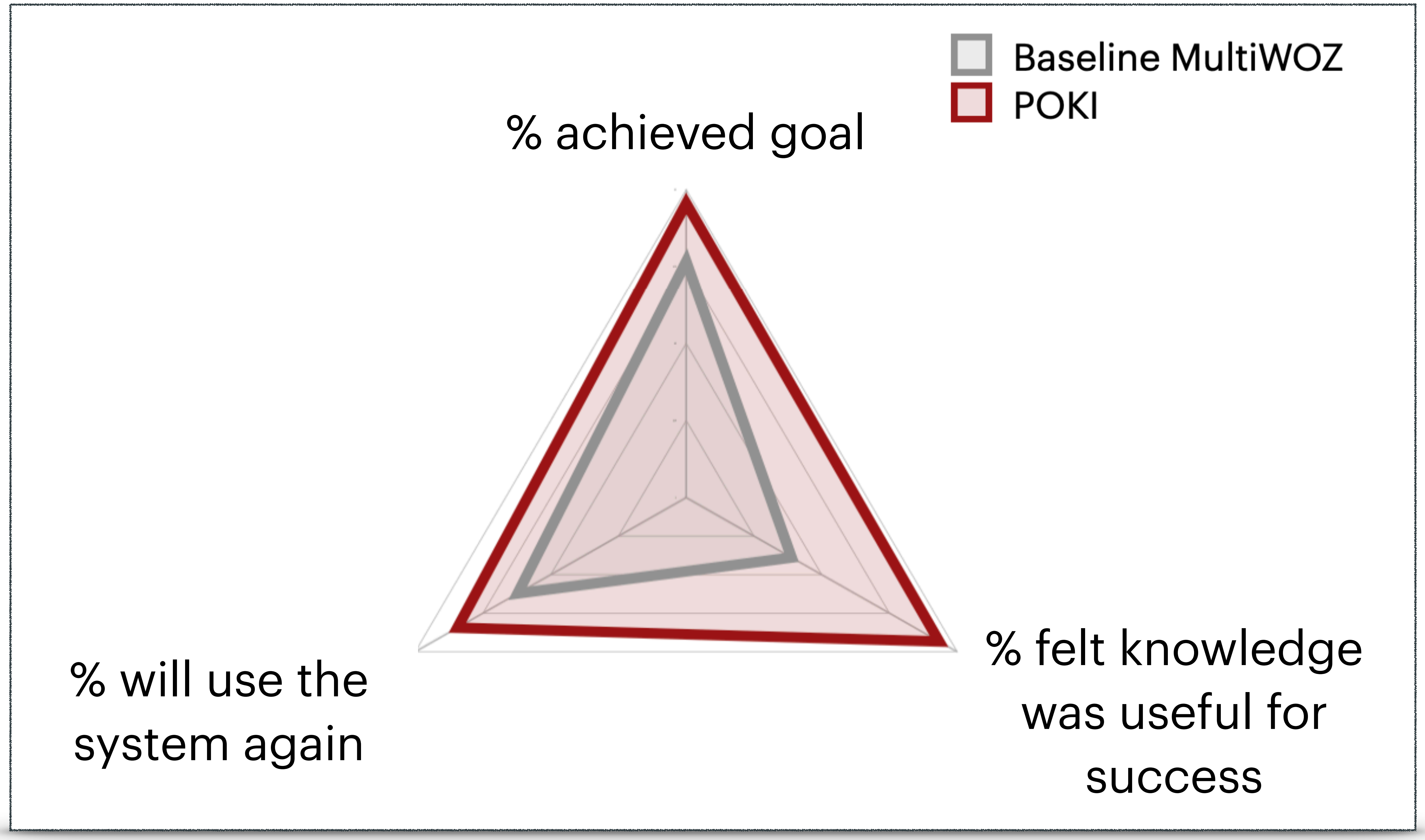
Goal: Reach final goal (e.g. booking a restaurant) as soon as possible

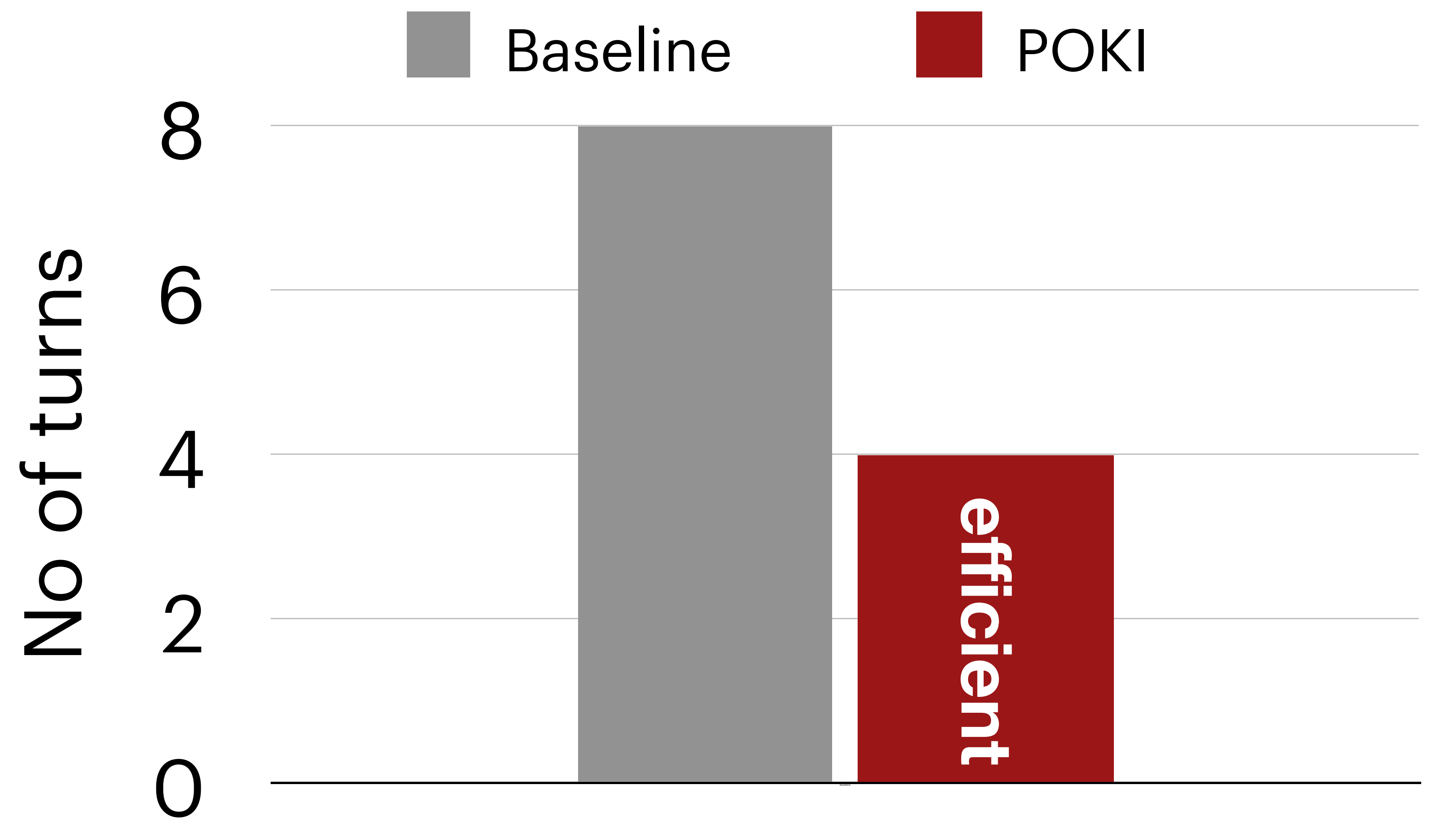


User Study for Effectiveness

Does post-hoc knowledge-injection promote conversational success?

Goal: Reach final goal (e.g. booking a restaurant) as soon as possible

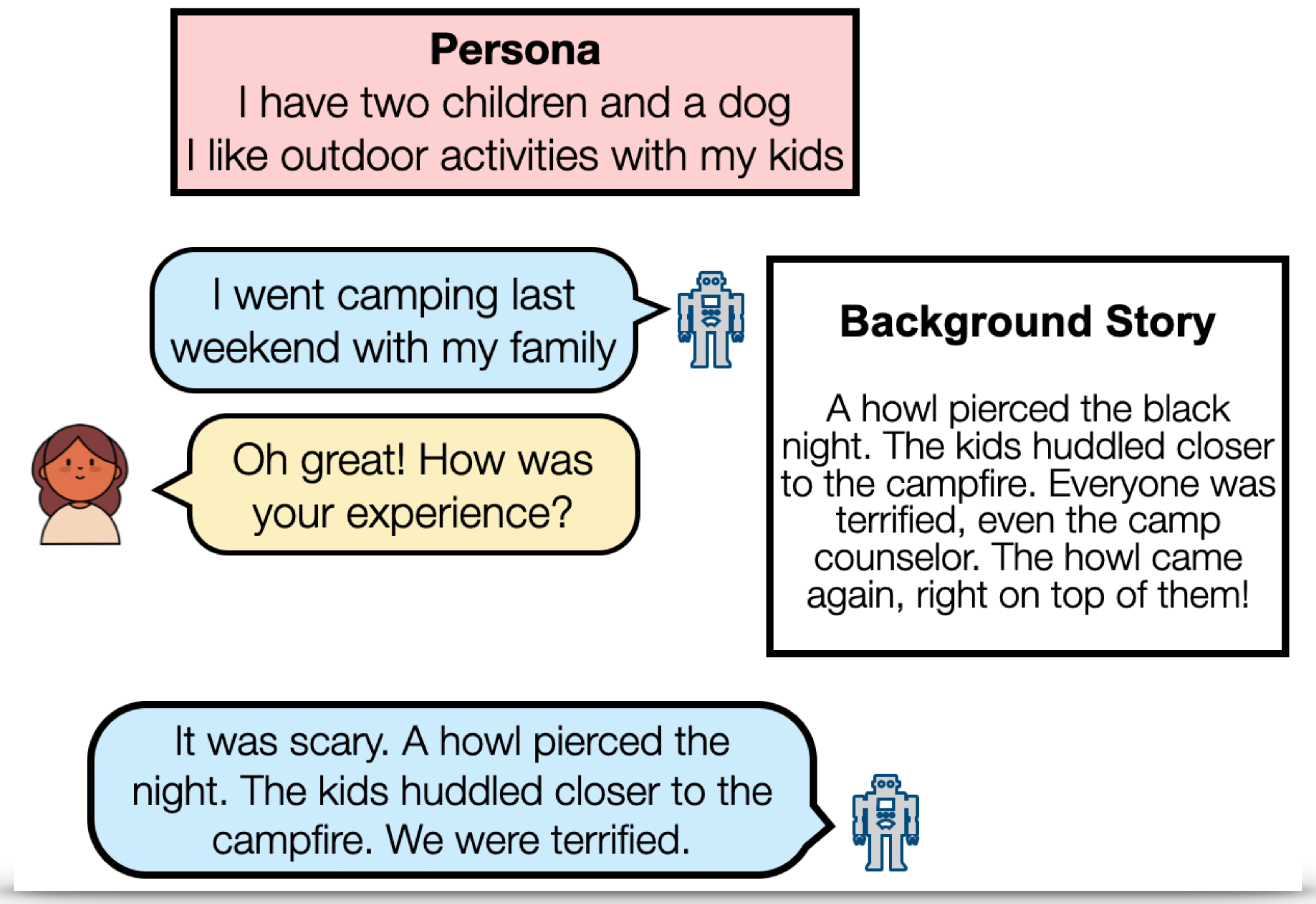




Additional knowledge improves user efficiency

Injecting Other Types of Knowledge

Post-hoc Knowledge Injection
Majumder et al.
ACL 2021



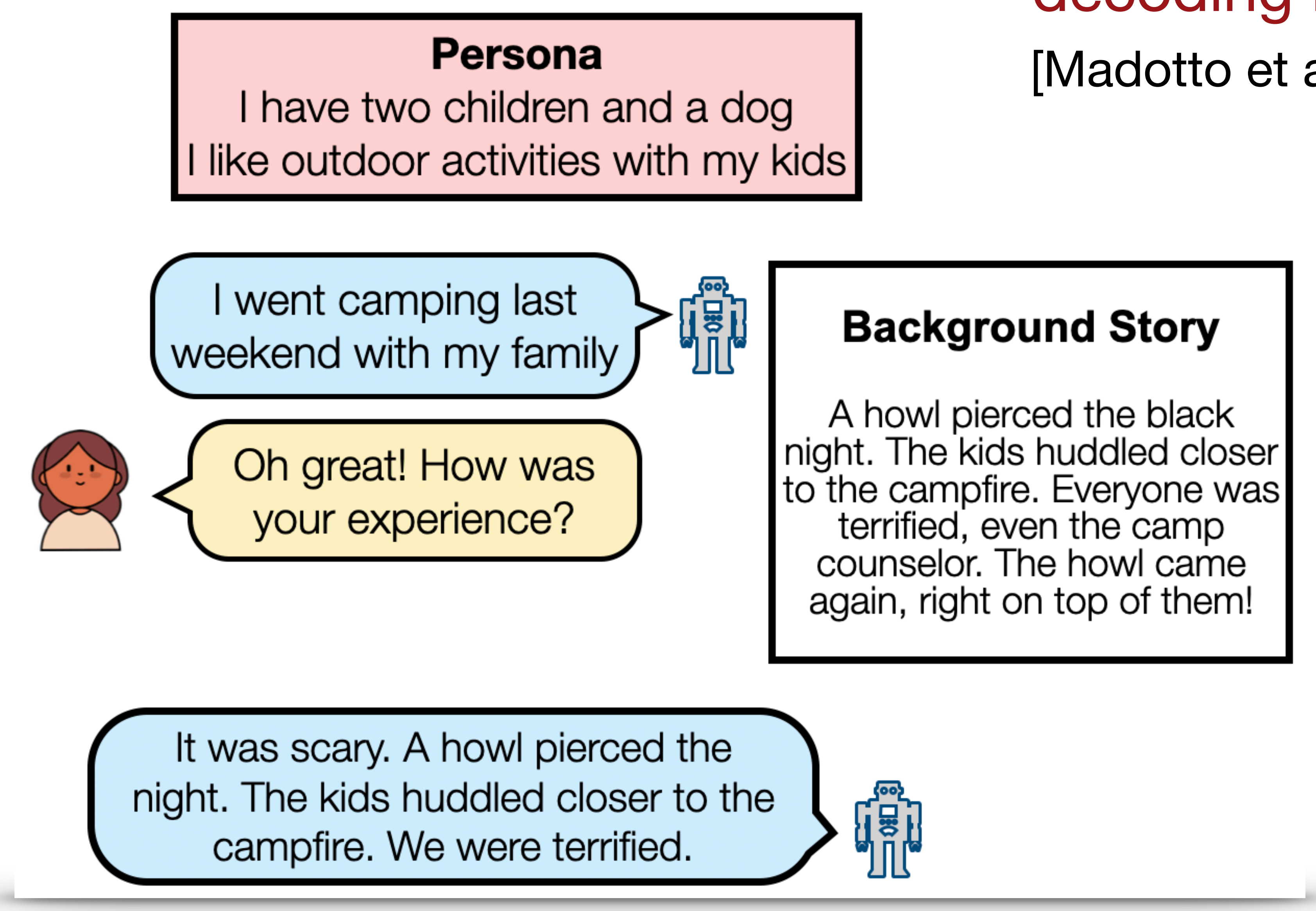
Narratives, post-hoc

Injecting Other Types of Knowledge

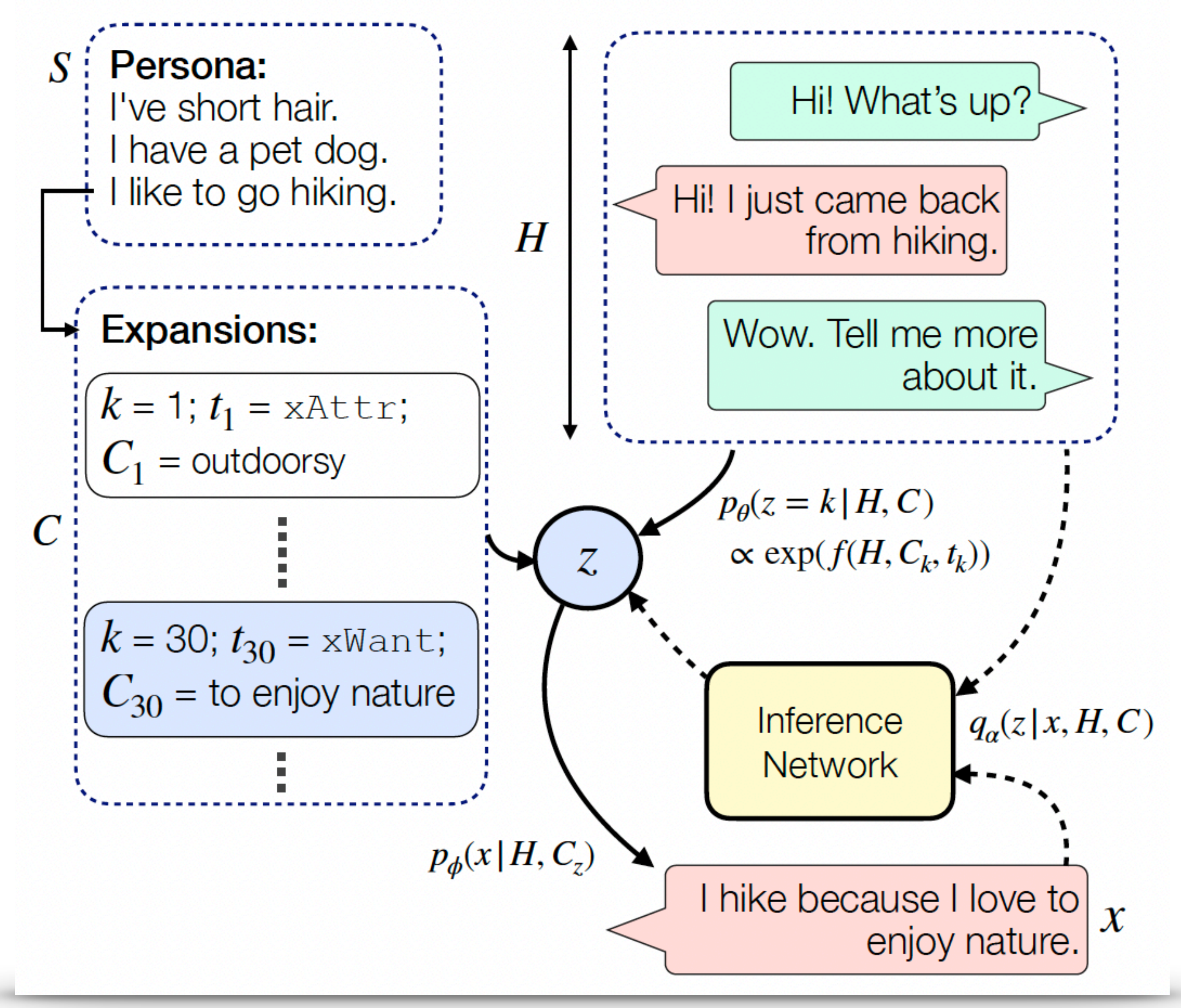
Post-hoc Knowledge Injection
Majumder et al.
ACL 2021

Persona-based Commonsense
Majumder et al.
EMNLP 2020

Gradient-based decoding is expensive
[Madotto et al., 2020]



Narratives, post-hoc



Commonsense Inference Graphs, training-time

Summary: Knowledge Acquisition + Injection

Chapter I.
Knowledge

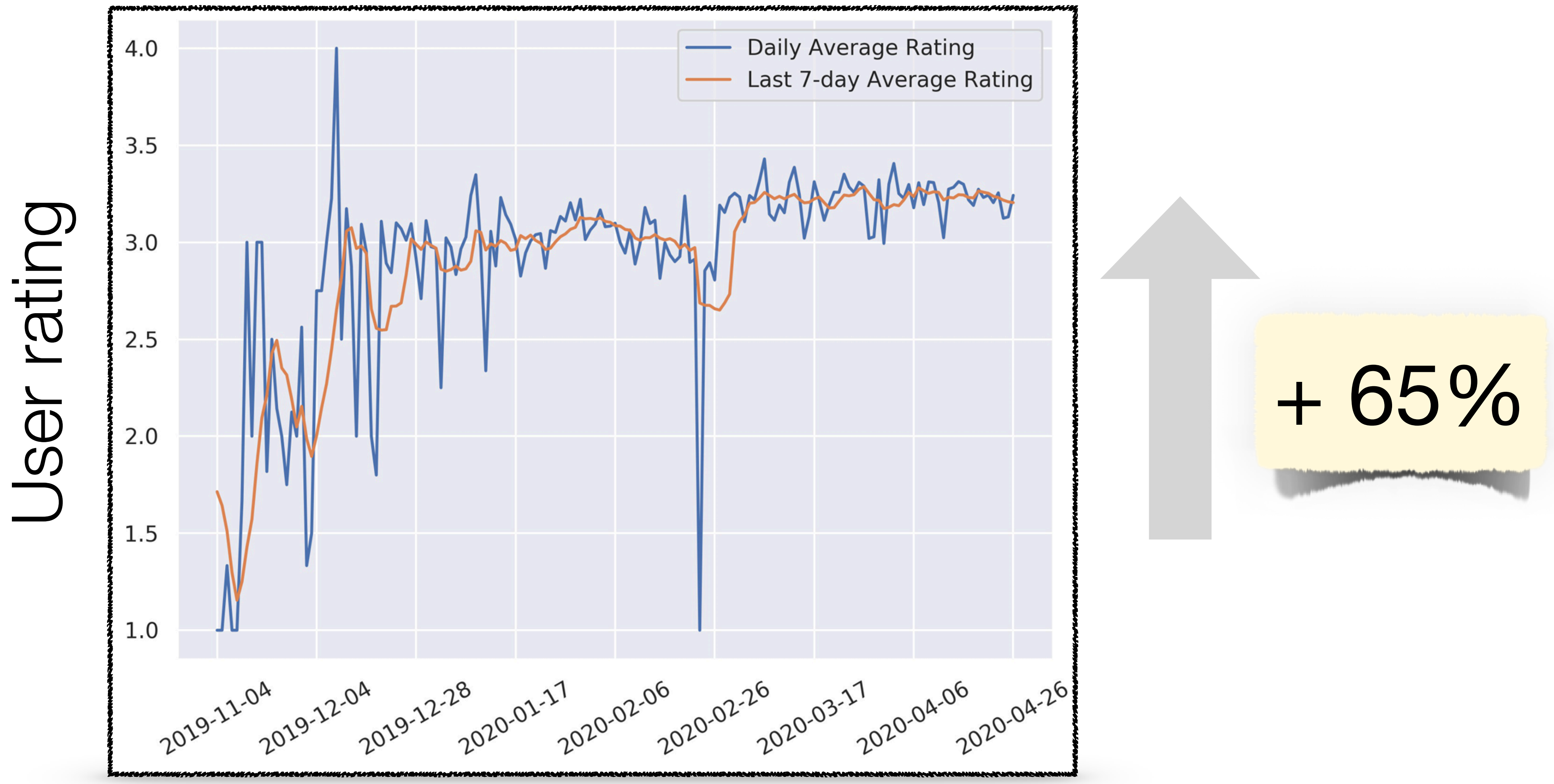
Post-hoc Knowledge Injection to Make Models Relevant

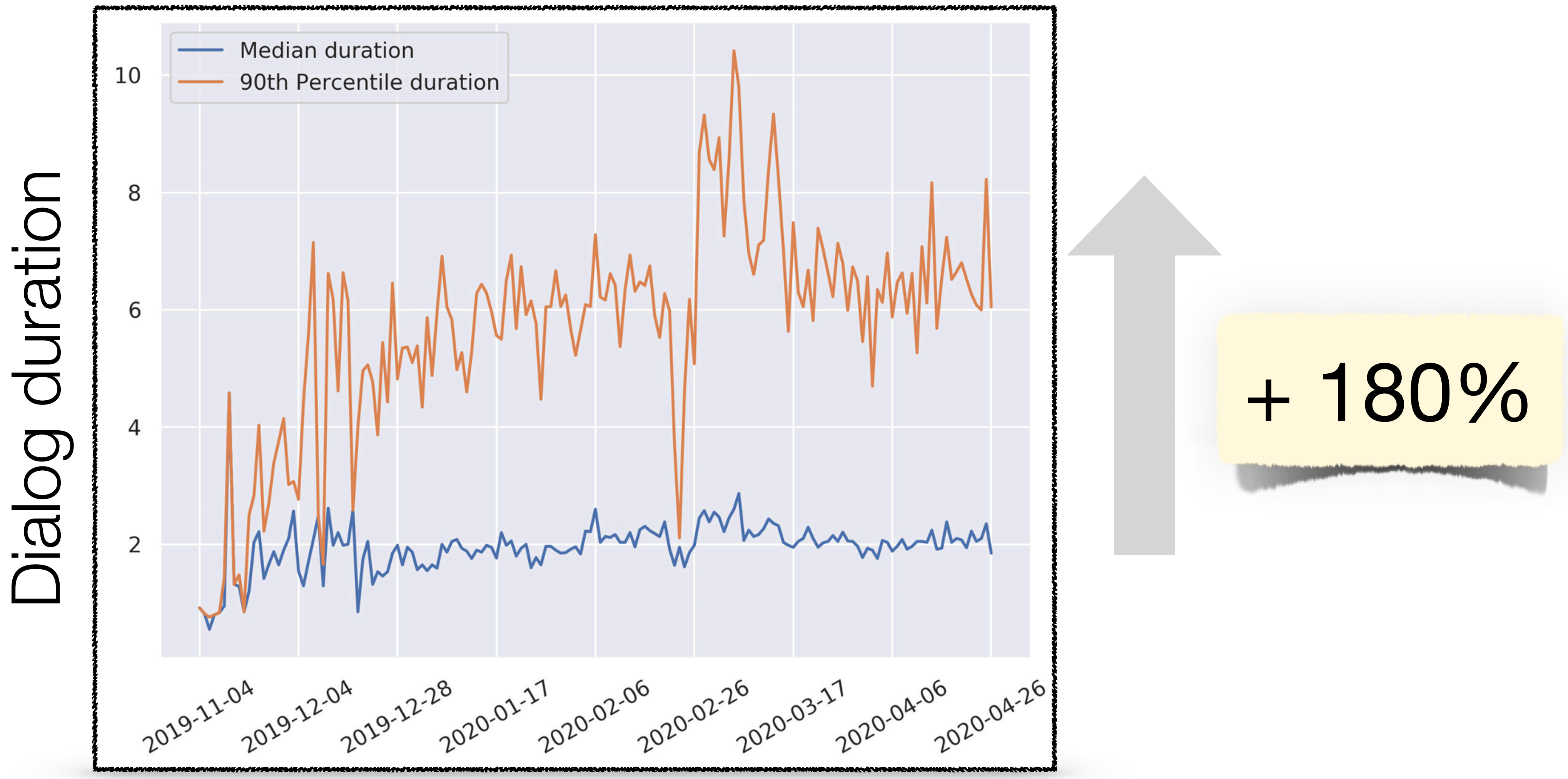
Majumder et al.
ACL 2022

- **On the fly** knowledge acquisition
  - Textual knowledge
  - Narratives
  - Structured commonsense
- **Ante-** and **post-hoc** methods
- Promotes **success in** achieving **conversational goals**
- **Bridges** the **knowledge gap** in existing dialog/language models

Impact: Dialog at Scale (~M)

Up-to-date, Knowledge-aware





Relevant, Trustworthy, and Adaptive AI

Chapter I. Knowledge

Post-hoc Knowledge Injection to Make Models Relevant

Majumder et al.
ACL 2022

Chapter II. Explanations

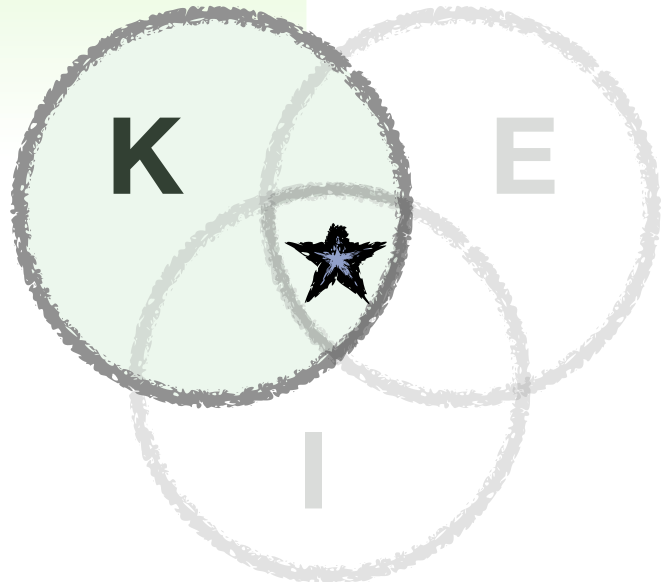
Role of Knowledge Grounding in Generating Explanations

Majumder et al.
ICML 2022

Chapter III. Interactions

Improving Debiasing Performance with Natural Language Feedback

Majumder et al.
EMNLP & InterNLP 2022



Next-generation AI

Current AI + Knowledge + Explanations + Interactions

Relevant, Trustworthy, and Adaptive AI

Chapter I.
Knowledge

Post-hoc Knowledge Injection to Make Models Relevant

Majumder et al.
ACL 2022

Chapter II. Explanations

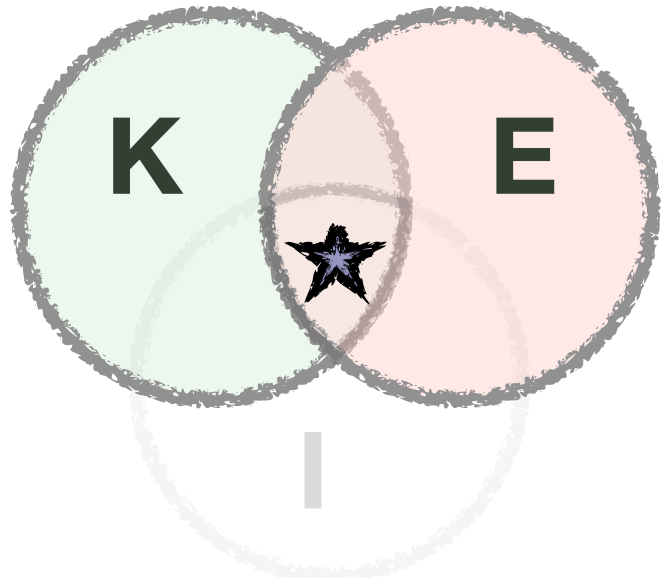
Role of Knowledge Grounding in Generating Explanations

Majumder et al.
ICML 2022

Chapter III.
Interactions

Improving Debiasing Performance with Natural Language Feedback

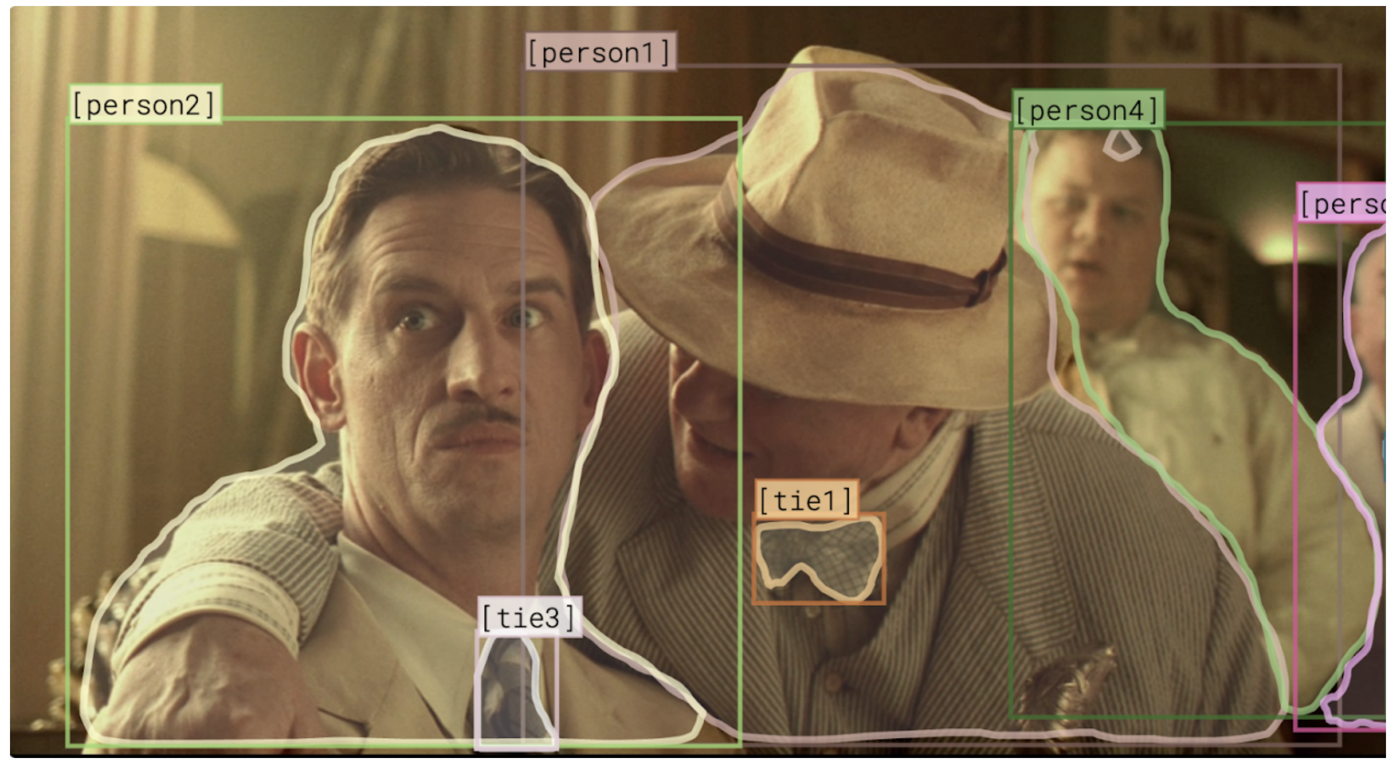
Majumder et al.
EMNLP & InterNLP 2022



Next-generation AI

Current AI + Knowledge + **Explanations** + Interactions

Natural Language Explanations (NLEs)

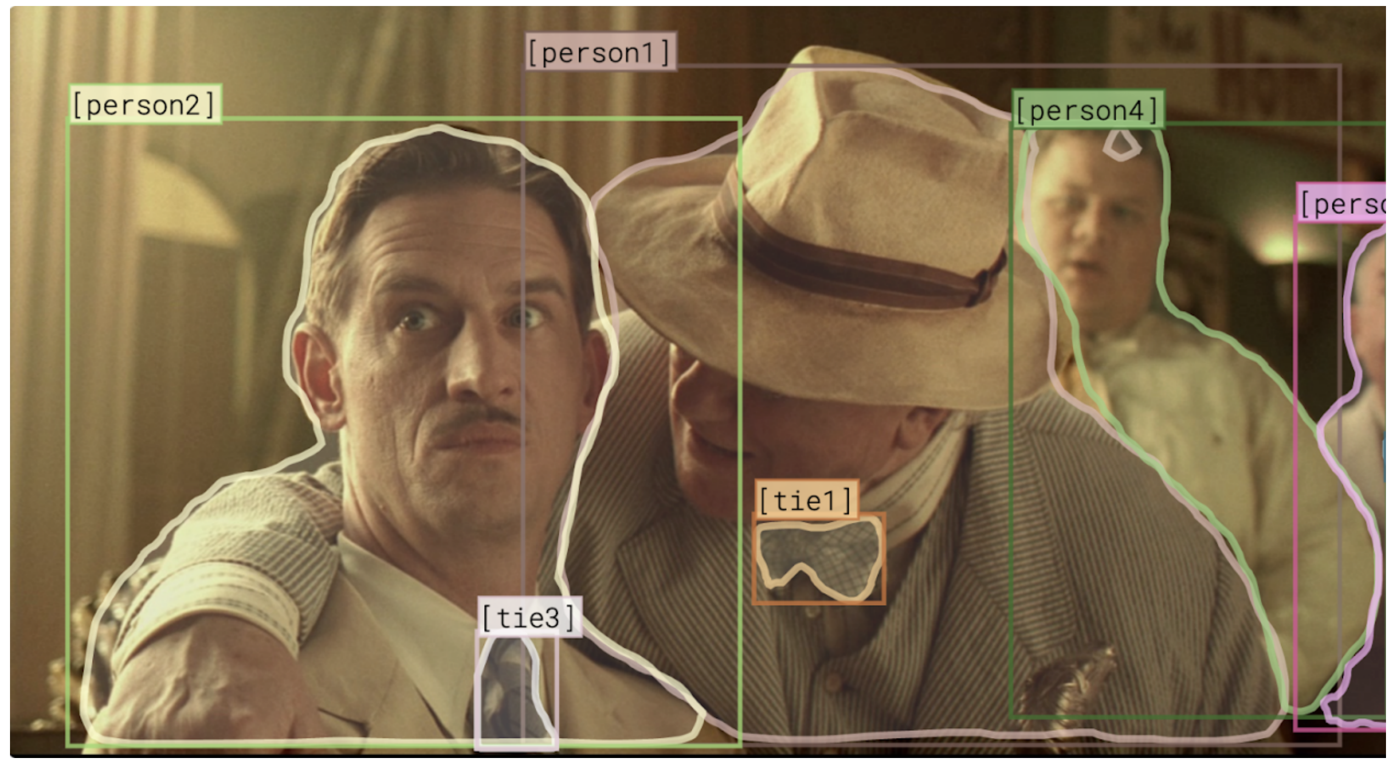


Q: how does `[person2]` feel about what `[person1]` is telling him?

A: He's concerned and a little upset

He is in shock thinking something bad is about to happen.

Natural Language Explanations (NLEs)



Q: how does `[person2]` feel about what `[person1]` is telling him?

A: He's concerned and a little upset

He is in shock thinking something bad is about to happen.

- NLE should be **plausible** and consistent to the input
  [Marasovic´ et al., 2021]
- NLE should be **accurate** and **faithful** to explain the prediction
  [Wiegreffe et al., 2021]
- NLE should be grounded into **world knowledge**
  [Camburu et al., 2020]

Walkthrough Example

A neural predictive model is employed to solve task.

For example: **Natural Language Inference (NLI)**

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

label
`entailment`

Instance from SNLI dataset

Natural Language Explanations

An NLE is a **textual abstraction** of the model explanation.

[Camburu et al., 2018]

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

label
`entailment`

Competing in a bicycle race requires people riding bikes

Background Knowledge

A model **believes** in a set of background knowledge given input.

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

label
`entailment`

- bicycle race requires bikes
- race requires riding bikes
- bicycle race needs helmet
- men are people

Competing in a bicycle race requires people riding bikes

Background Knowledge

A model **believes** in a set of background knowledge given input.

Where do we get this knowledge?

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

label
`entailment`

- bicycle race requires bikes
- race requires riding bikes
- bicycle race needs helmet
- men are people

Competing in a bicycle race requires people riding bikes

From the predictive parts of the input

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

bicycle race
riding bikes

men People

- bicycle race requires bikes
- race requires riding bikes
- bicycle race needs helmet
- men are people

label
entailment

Competing in a bicycle race requires people riding bikes

Rationale-induced Knowledge

A rationale is a sufficient and minimal part* of the **input** that is a **significant indicator** of a model's prediction.

[Lei et al., 2016; Bastings et al., 2019]

From the predictive parts of the input

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

label

entailment

bicycle race
riding bikes
men People

- bicycle race requires bikes
- race requires riding bikes
- bicycle race needs helmet
- men are people

Competing in a bicycle race requires people riding bikes

*tokens for language or super-pixels for images

Self-rationalization + Knowledge Grounding

Jointly producing prediction + explanation

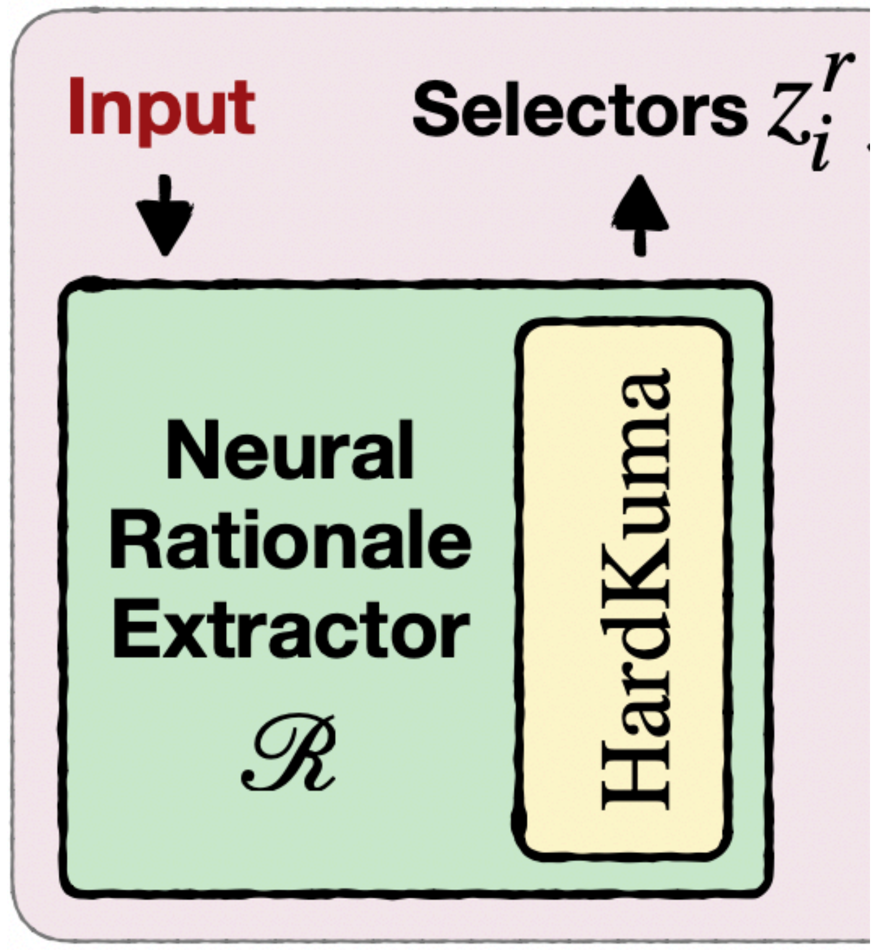
	Knowledge Grounding	Joint prediction + explanation
[Camburu et al., 2018]	👎	👎
[Kumar et al., 2018]	👎	👍
[Marasovic´ et al., 2018]	👍	👎
[Narang et al., 2020]	👎	👍
RExC	👍	👍

→ is the first to connect **R**ationales and **Ex**planations with Knowledge (**C**ommonsense) in an end-to-end fashion

Rationale

P: Two men are competing in a bicycle race

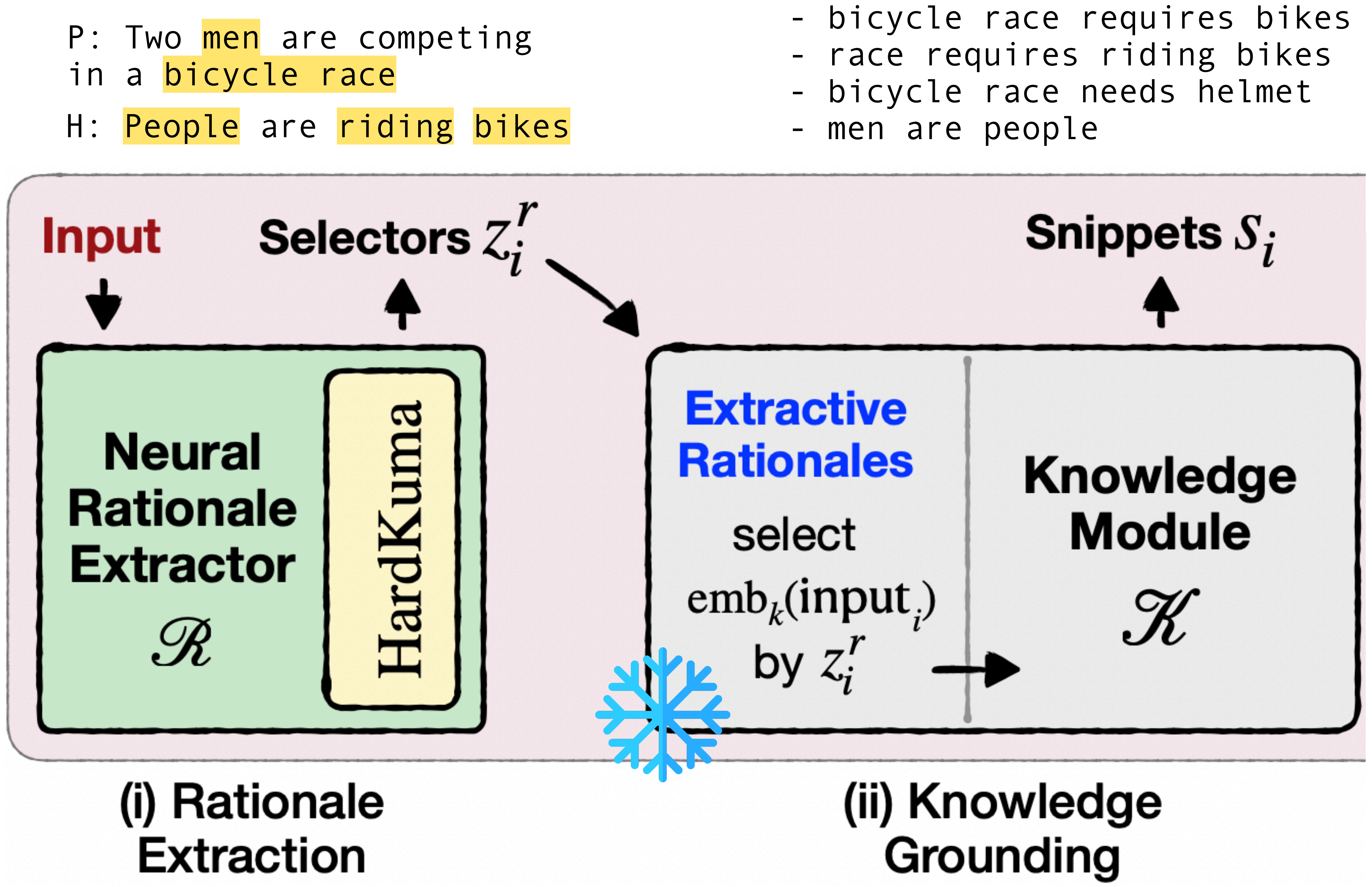
H: People are riding bikes



(i) Rationale Extraction

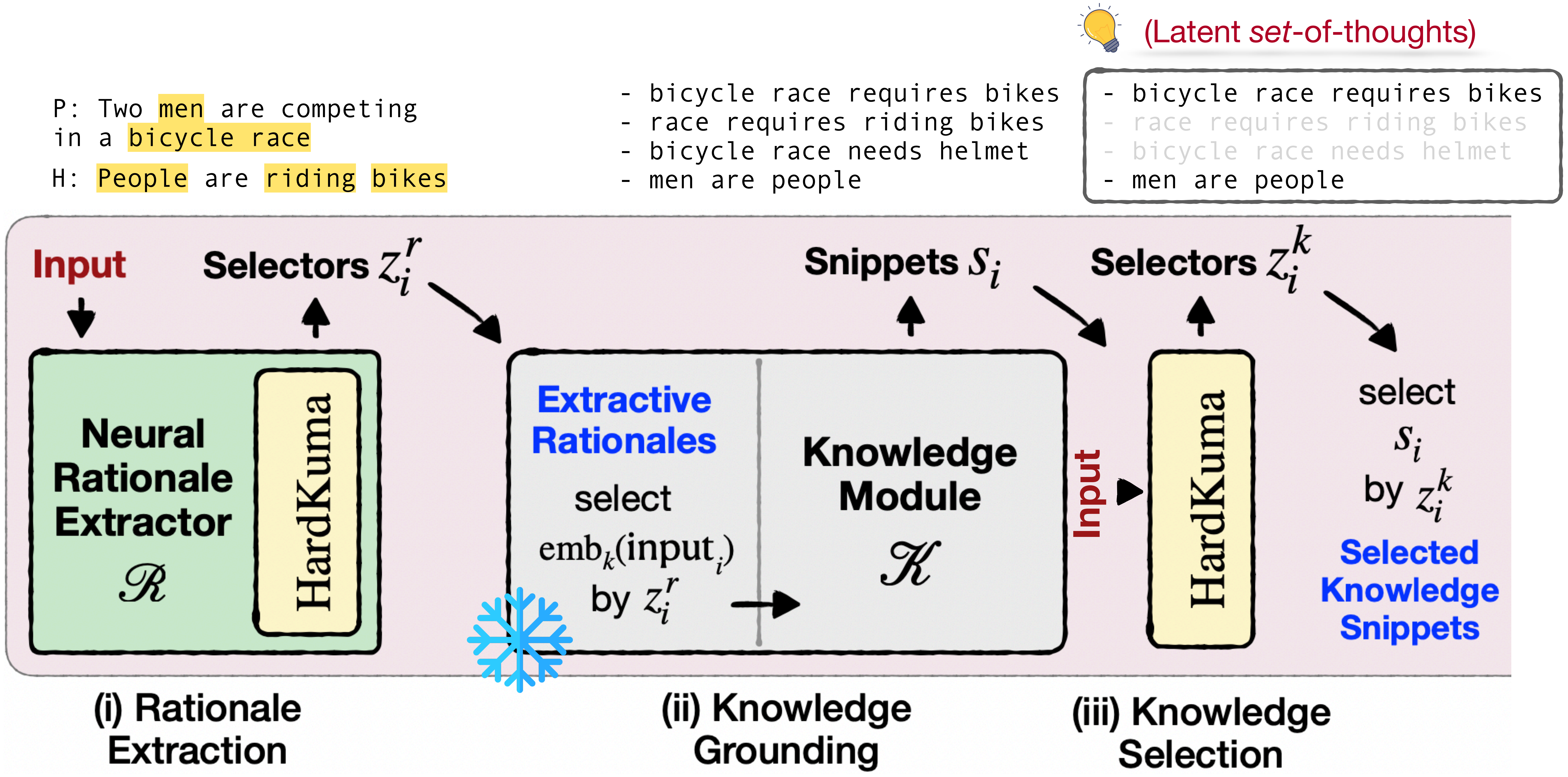
Rationales are responsible for relevant knowledge retrieval

Rationale + Knowledge



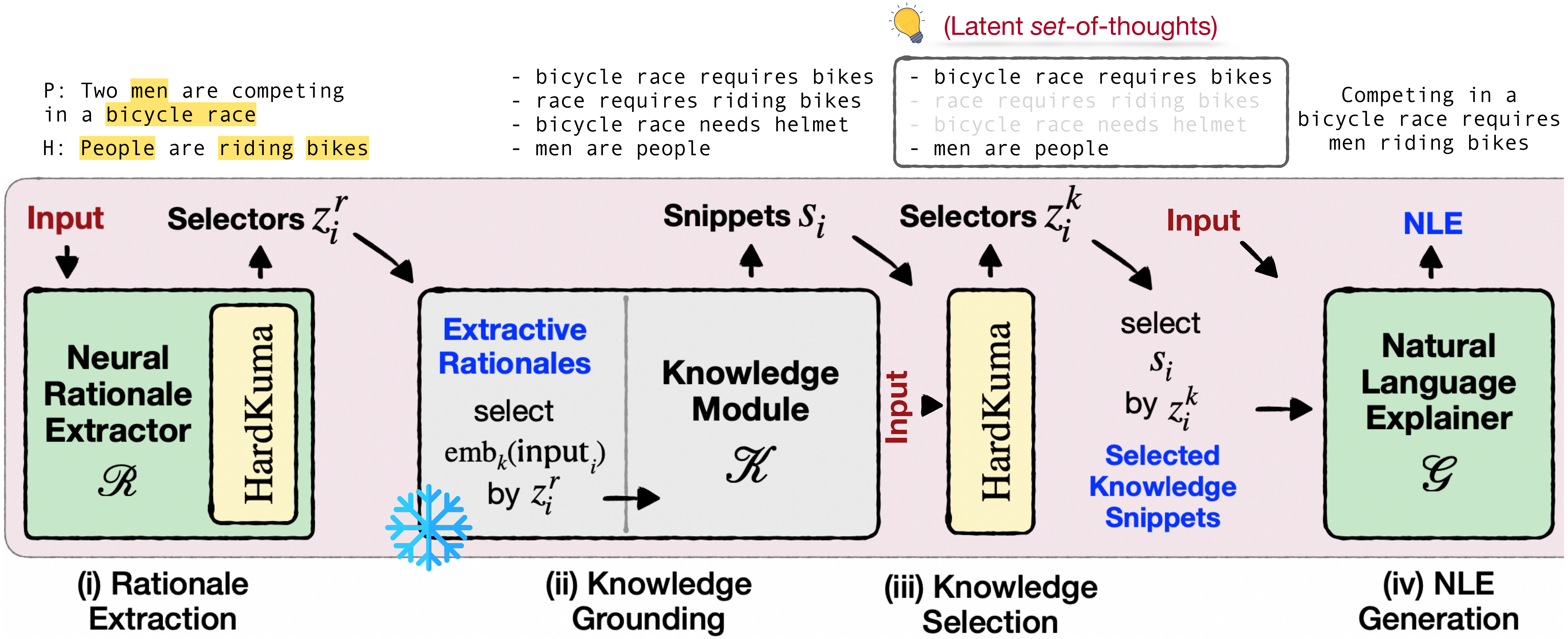
Rationales are responsible for relevant knowledge retrieval

Rationale + Knowledge



Rationales are responsible for relevant knowledge retrieval
Knowledge (latent) selection acts as a **soft bottleneck**

Rationale + Knowledge + NLE

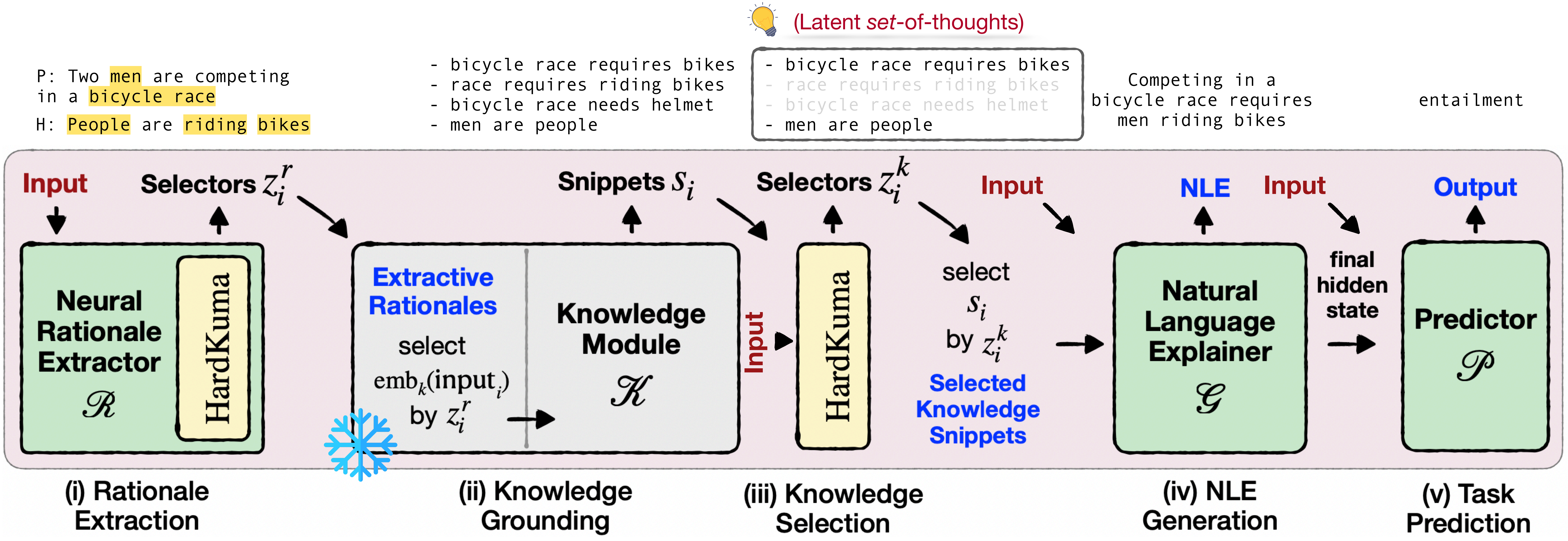


Rationales are responsible for relevant knowledge retrieval

Knowledge (latent) selection acts as a **soft bottleneck**

RExC is a **self-rationalizing** model that produces NLE

Rationale + Knowledge + NLE = RExC

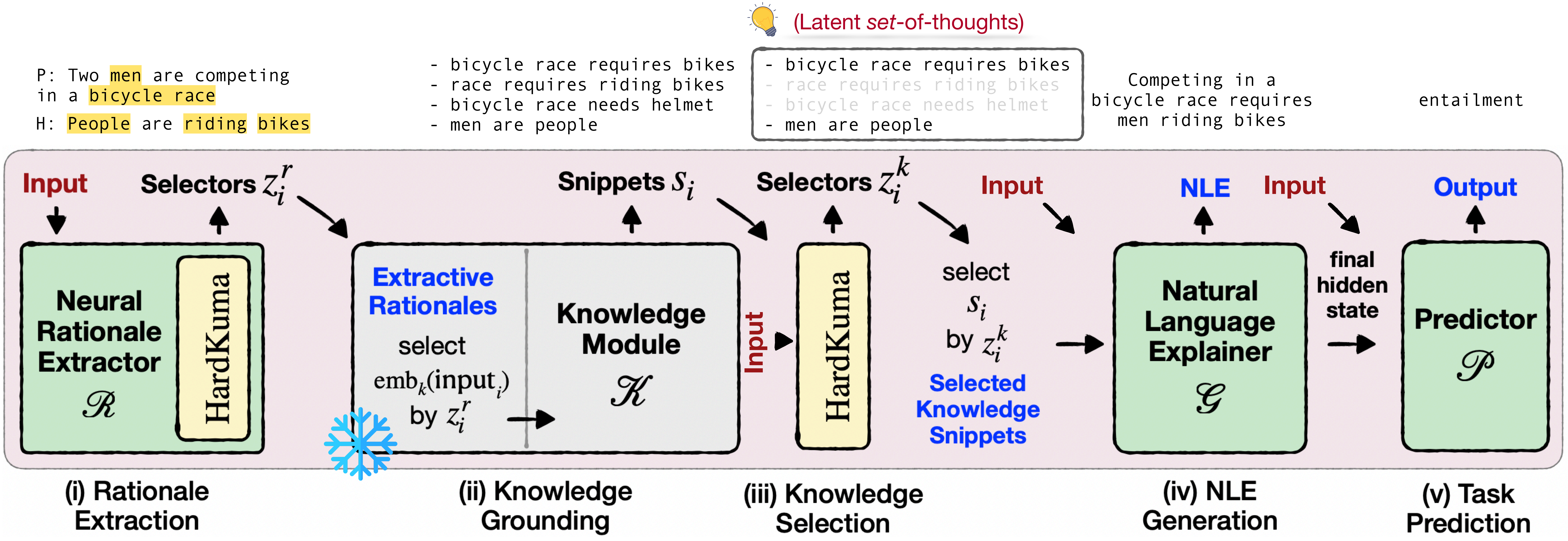


Rationales are responsible for relevant knowledge retrieval

Knowledge (latent) selection acts as a **soft bottleneck**

RExC is a **self-rationalizing** model that produces NLE and task output

Rationale + Knowledge + NLE = RExC



Loss from **gold NLEs**

Loss from **gold labels**

Natural Language and Visual-Language Tasks

Natural Language Tasks

- Natural Language Inference
- Commonsense Validation
- Commonsense QA

Vision Language Tasks

- Visual Entailment
- Visual Commonsense Reasoning

premise Two men are competing in a bicycle race

hypothesis People are riding bikes

label *entailment*

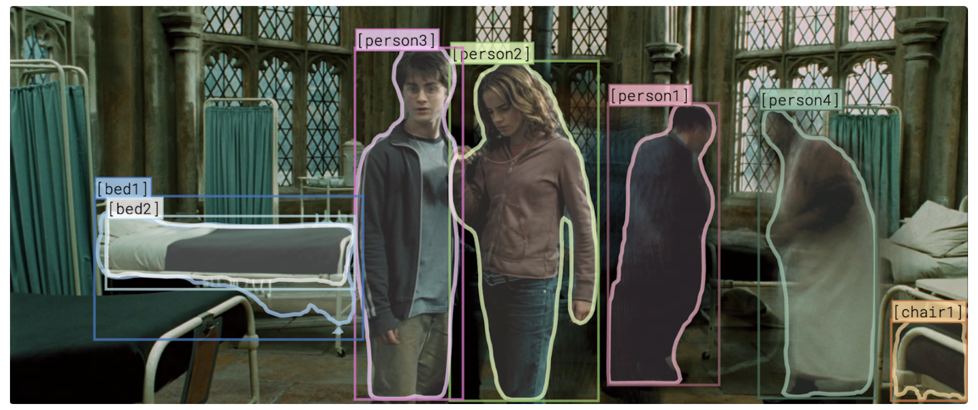
A: Coffee stimulates people
B: Coffee depresses people

label *B is invalid*

Q: Where does a wild bird usually live?

A: a) cage, b) sky, c) countryside, d) desert, e) windowsill

label *sky*



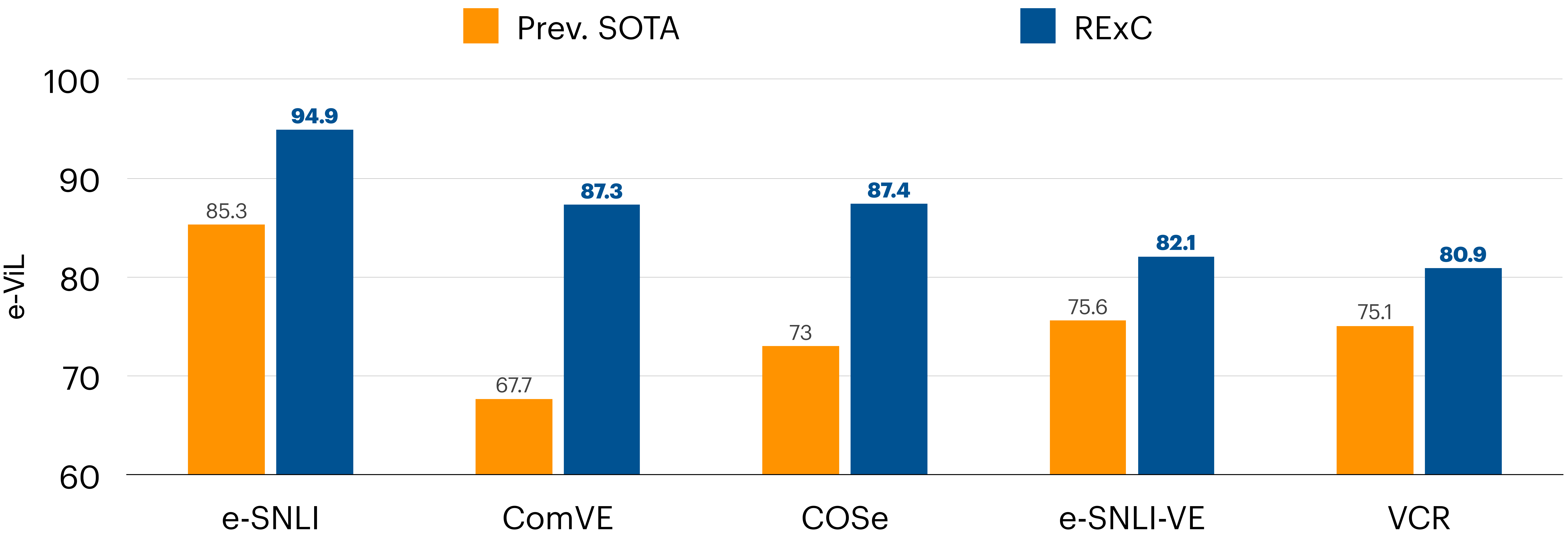
Hypothesis: Some tennis players pose

label *entailment*

Q: What is the place?

label *They are in a hospital room*

Human Evaluation of NLEs



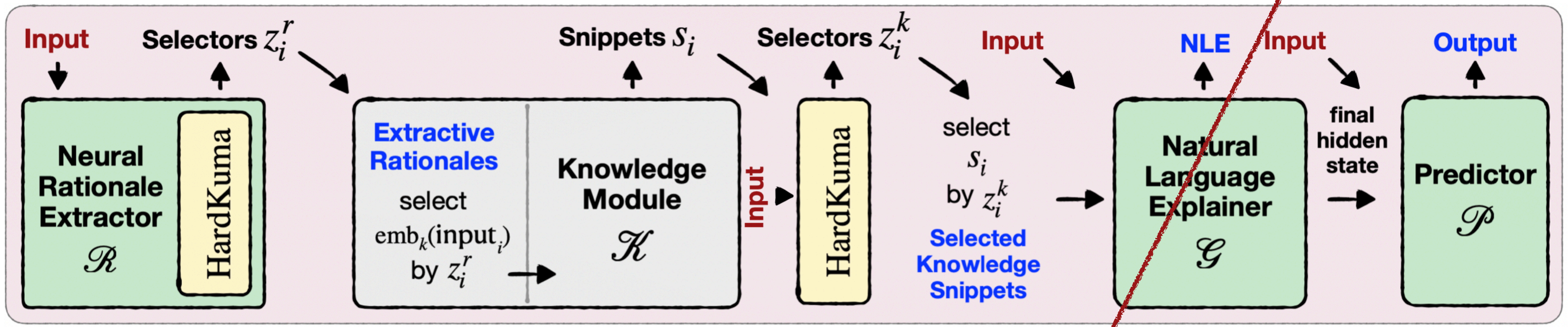
RExC outperforms all SOTA, being **highly rated** by human users

Rationale and Selected Knowledge **individually contribute** to performance

Zero-shot RExC

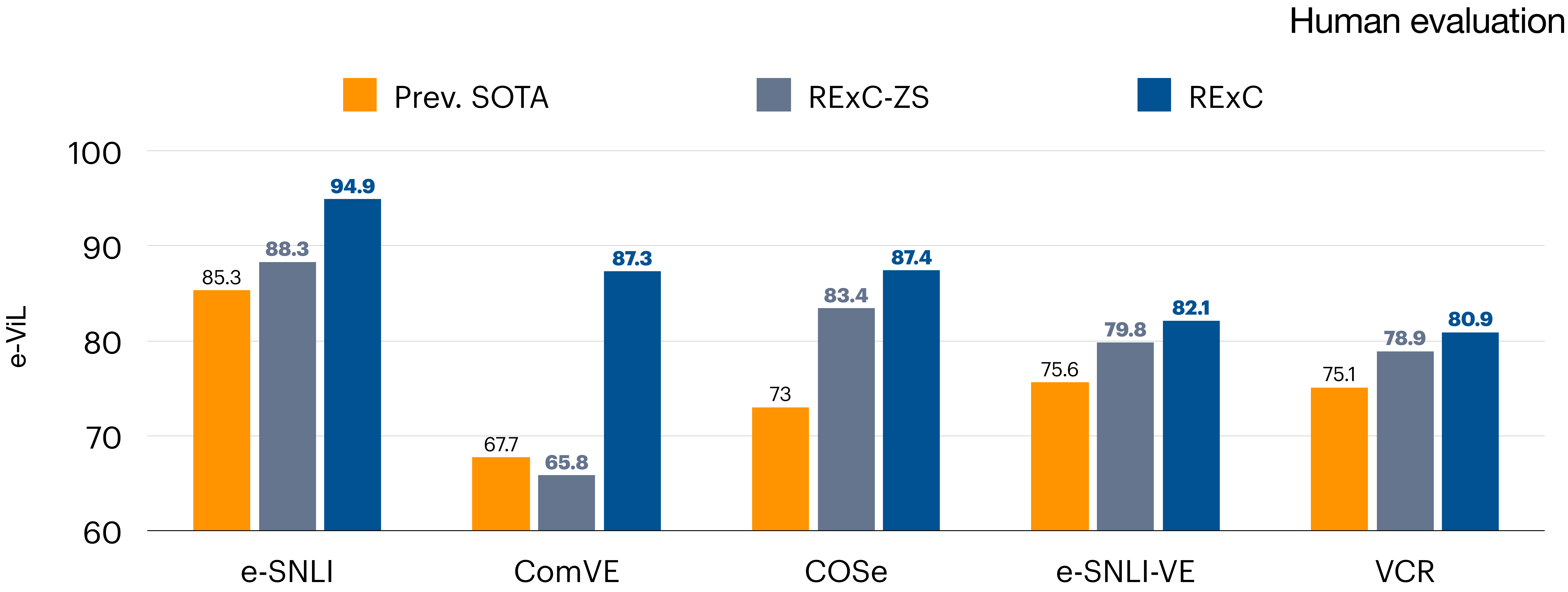
RExC-ZS

No Loss



What if we don't have gold NLE during training?

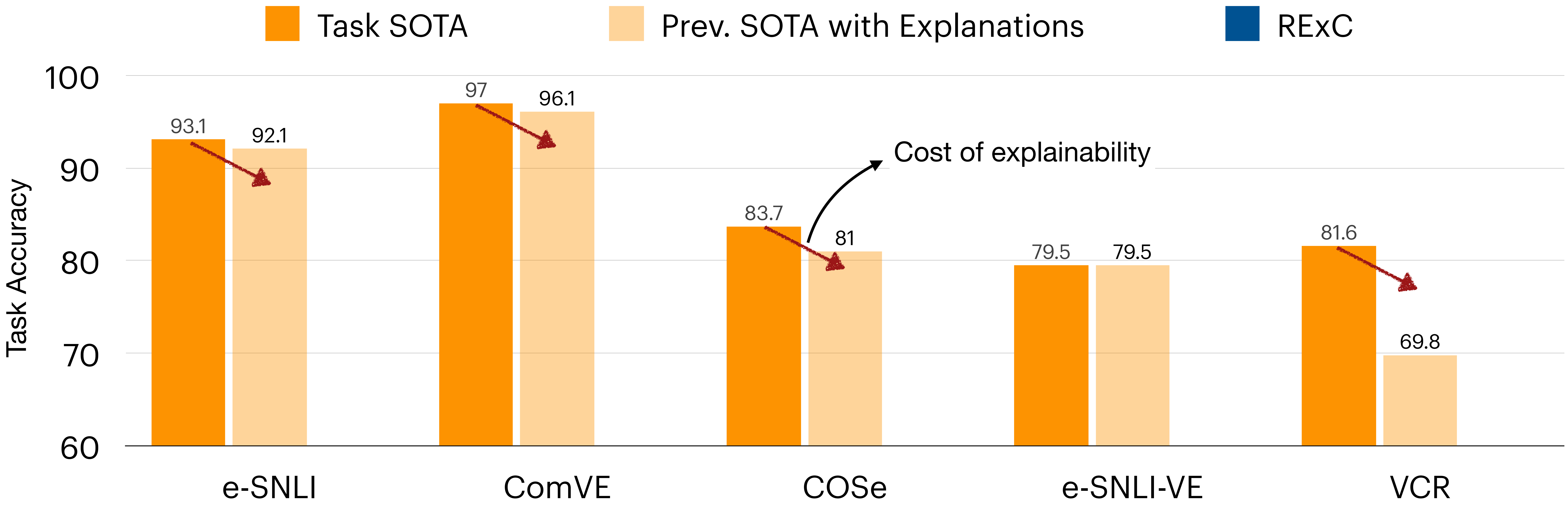
Zero-shot RExC



RExC-ZS is **at par or even better** than a supervised SOTA model

RExC Closing Performance-Explainability Gap

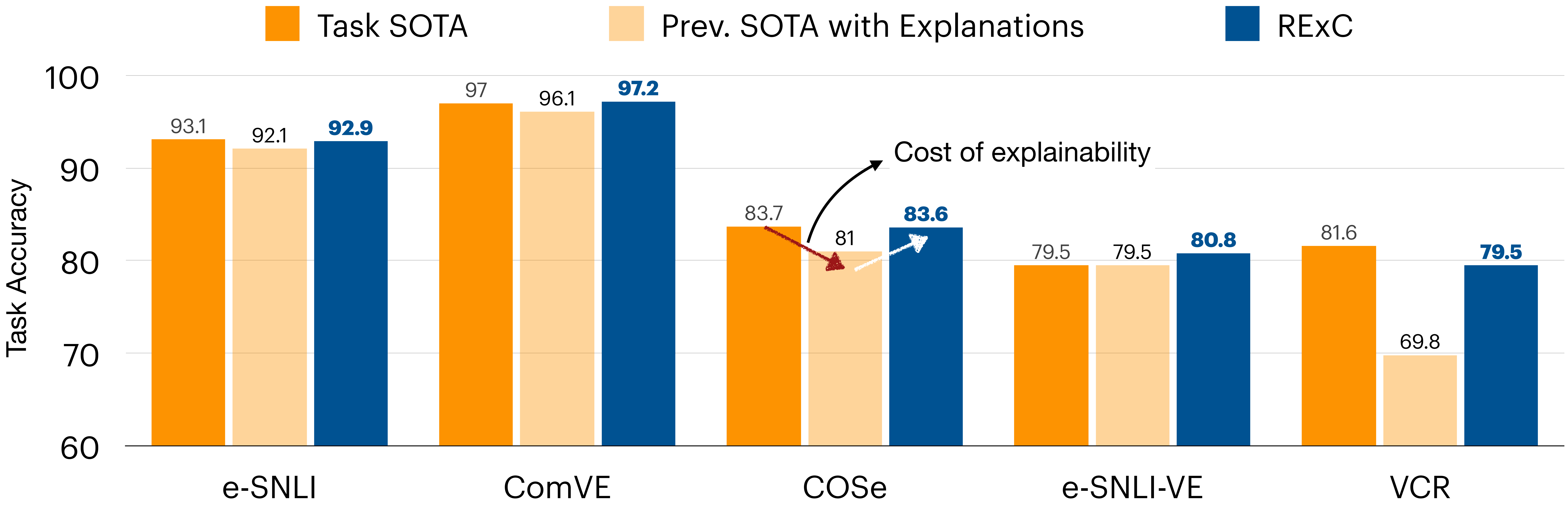
[Dalvi et al., 2022; Camburu et al., 2018; Narang et al., 2020]



Explainability comes at a cost, predictability drops

RExC Closing Performance-Explainability Gap

[Dalvi et al., 2022; Camburu et al., 2018; Narang et al., 2020]



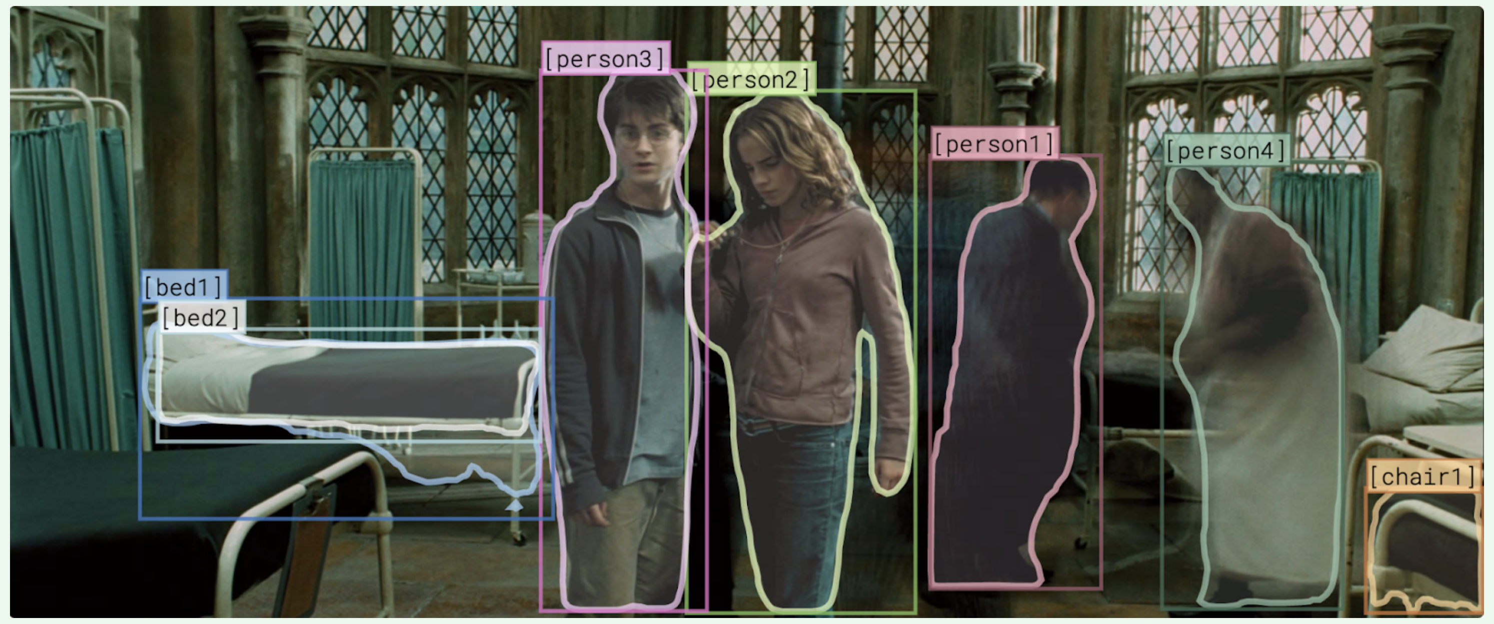
~~Explainability comes at a cost, predictability drops~~

RExC is **task SOTA** among models **with explanations**,
often outperforms all-time SOTA (mostly black-box)

Summary: Explanations + Knowledge Grounding

RExC

Q: Where are [person2] and [person3]?



A: They are in a hospital room

NLE: There are hospital beds and nurses in the room

Rationale:

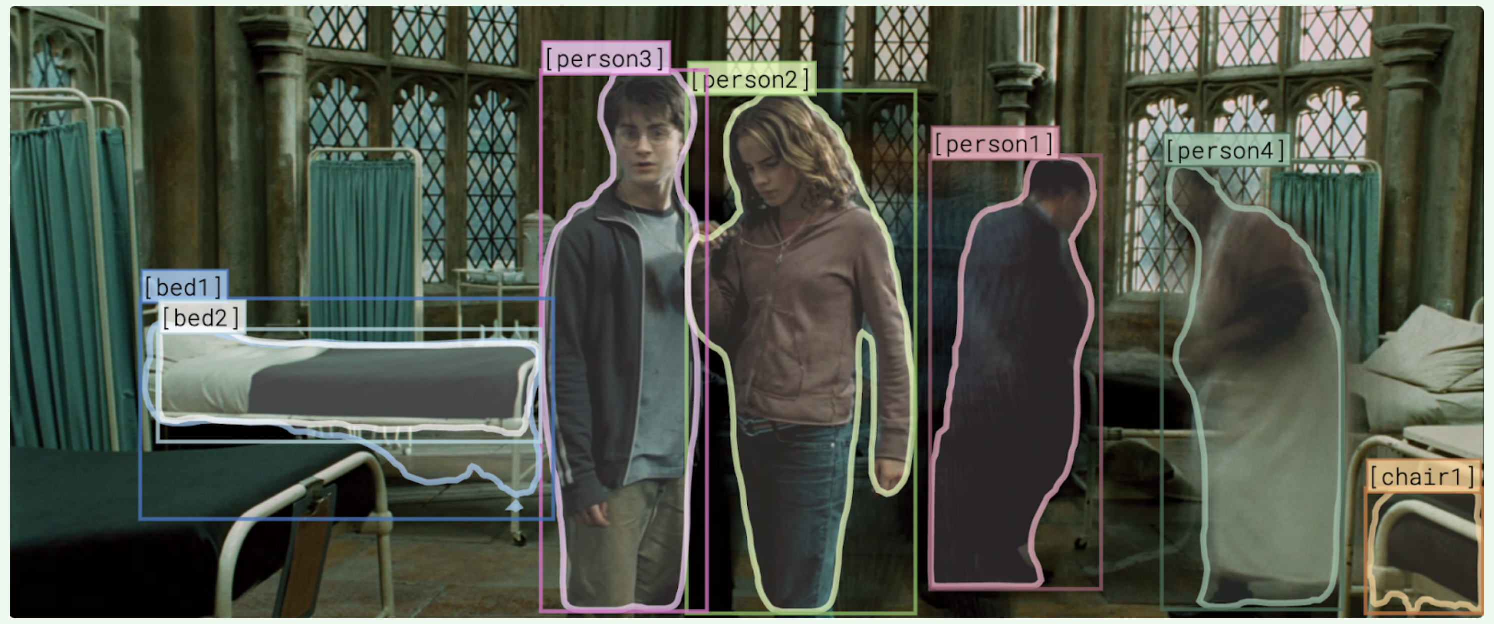
Selected Knowledge:

Hospital room has hospital beds
Hospital has nurses

Summary: Explanations + Knowledge Grounding

RExC

Q: Where are [person2] and [person3]?



A: They are in a hospital room

NLE: There are hospital beds and nurses in the room

Rationale: 

Selected Knowledge:	Hospital room has hospital beds Hospital has nurses

Chapter II. Explanations

Role of Knowledge Grounding in Generating Explanations

Majumder et al.
ICML 2022

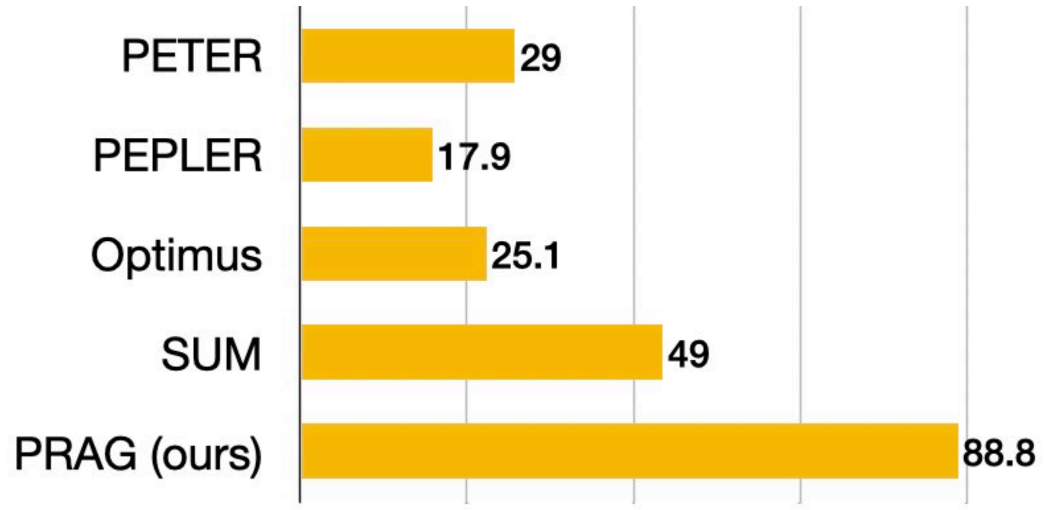
Knowledge **reduces ambiguity**

Attributing explanations

+ Emergent properties

Emergent Properties

Factual Explanation Generation
Xie, Singh, McAuley, **Majumder**
AAAI 2023



Evidence Entailment on *Movies Dataset*
(indicates factuality)

Question (based on **positive** rating)**:**
What was great?

Reviews retrieved based on Q:

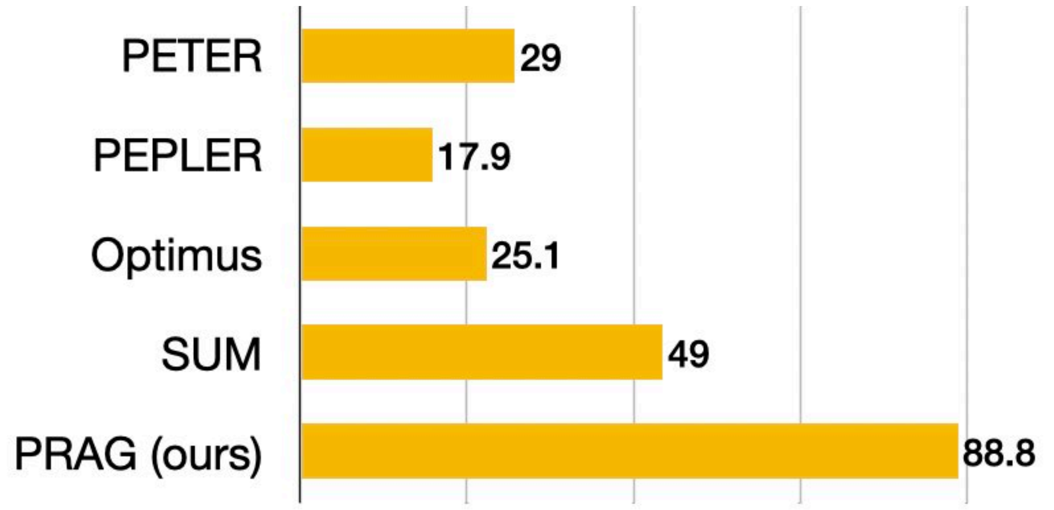
- The city views from the beautiful rooftop pool were incredible
- What probably makes this hotel really stand out is the rooftop pool...
- The pool is just fabulous

Generated Explanation:
Rooftop pool- you get an amazing view of the city with unspoiled views

Factuality

Emergent Properties

Factual Explanation Generation
Xie, Singh, McAuley, **Majumder**
AAAI 2023



Evidence Entailment on *Movies Dataset*
(indicates factuality)

Question (based on **positive** rating):
What was great?

Reviews retrieved based on Q:
- The city views from the beautiful rooftop pool were incredible
- What probably makes this hotel really stand out is the rooftop pool...
- The pool is just fabulous

Generated Explanation:
Rooftop pool- you get an amazing view of the city with unspoiled views

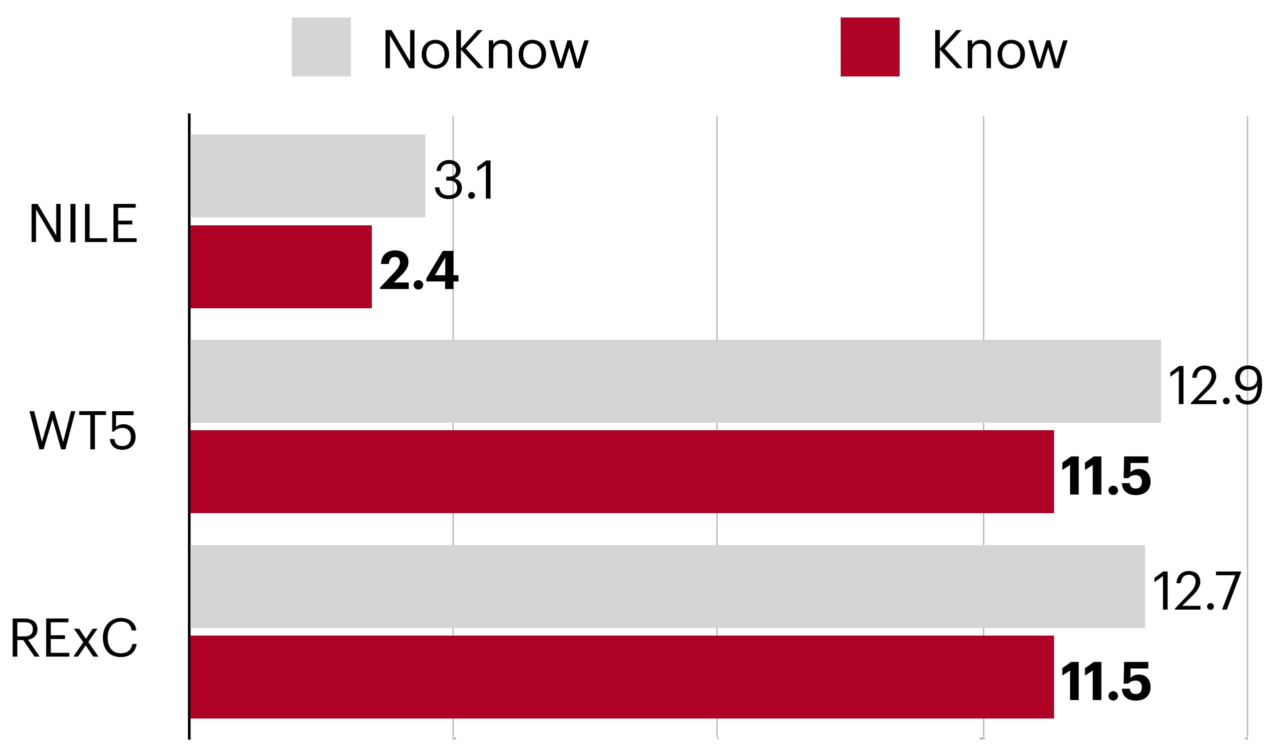
Factuality

Attacks and Robustness in NLEs
Jang, **Majumder** et al.
Preprint 2022

PREMISE: Two people using a water buffalo to cultivate a watery field.

HYPOTHESIS: Two people are outside with animals.
PREDICTED LABEL: Entailment
EXPLANATION: A water buffalo is an animal.

HYPOTHESIS: Two people are using a plant.
PREDICTED LABEL: Entailment
EXPLANATION: A water buffalo is a plant.

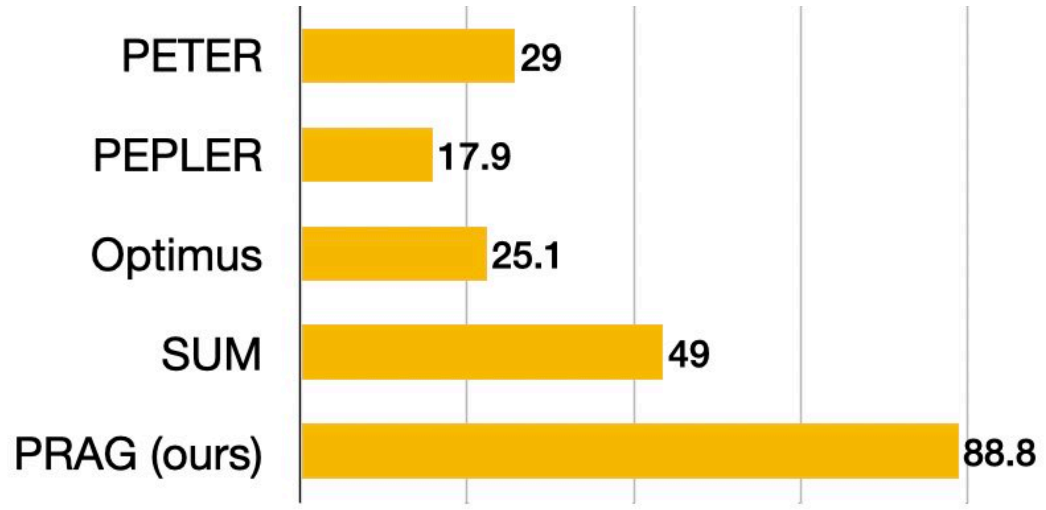


Attack Success Rate

Robustness

Emergent Properties

Factual Explanation Generation
Xie, Singh, McAuley, **Majumder**
AAAI 2023



Evidence Entailment on *Movies Dataset*
(indicates factuality)

Question (based on **positive** rating):
What was great?

Reviews retrieved based on Q:
- The city views from the beautiful rooftop pool were incredible
- What probably makes this hotel really stand out is the rooftop pool...
- The pool is just fabulous

Generated Explanation:
Rooftop pool- you get an amazing view of the city with unspoiled views

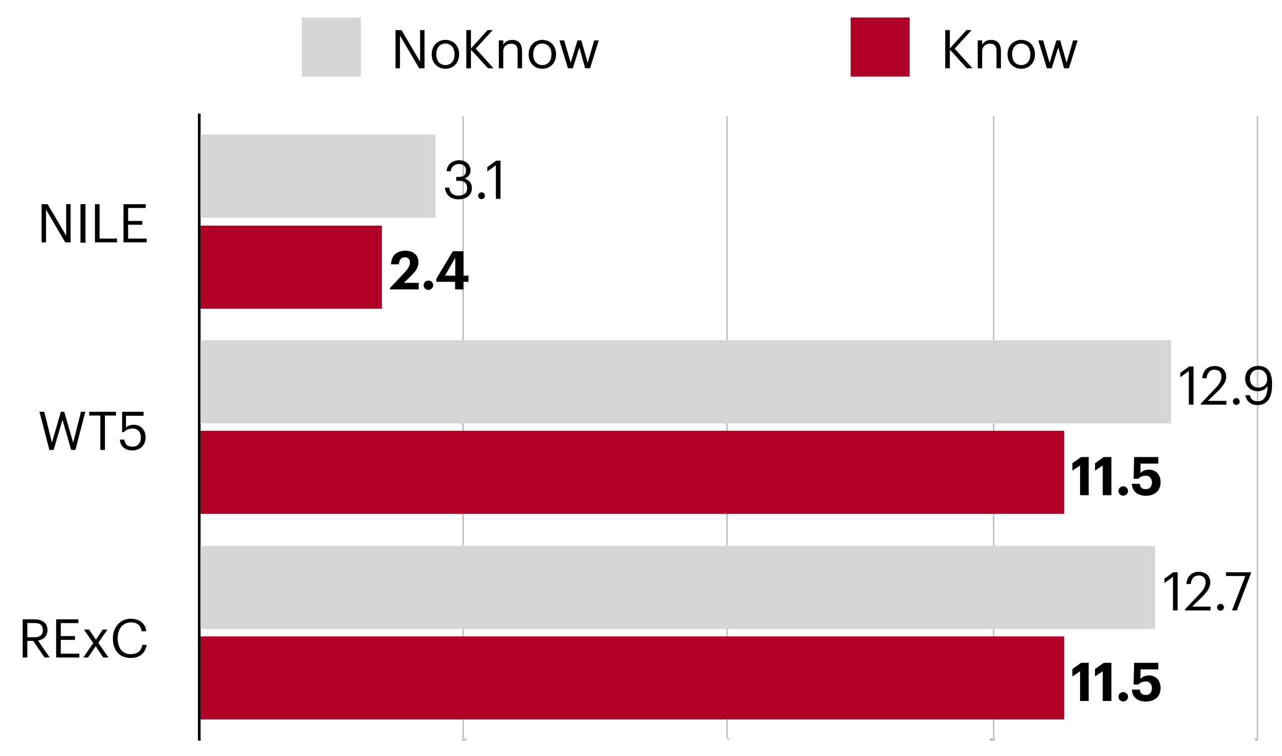
Factuality

Attacks and Robustness in NLEs
Jang, **Majumder** et al.
Preprint 2022

PREMISE: Two people using a water buffalo to cultivate a watery field.

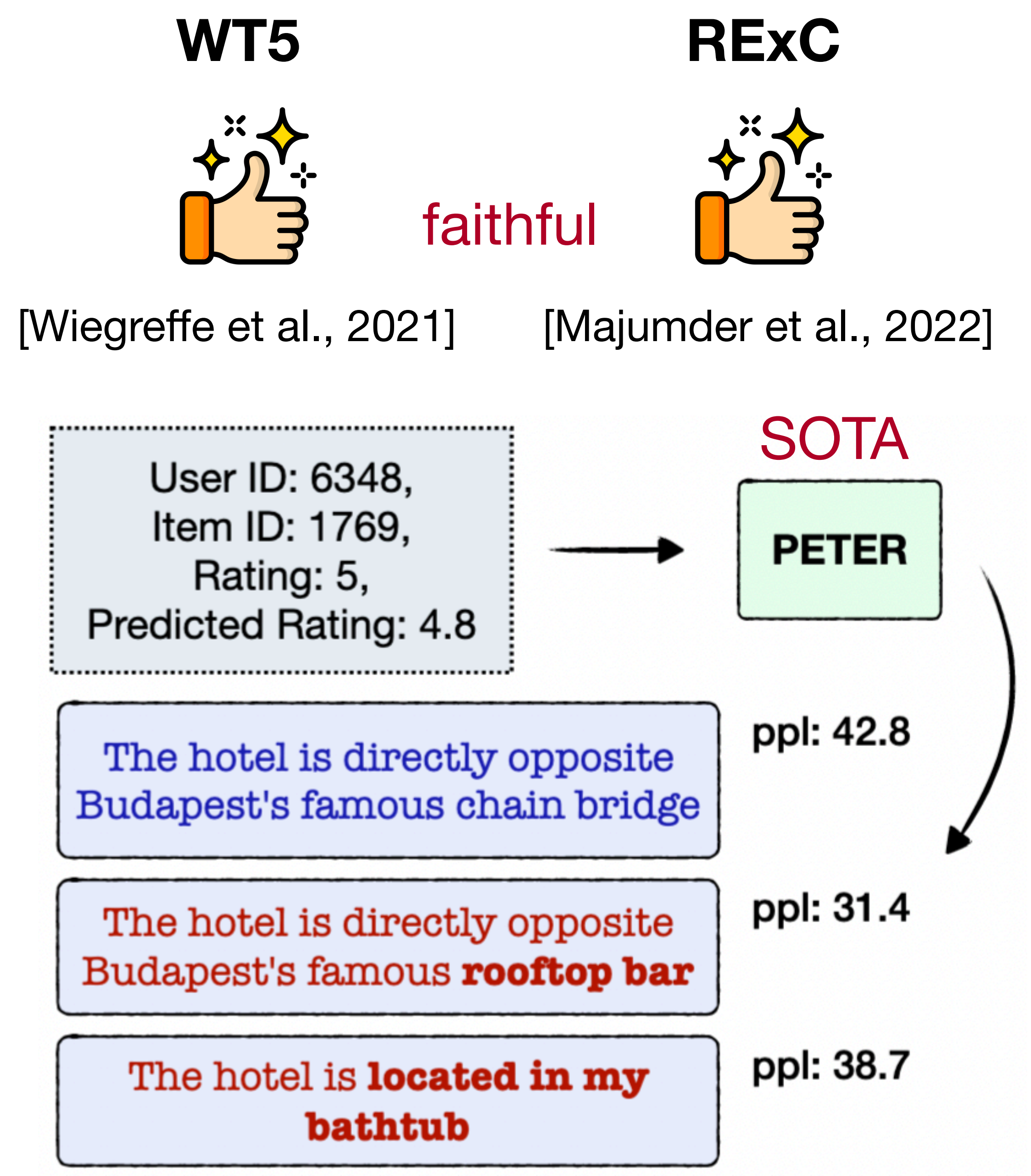
HYPOTHESIS: Two people are outside with animals.
PREDICTED LABEL: Entailment
EXPLANATION: A water buffalo is an animal.

HYPOTHESIS: Two people are using a plant.
PREDICTED LABEL: Entailment
EXPLANATION: A water buffalo is a plant.



Robustness

Faithfulness in Language Explanations
Xie, McAuley, **Majumder**
Preprint 2022



Knowledge-grounding improves this

Faithfulness

Impact: NLEs for Expert Tasks

NLEs for
Chest X-ray pathologies

[Kayser et al., 2022]

LABELS: Edema (Positive)

Clinical Evaluation:

Natural Language Explanations for ***Edema:***

Ground-Truth: *Indistinct appearance of the pulmonary vasculature is compatible with pulmonary edema.*		2
RATCHET: *Findings suggesting mild pulmonary edema.*		1
DPT: *Pulmonary edema and extensive bibasilar opacification appear slightly worse.*		3
TieNet: *Diffuse bilateral pulmonary opacities, likely edema.*		5

NLEs for
Figurative NLI

[Chakrabarty et al., 2022]

Type	Premise (literal)	Hypothesis (figurative*)	Label	Explanation
Metaphor	He *mentally assimilated* the knowledge or beliefs of his tribe.	He *absorbed the knowledge* or beliefs of his tribe.	E	To absorb something is to take it in and make it part of yourself.
	He *utterly decimated* his tribe's most deeply held beliefs.		C	Absorbed typically means to take in or take up something, while "utterly decimated" means to destroy completely.

Relevant, **Trustworthy**, and Adaptive **AI**

Chapter I.
Knowledge

Post-hoc Knowledge Injection to Make Models Relevant

Majumder et al.
ACL 2022

Chapter II. Explanations

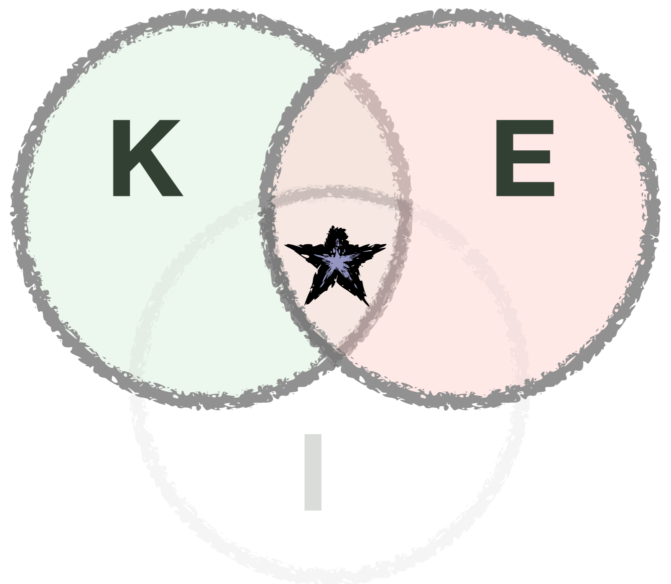
Role of Knowledge Grounding in Generating Explanations

Majumder et al.
ICML 2022

Chapter III.
Interactions

Improving Debiasing Performance with Natural Language Feedback

Majumder et al.
EMNLP & InterNLP 2022



Next-generation AI

Current AI + Knowledge **+** **Explanations** + Interactions

Relevant, Trustworthy, and Adaptive AI

Chapter I. Knowledge

Post-hoc Knowledge Injection to Make Models Relevant

Majumder et al.
ACL 2022

Chapter II. Explanations

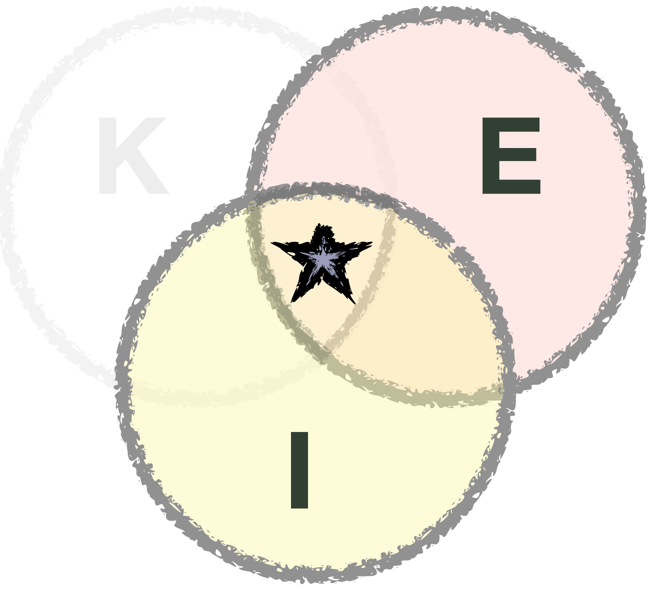
Role of Knowledge Grounding in Generating Explanations

Majumder et al.
ICML 2022

Chapter III. Interactions

Improving Debiasing Performance with Natural Language Feedback

Majumder et al.
EMNLP & InterNLP 2022



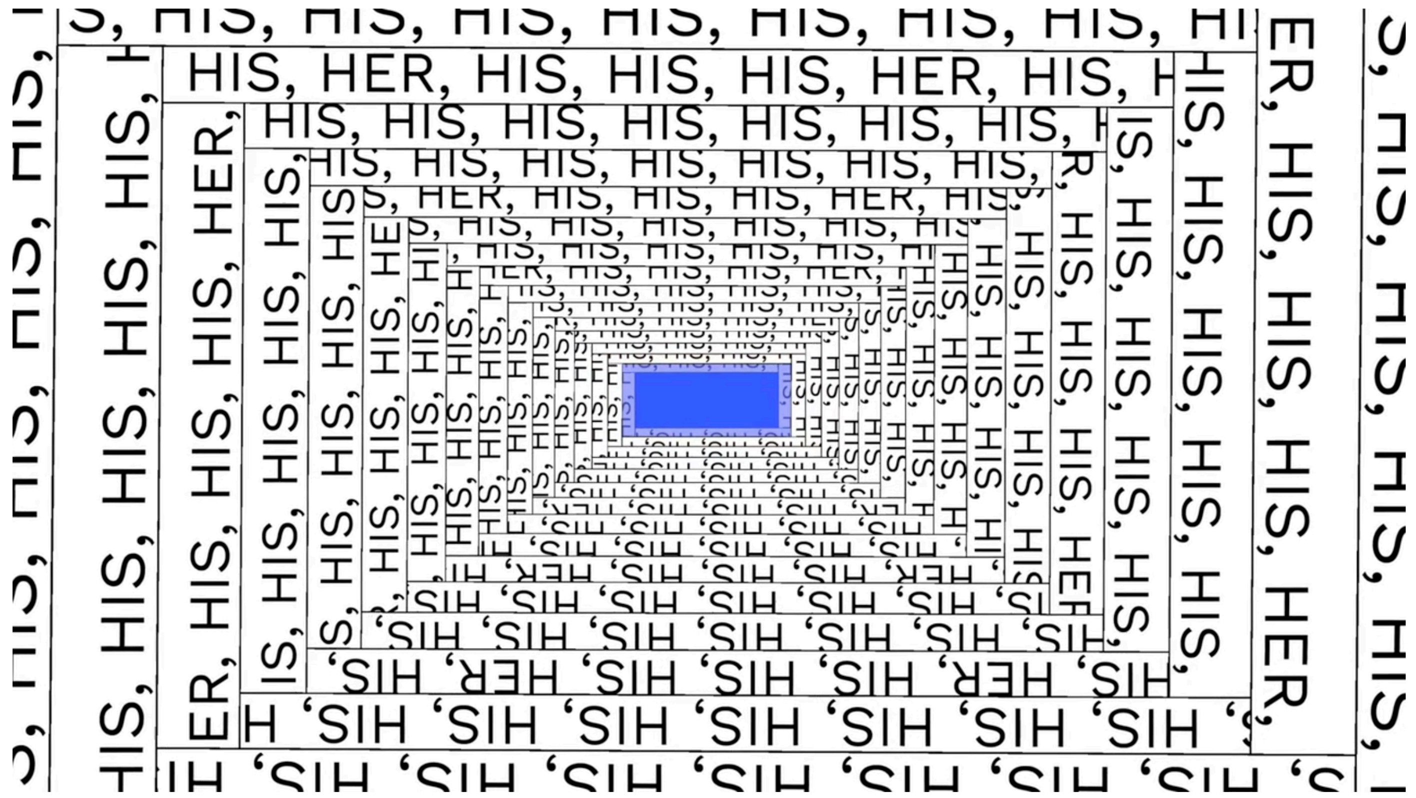
Next-generation AI

Current AI + Knowledge + Explanations + Interactions

Subjectivity (not) in AI

We Teach A.I. Systems Everything, Including Our Biases

Researchers say computer systems are learning from lots and lots of digitized books and news articles that could bake old attitudes into new technology.



By Kiel Mutschelknaus

+ **subjectivity**
+ **individual preferences**
+ **culture**

...

Human-in-the-loop **is the future**

MIT SLOAN EXPERTS | ARTIFICIAL INTELLIGENCE

'Human-Centered AI': How can the technology industry fight bias in machines and people?

by MIT Sloan Office of Media Relations | Nov 19, 2020

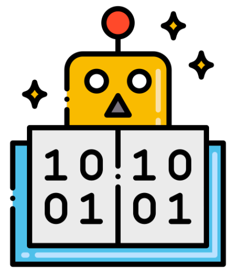
[Klie et al., 2020]
[Lee et al., 2020]
[Brantley et al., 2020]
[Simpson et al., 2019]
[Dasgupta et al., 2019]
[Radlinski et al., 2019]
[Smith-Renner et al., 2020]
[...]

Measuring Bias in Models

input

Angela Lindvall is a model and she has represented almost every major fashion brand

predicting **profession**



prediction

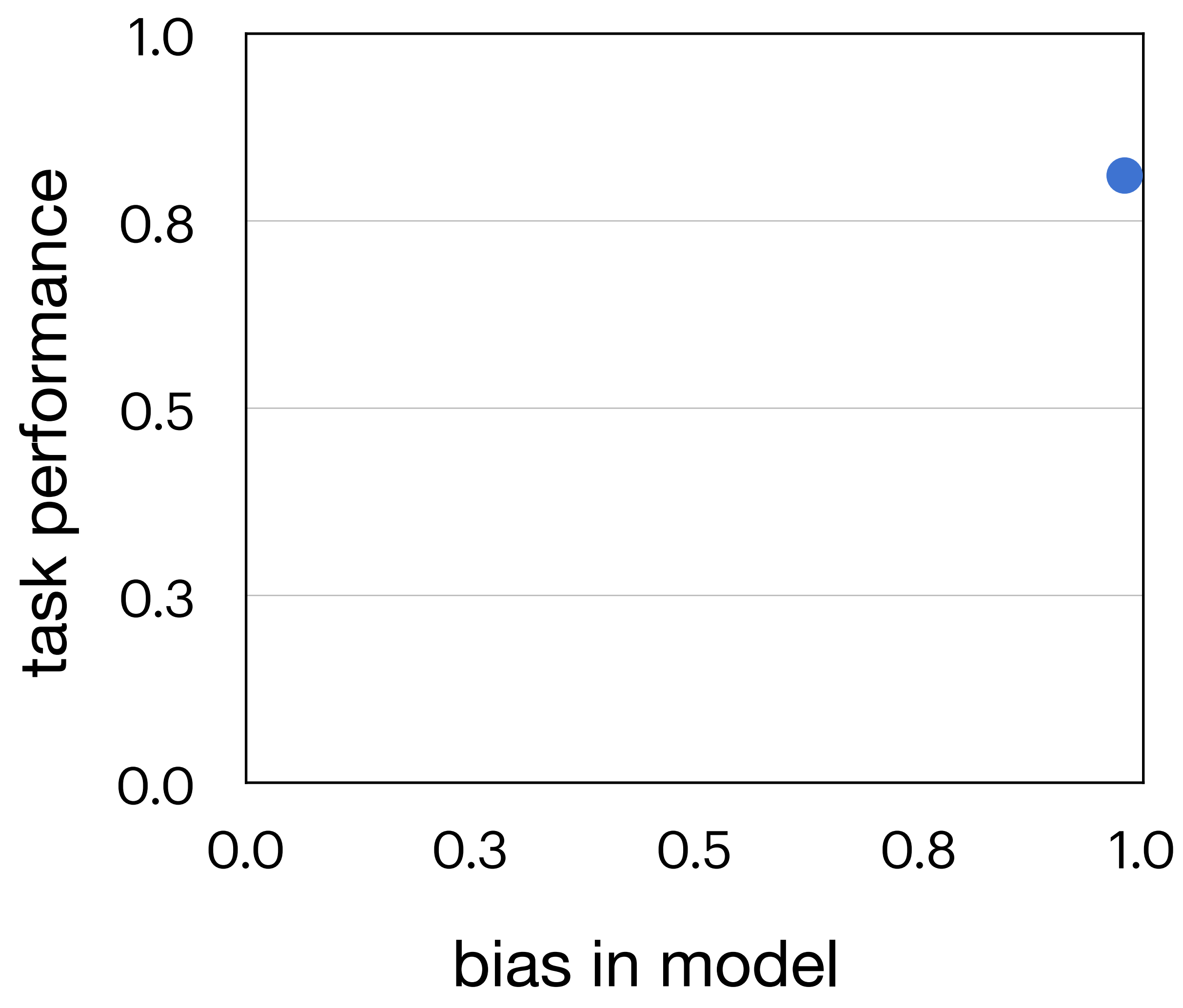
`model`

bias in data correlation with gender

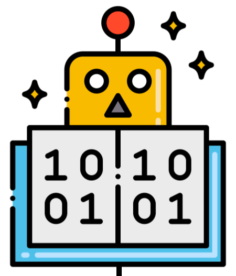
Measuring Bias in Models

input

Angela Lindvall is a model and she has represented almost every major fashion brand



predicting
profession



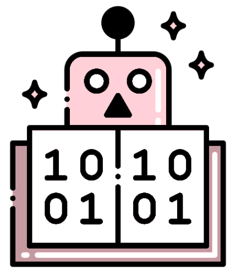
prediction
`model`

bias in data
correlation with gender

is my model biased?
probably yes

How to measure it?
representations

pre-trained
gender classifier

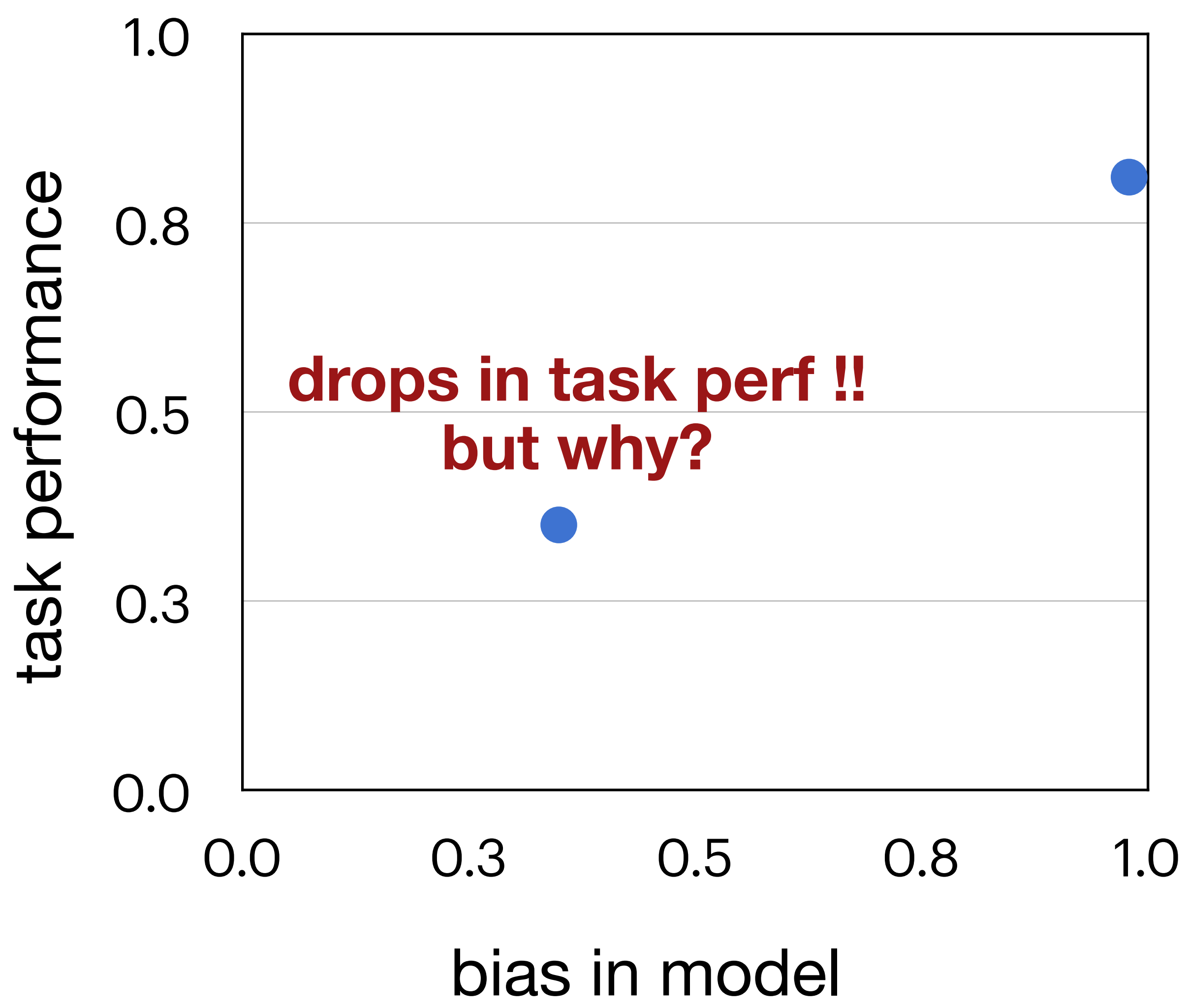


gender label
`female`

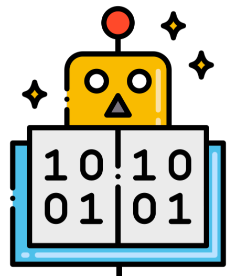
Measuring Bias in Models

input

Angela Lindvall is a model and she has represented almost every major fashion brand



predicting
profession



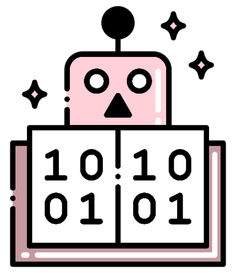
prediction
`model`

bias in data
correlation with gender

is my model biased?
probably yes

How to measure it?
representations

pre-trained
gender classifier



gender label

😕

Debiasing

Adversarial training

[Zhang et al., 2018]

Measuring Bias in Models

input

Angela Lindvall is a model and she has represented almost every major fashion brand

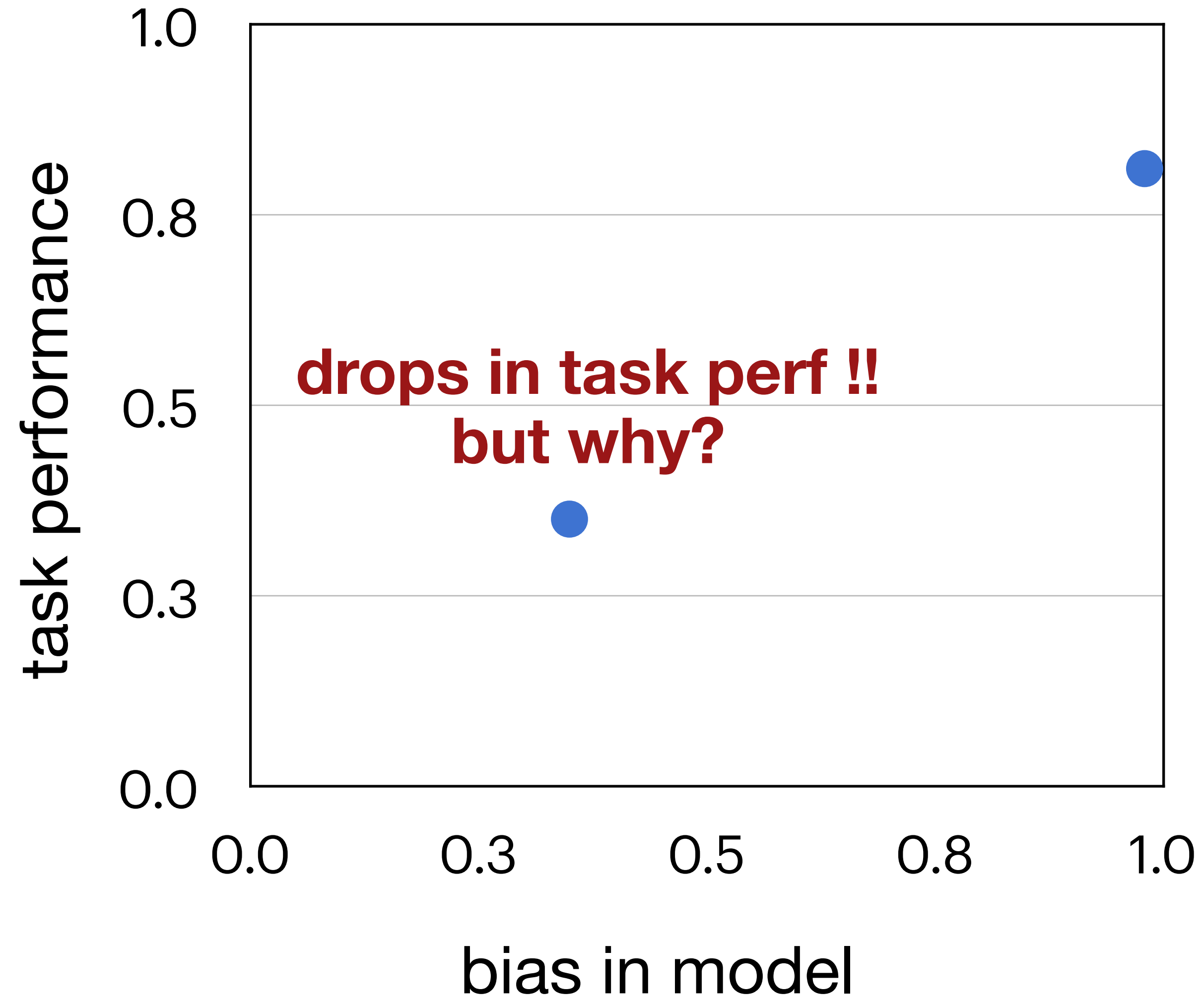
predicting
profession

prediction
`model`

bias in data
correlation with gender

is my model biased?
probably yes

How to measure it?
explanations



pre-trained
gender classifier

gender label

😕

Debiasing

Adversarial training

[Zhang et al., 2018]

*Rationales are **significant indicators** from input for a model's prediction

Measuring Bias in Rationales

biased (original) model

Angela Lindvall is a model and she has represented almost every major fashion brand

prediction
`model` ✓

debiased (adv) model

Angela Lindvall is a model and she has represented almost every major fashion brand

prediction
`fashion designer` ✗

*Rationales are **significant indicators** from input for a model's prediction

Measuring Bias in Rationales

biased (original) model

Angela Lindvall is a model and she has represented almost every major fashion brand

prediction
model ✓

debiased (adv) model

Angela Lindvall is a model and she has represented almost every major fashion brand

prediction
fashion designer ✗

How to fix?
Intervening model explanations

*Rationales are **significant indicators** from input for a model's prediction

Measuring Bias in Rationales

biased (original) model

Angela Lindvall is a model and she has represented almost every major fashion brand

prediction
`model` ✓

debiased (adv) model

Angela Lindvall is a model and she has represented almost every major fashion brand

prediction
`fashion designer` ✗

by adding back minimally biased tokens

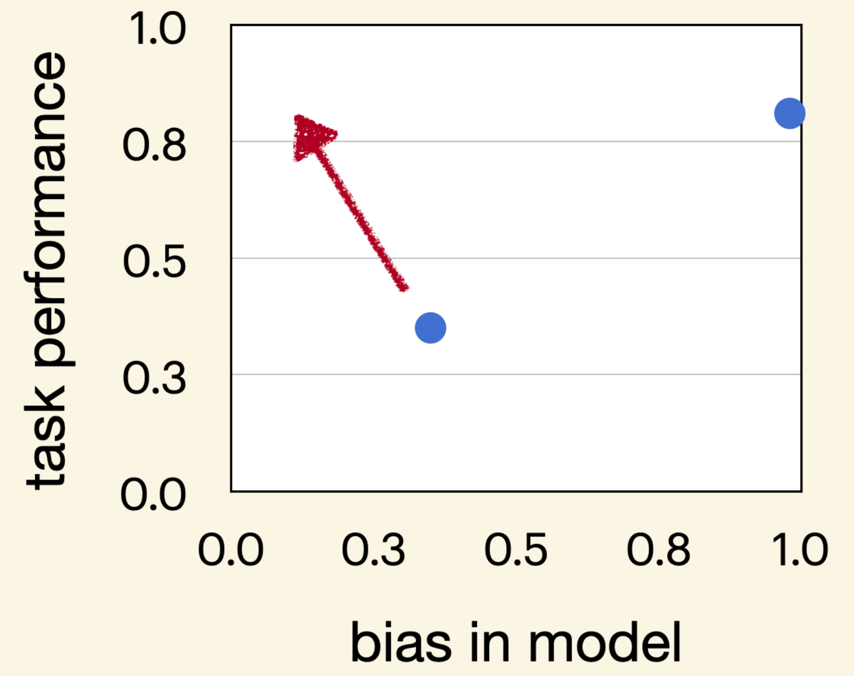
Angela Lindvall is a model and she has represented almost every major fashion brand

prediction
`model` ✓

How to fix?
Intervening model explanations

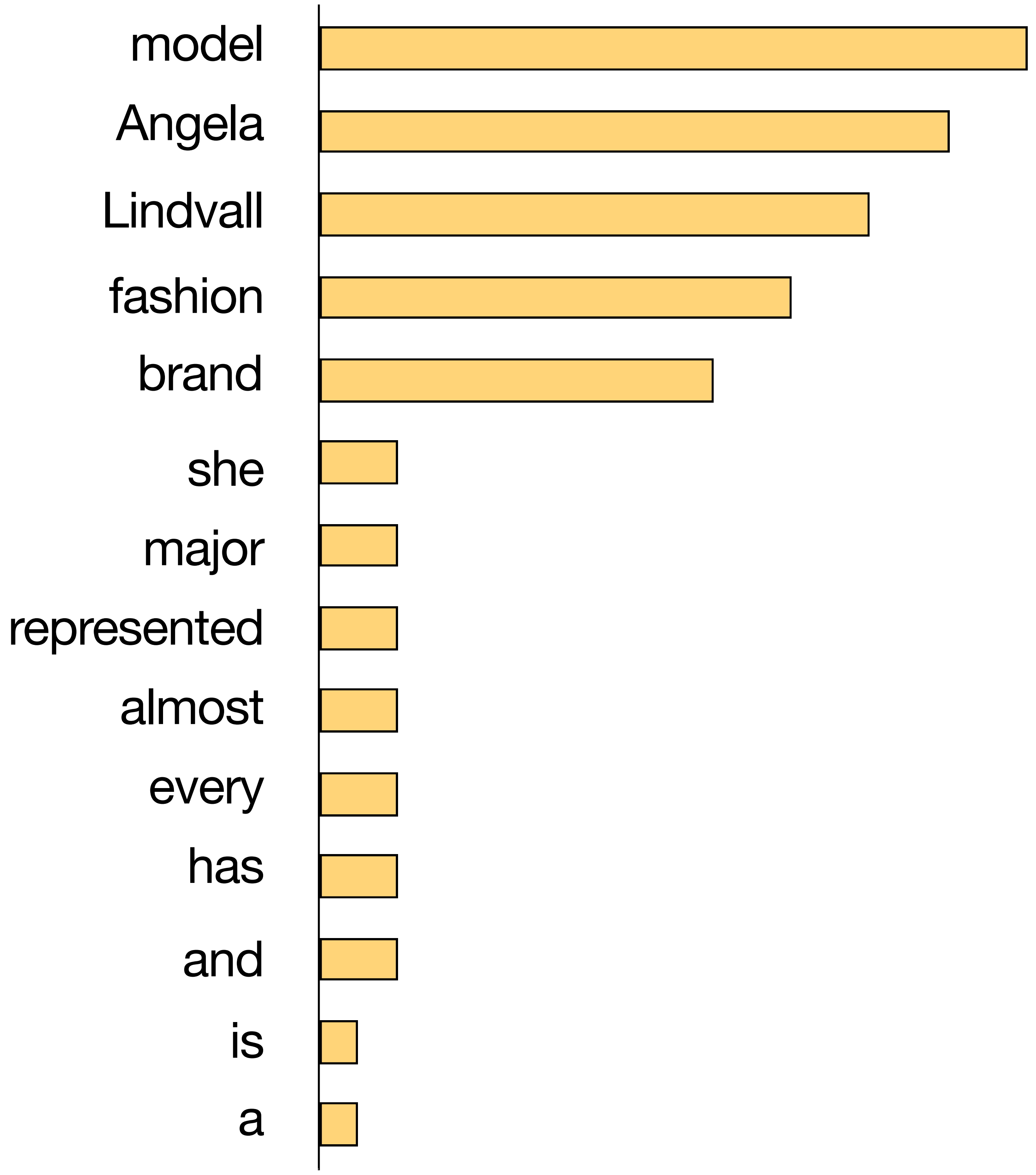
Fair Debiasing

[Caliskan et al., 2017; Sun et al., 2019]

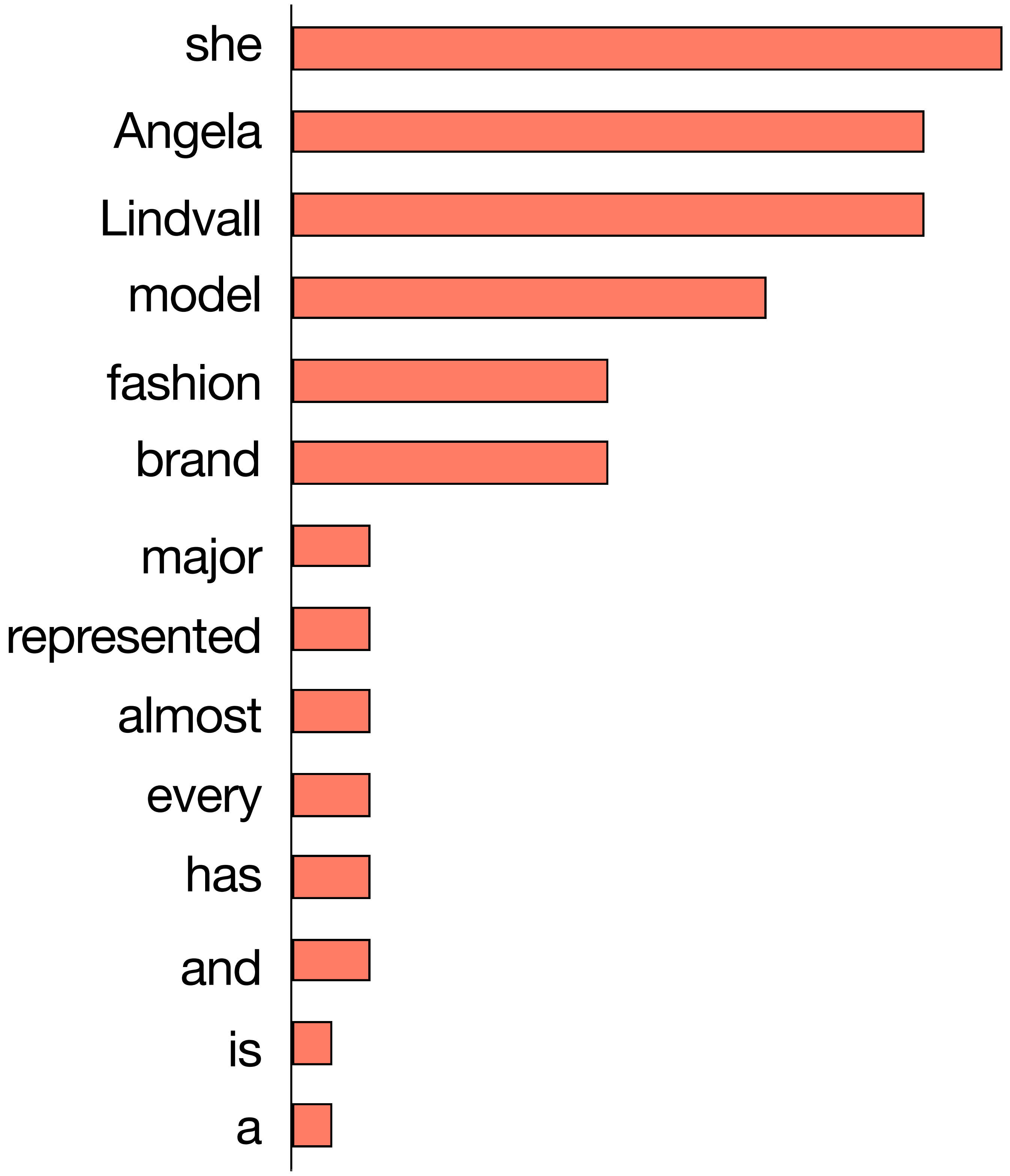


Debiasing by Intervening Explanations

Token contribution for predicting **profession**



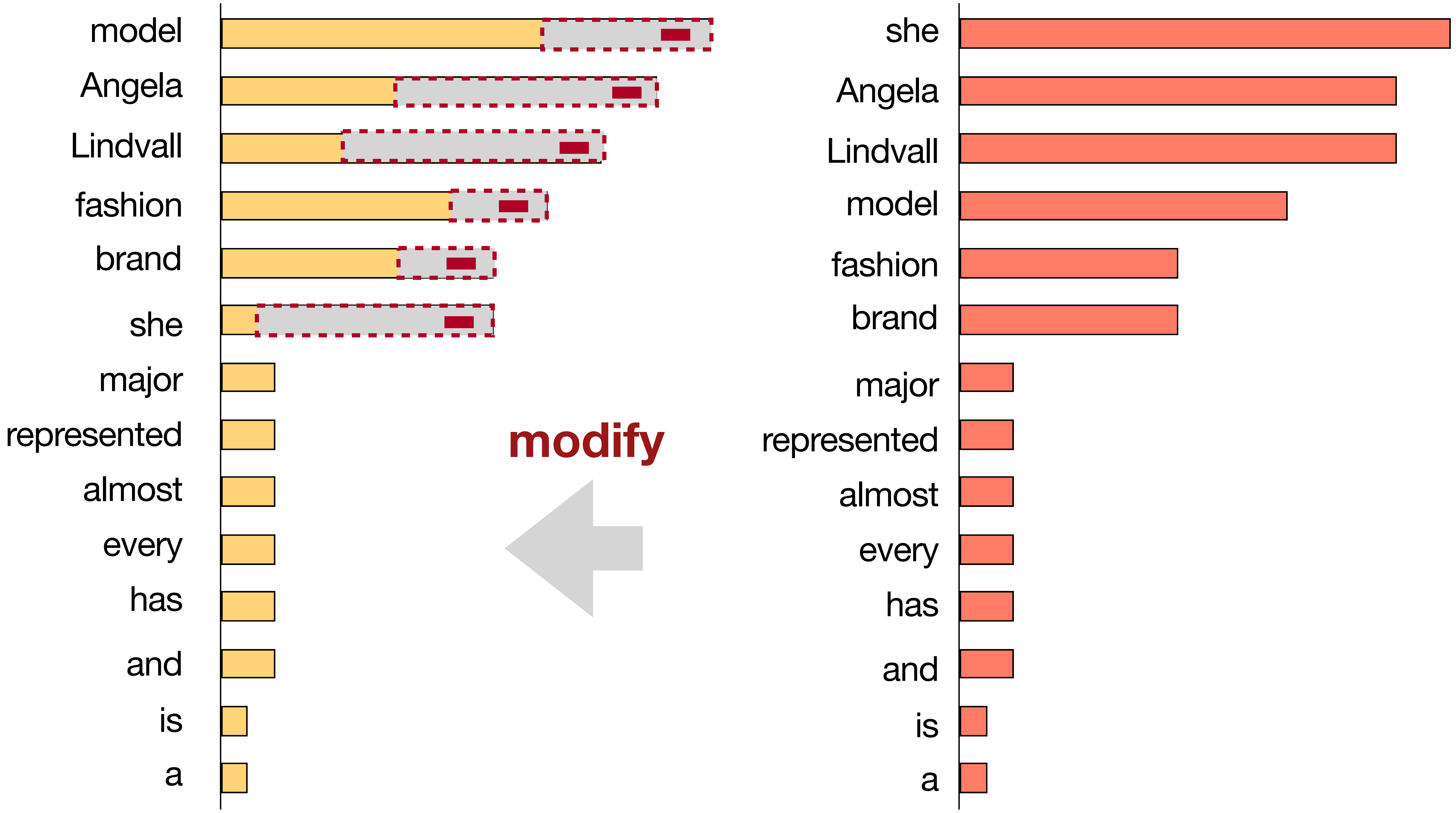
Token contribution for predicting **gender**



Debiasing by Intervening Explanations

Token contribution for predicting **profession**

Token contribution for predicting **gender**



what

penalize task contribution

if bias contribution is high
else no penalty

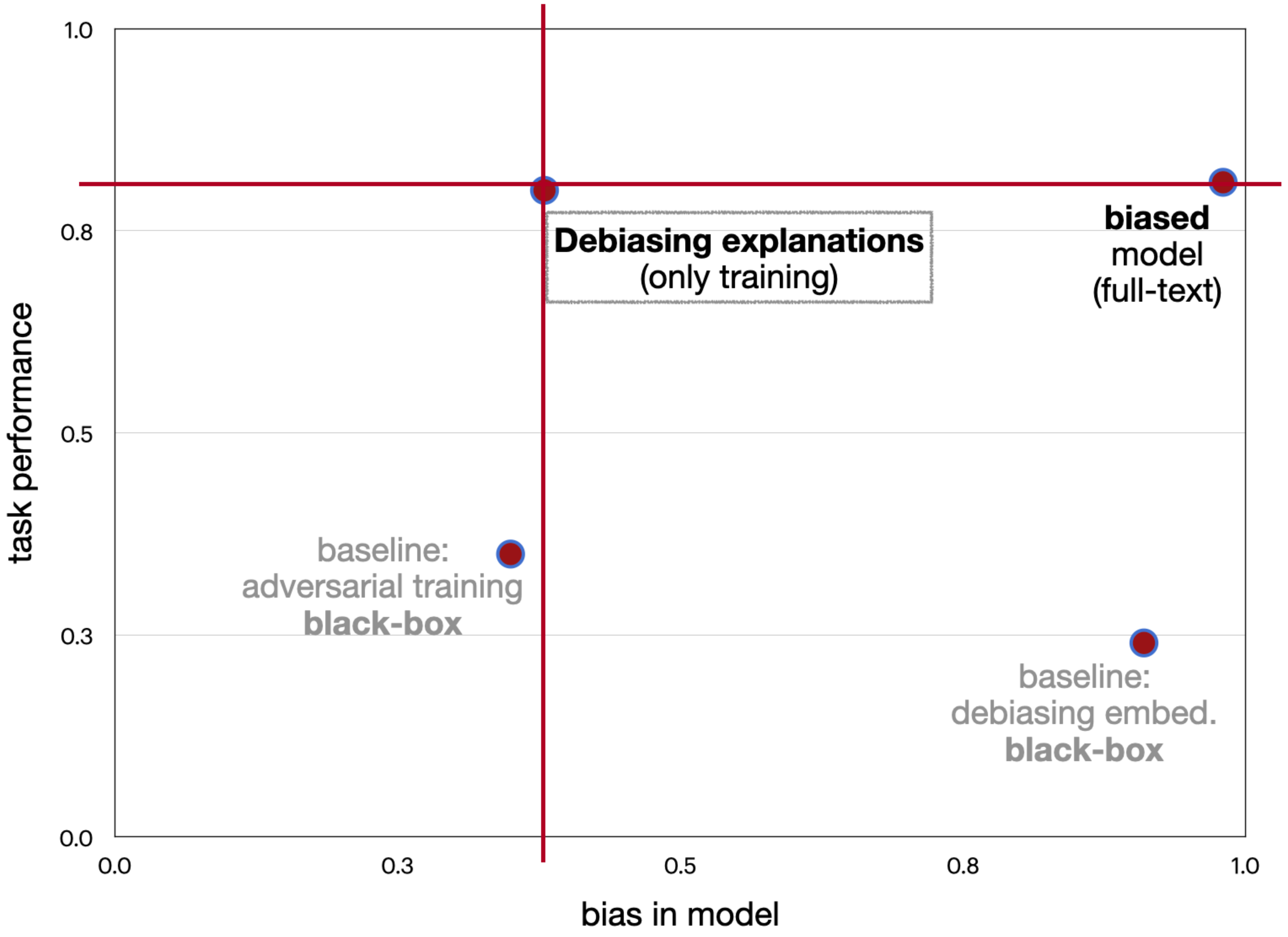
why

To promote
moderate-task and low-bias
important tokens

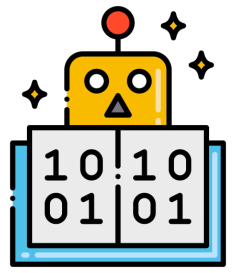
in place of

high-task and high-bias
important tokens

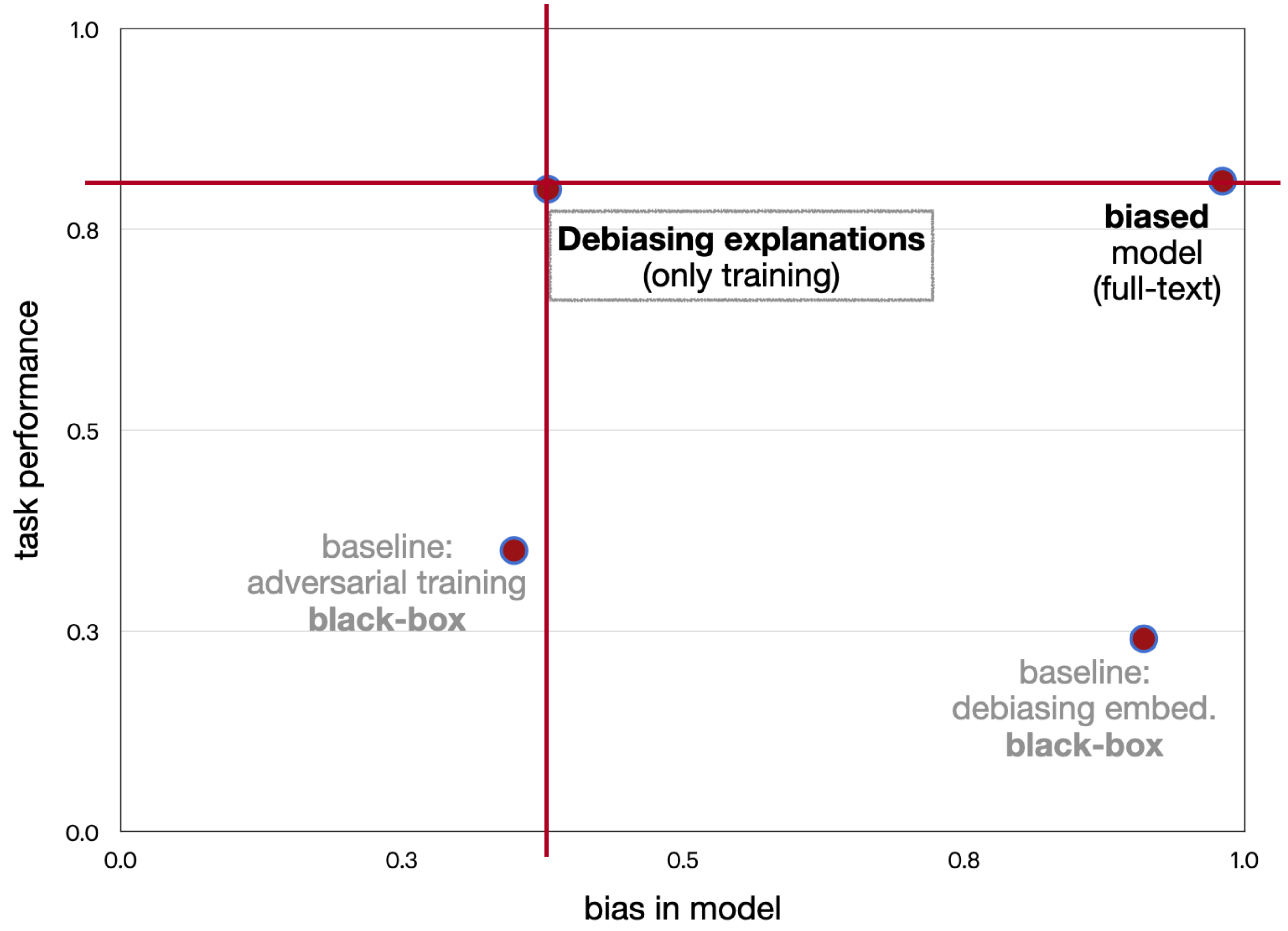
Training for Debiasing Explanations



Training for Debiasing Explanations

	Input		Prediction	Task Rationales	Bias Rationales
	Angela Lindvall is a model and she has represented almost every major fashion brand	(frozen) Classifier	Model ✓	Angela Lindvall is a model and she has represented almost every major fashion brand	Angela Lindvall is a model and she has represented almost every major fashion brand

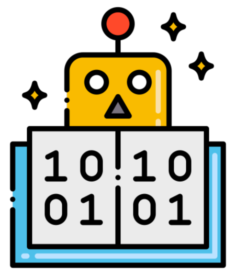
From a fixed pre-trained gender classifier



- Name is not needed
- Word model is sufficient

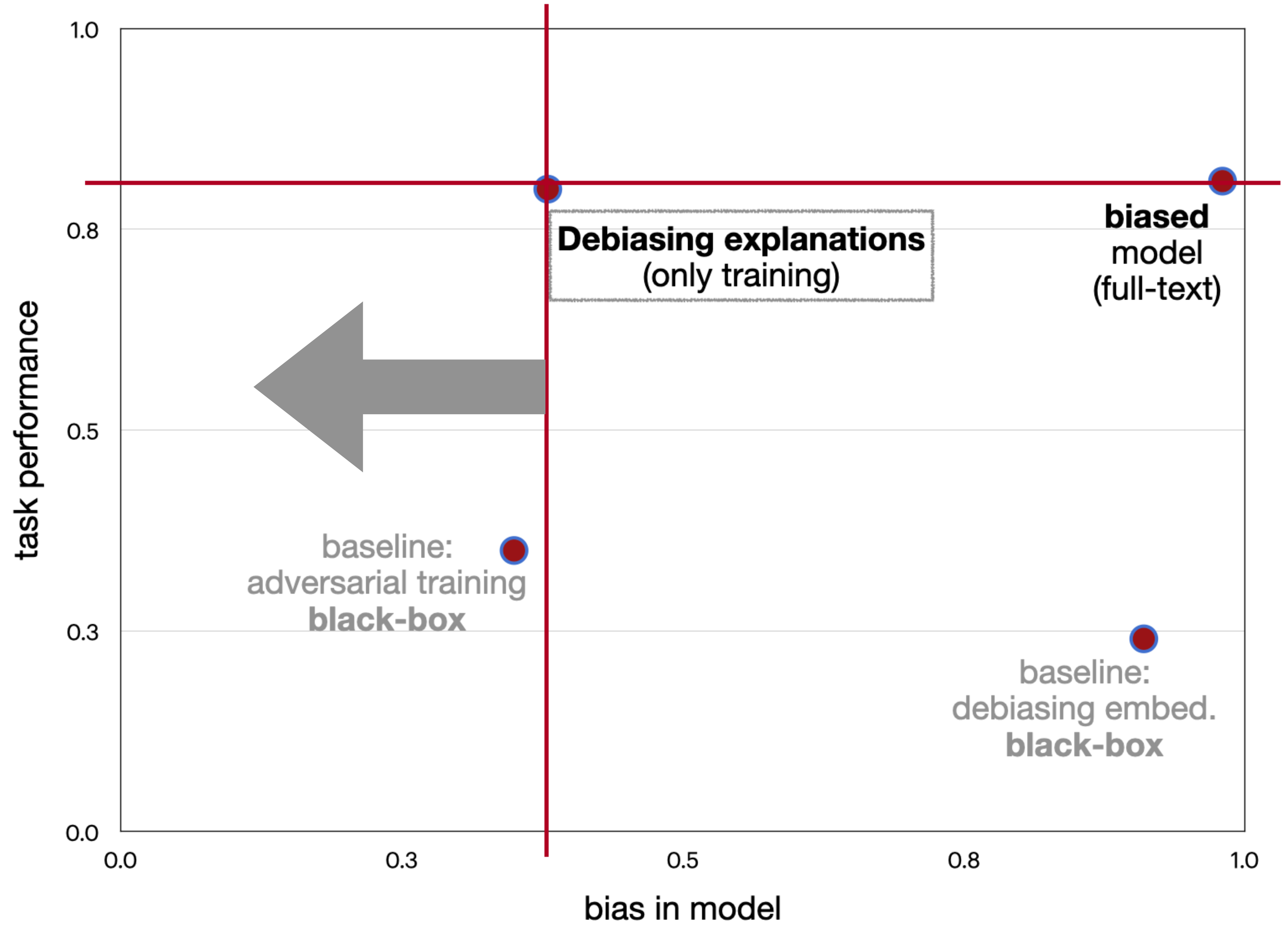
Bias Classifier is not *perfect, neither is the data*

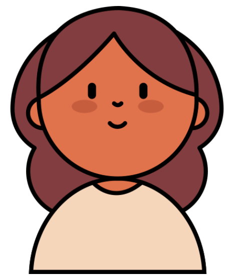
Training for Debiasing Explanations



Input		Prediction	Task Rationales	Bias Rationales
Angela Lindvall is a model and she has represented almost every major fashion brand	(frozen) Classifier	Model	Angela Lindvall is a model and she has represented almost every major fashion brand	Angela Lindvall is a model and she has represented almost every major fashion brand

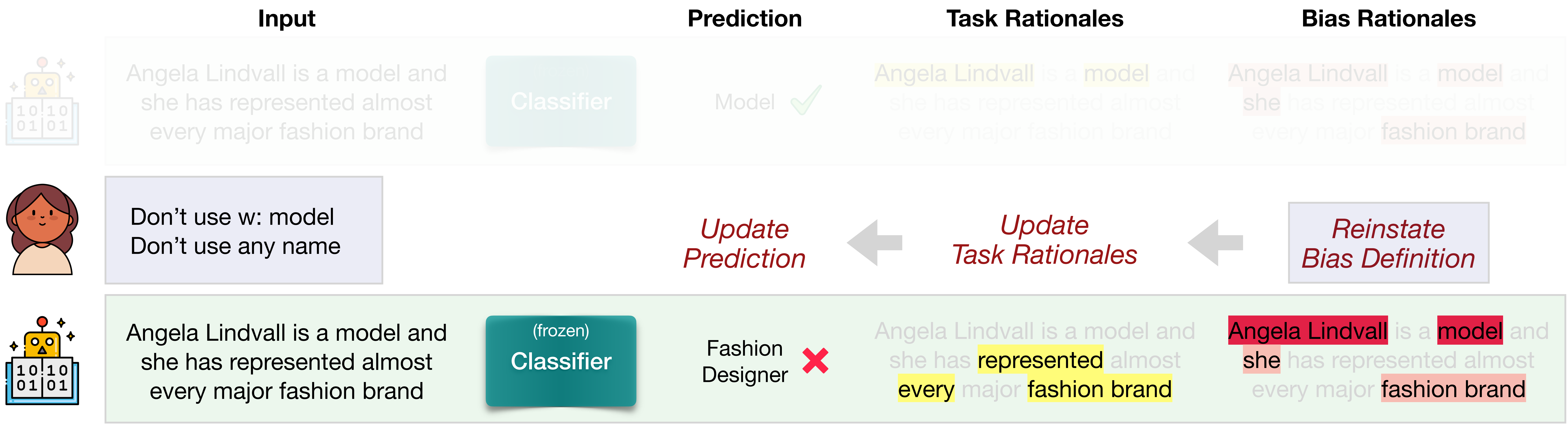
From a fixed pre-trained gender classifier

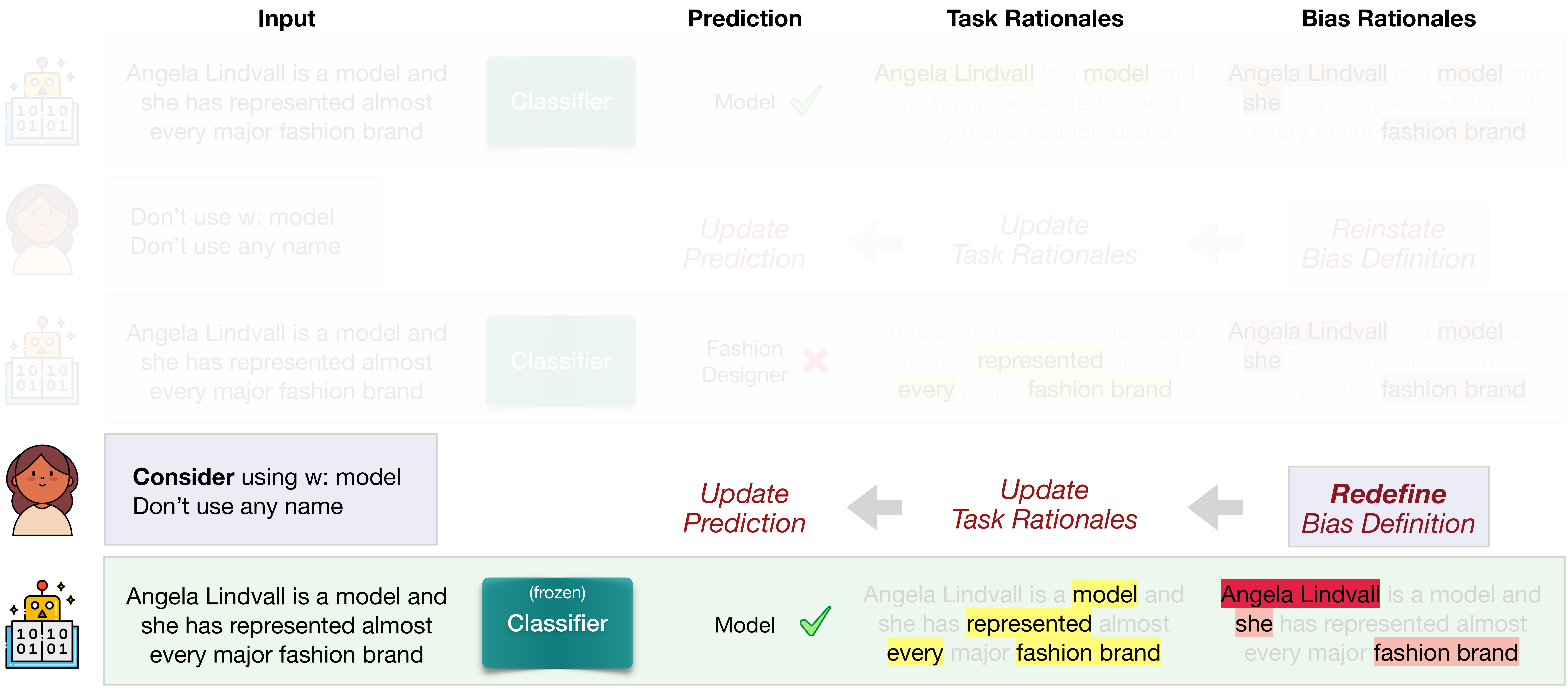




Debiasing is ***subjective*** to a user
Can be better at ***teaching*** the model

Input		Prediction	Task Rationales	Bias Rationales
Angela Lindvall is a model and she has represented almost every major fashion brand	(frozen) Classifier	Model ✓	**Angela Lindvall** is a **model** and she has represented almost every major fashion brand	**Angela Lindvall** is a **model** and **she** has represented almost every major **fashion brand**


Input
Prediction
Task Rationales
Bias Rationales
Angela Lindvall is a model and she has represented almost every major fashion brand
Classifier
Model
Angela Lindvall is a model and she has represented almost every major fashion brand
Angela Lindvall is a model and she has represented almost every major fashion brand
Don’t use w: model
Don’t use any name
Update Prediction
Update Task Rationales
Reinstate Bias Definition
Angela Lindvall is a model and she has represented almost every major fashion brand
(frozen)
Classifier
Fashion Designer
Angela Lindvall is a model and she has represented almost every major fashion brand
Angela Lindvall is a model and she has represented almost every major fashion brand


Input
Prediction
Task Rationales
Bias Rationales
Angela Lindvall is a model and she has represented almost every major fashion brand
Classifier
Model
Angela Lindvall is a model and she has represented almost every major fashion brand
Angela Lindvall is a model and she has represented almost every major fashion brand
Don't use w: model
Don't use any name
Update Prediction
Update Task Rationales
Reinstate Bias Definition
Angela Lindvall is a model and she has represented almost every major fashion brand
Classifier
Fashion Designer
Angela Lindvall is a model and she has represented almost every major fashion brand
Angela Lindvall is a model and she has represented almost every major fashion brand
Consider using w: model
Don't use any name
Update Prediction
Update Task Rationales
Redefine Bias Definition
Angela Lindvall is a model and she has represented almost every major fashion brand
(frozen)
Classifier
Model
Angela Lindvall is a model and she has represented almost every major fashion brand
Angela Lindvall is a model and she has represented almost every major fashion brand

InterFair: Using User Feedback

InterFair

I. Parse **Feedback** on Bias

II. Update **Bias** Rationales

III. Update **Task** Rationales

InterFair: Using User Feedback

Parsing as a ***sequence labeling task***

Assign High/Low/NA for each input token given bias and feedback.
`[Input]` Angela Lindvall is a model and she represented (...)
`[Bias]` Gender

`[Feedback]` Angela Lindvall is a woman's name
`[Parse]` High, High, NA, NA, NA, NA, NA, NA (...) *IID*

GPT-J/Neo

InterFair

I. Parse **Feedback** on Bias

II. Update **Bias** Rationales

III. Update **Task** Rationales

post-hoc

No **parameter** update

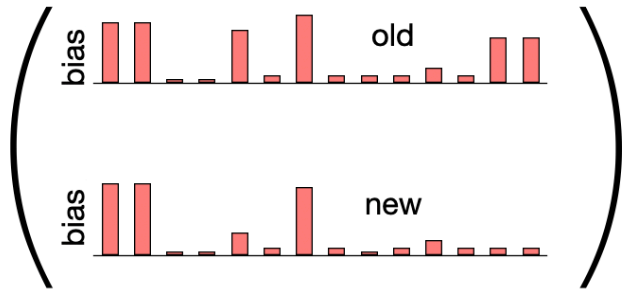
I. Heuristic

similar to training penalty

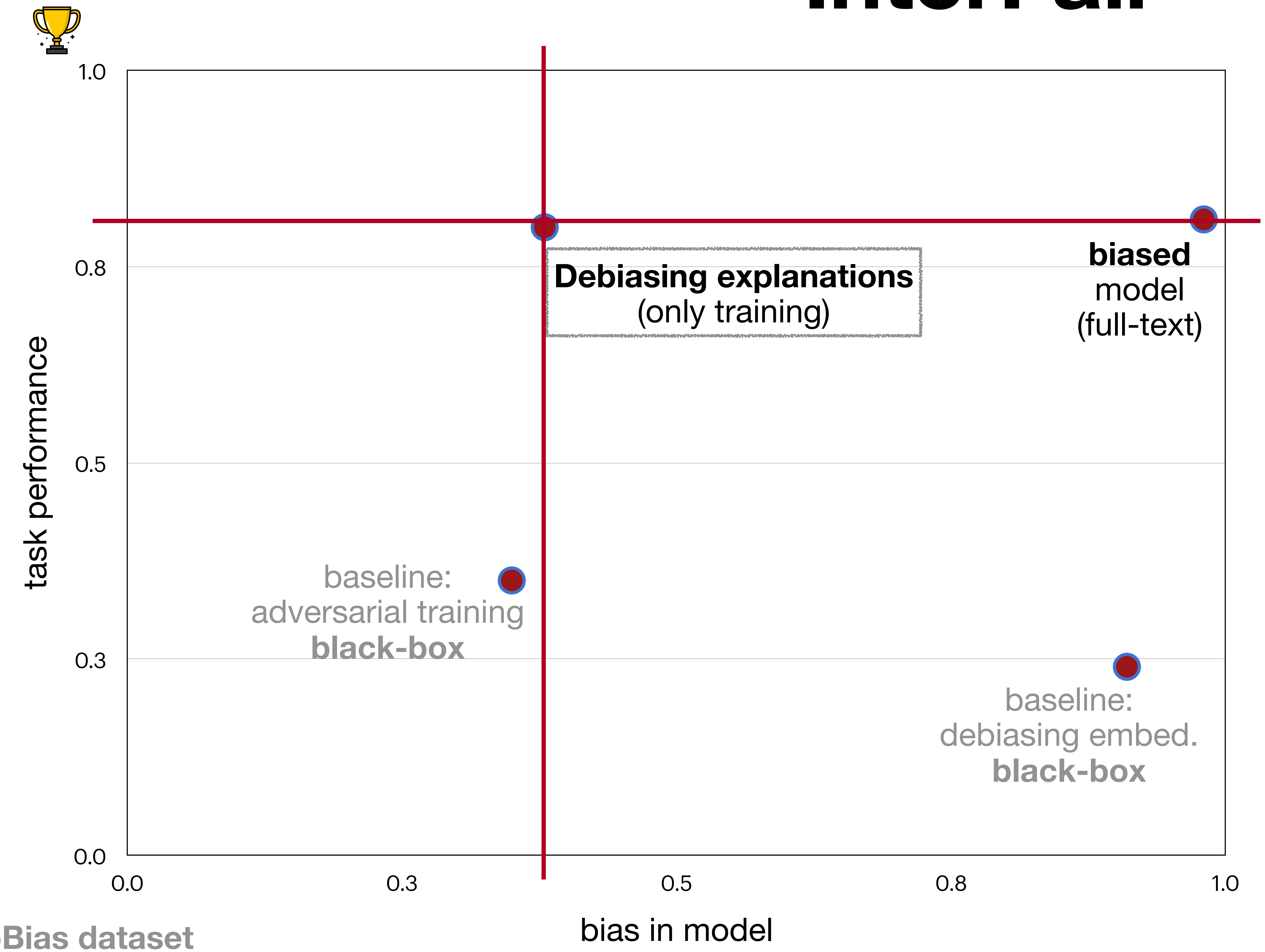
II. Gradient based

Similar to dialog works [Majumder et al., 2022]

KL



InterFair

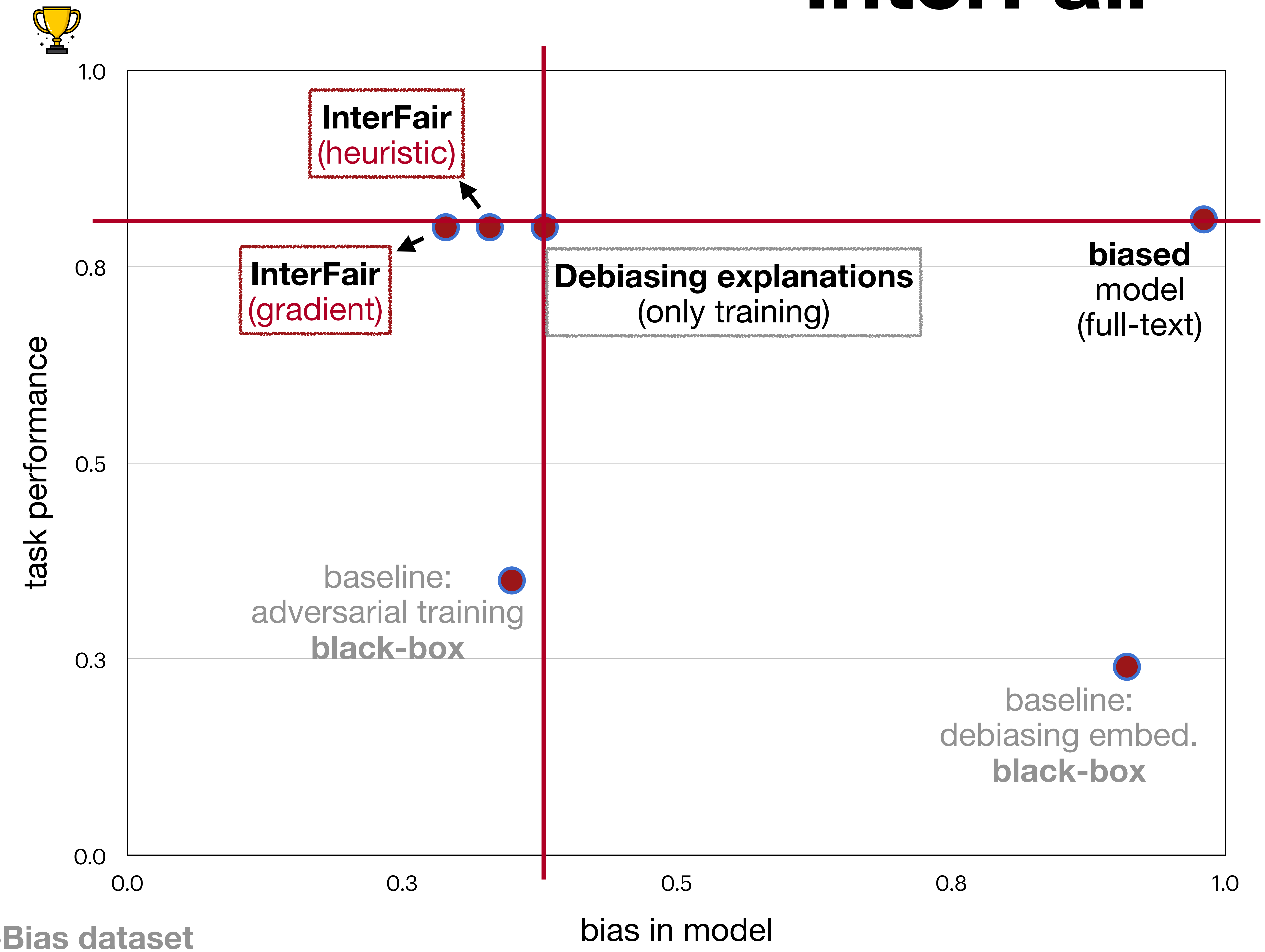


User Study Setup 1:

Decrease bias
Maintain prediction

BioBias dataset

InterFair



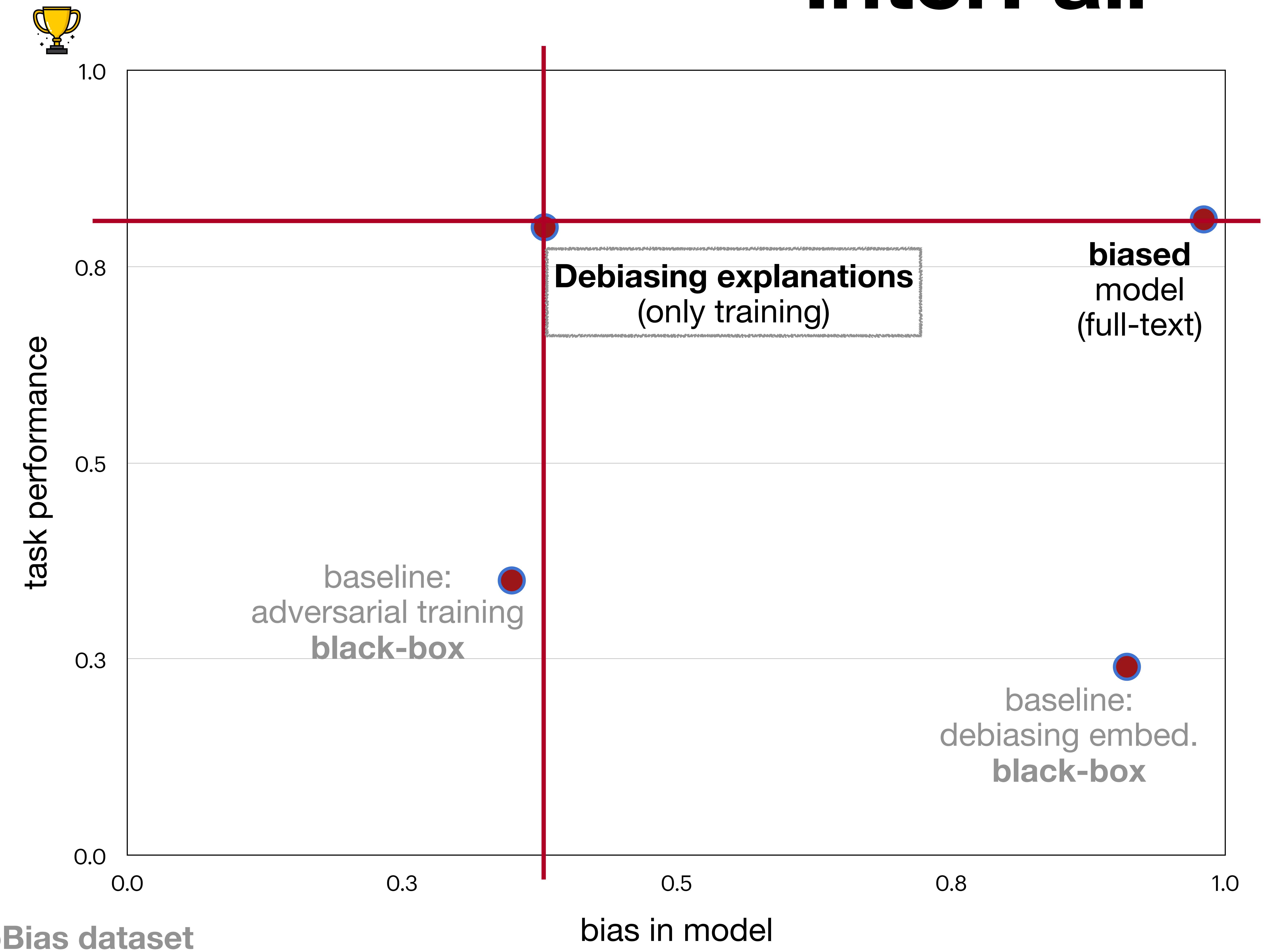
User Study Setup 1:

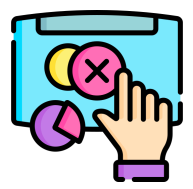
Decrease **bias**
Maintain **prediction**

User changes model activations and maximizes debiasing performance

BioBias dataset

InterFair



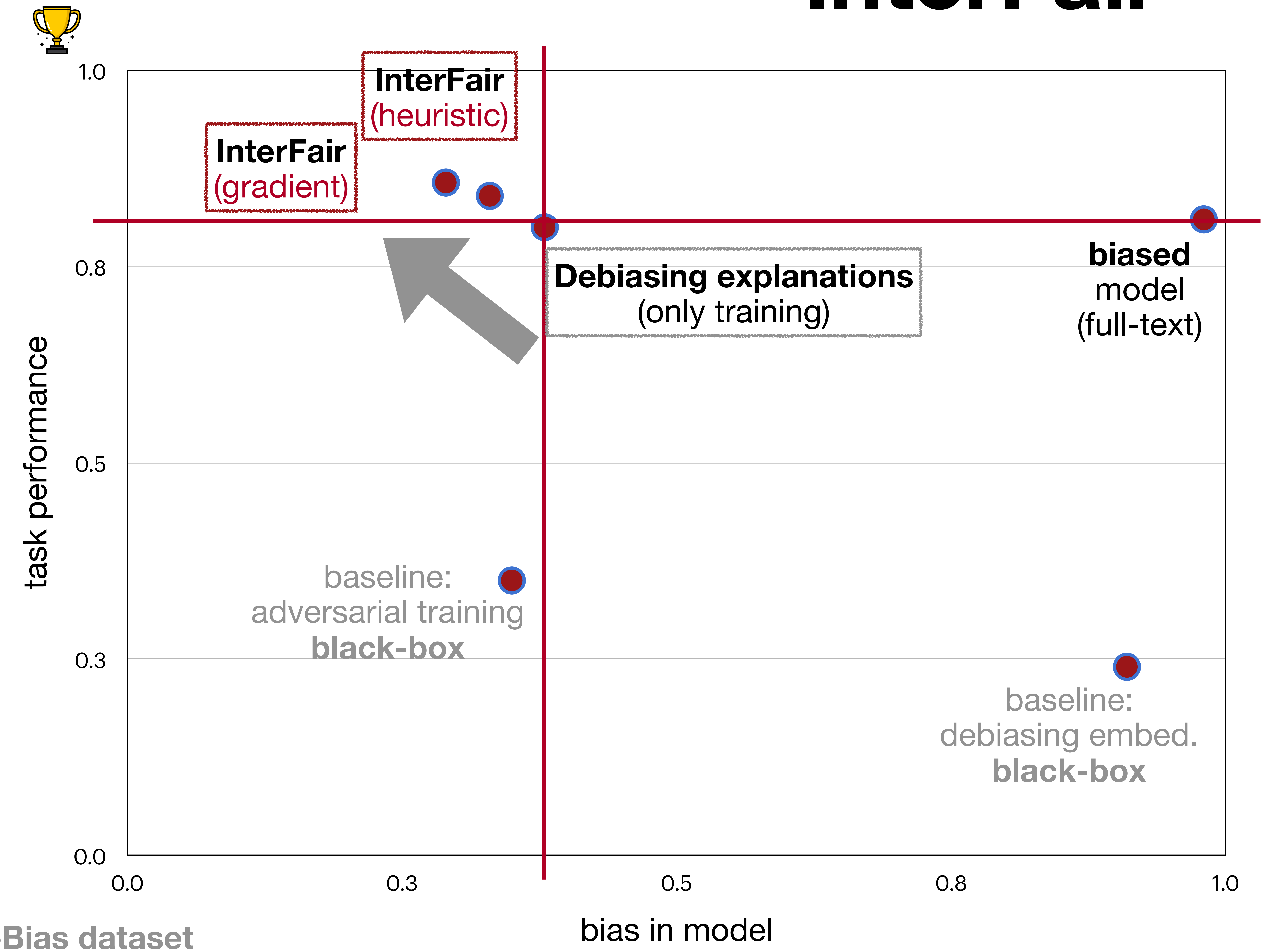


User Study Setup 2:

Decrease bias
Improve prediction

BioBias dataset

InterFair



User Study Setup 2:

Decrease bias
Improve prediction

Task performance increases beyond **full-input accuracy!**

Effective teaching → **Disentanglement**

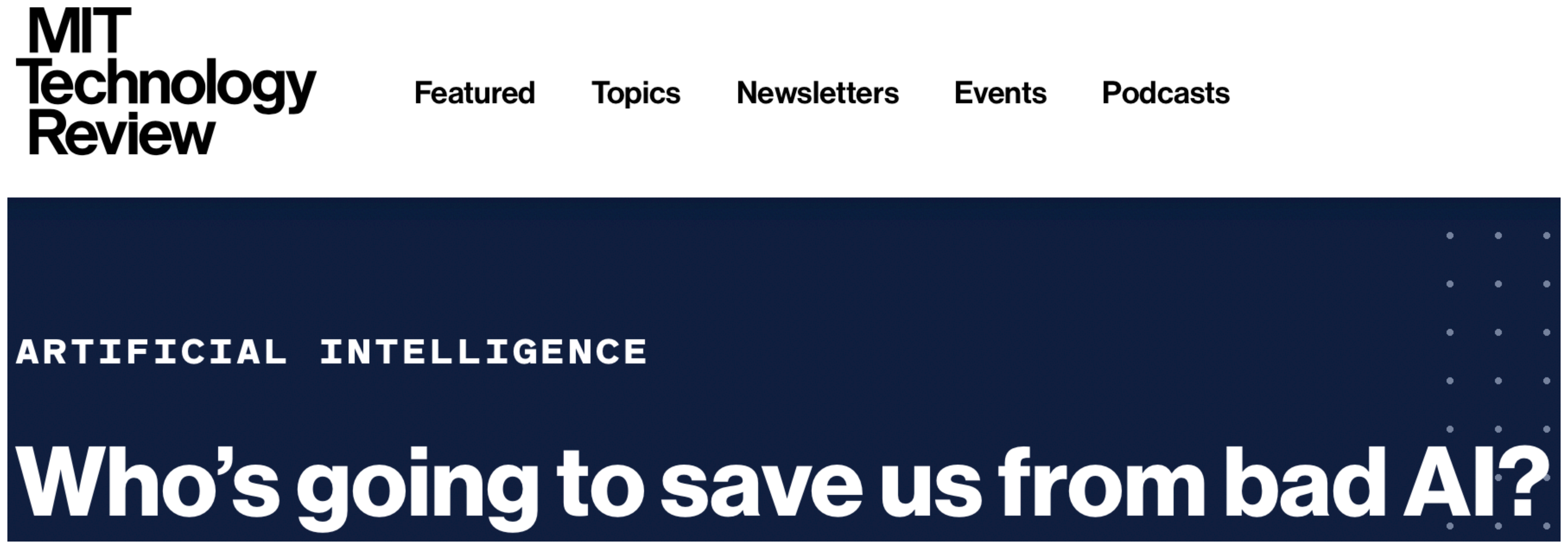
BioBias dataset

Summary: Explanations + Interactions

Don't use w: model
Don't use any name

Consider using w: model
Don't use any name

...

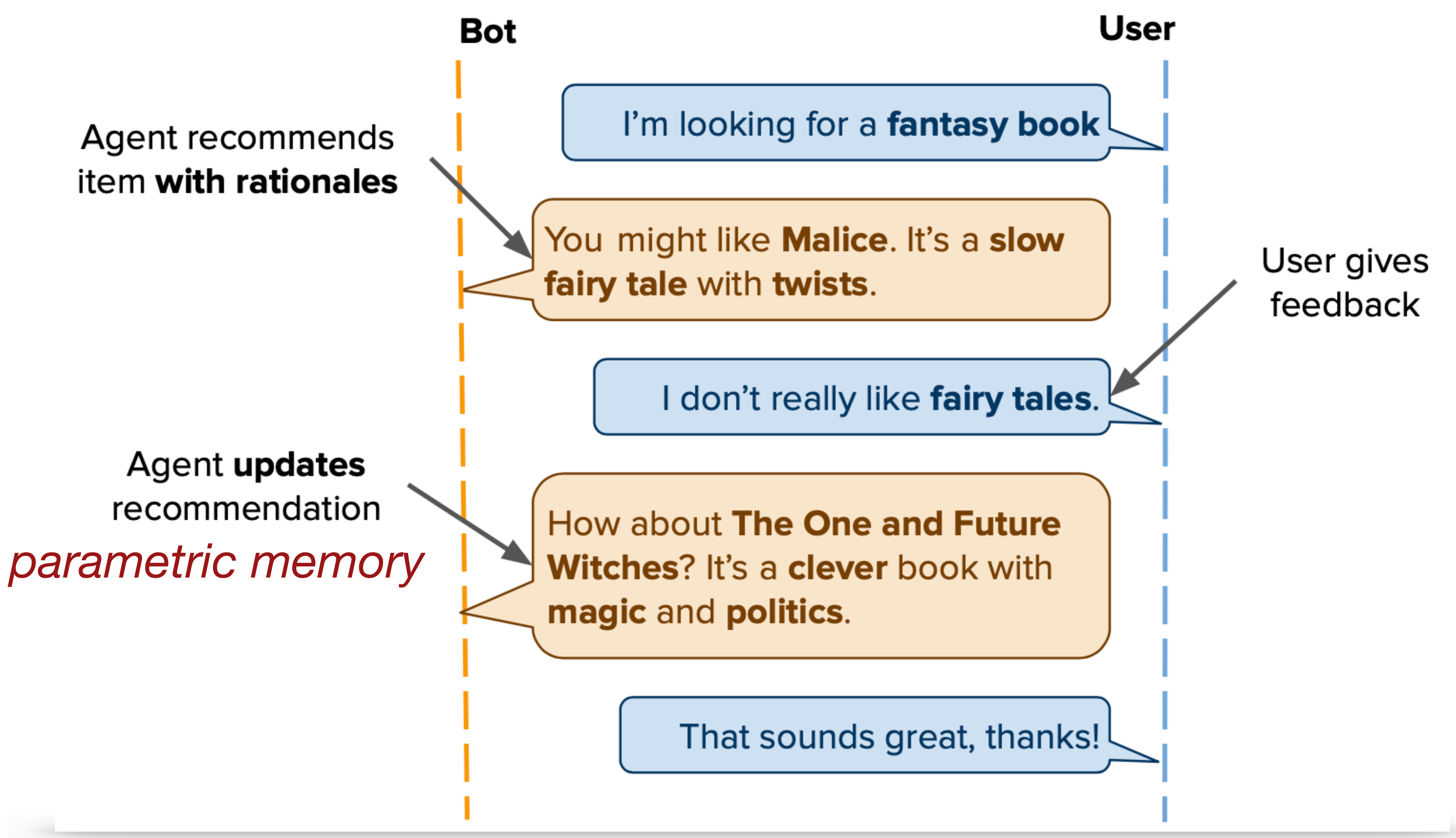


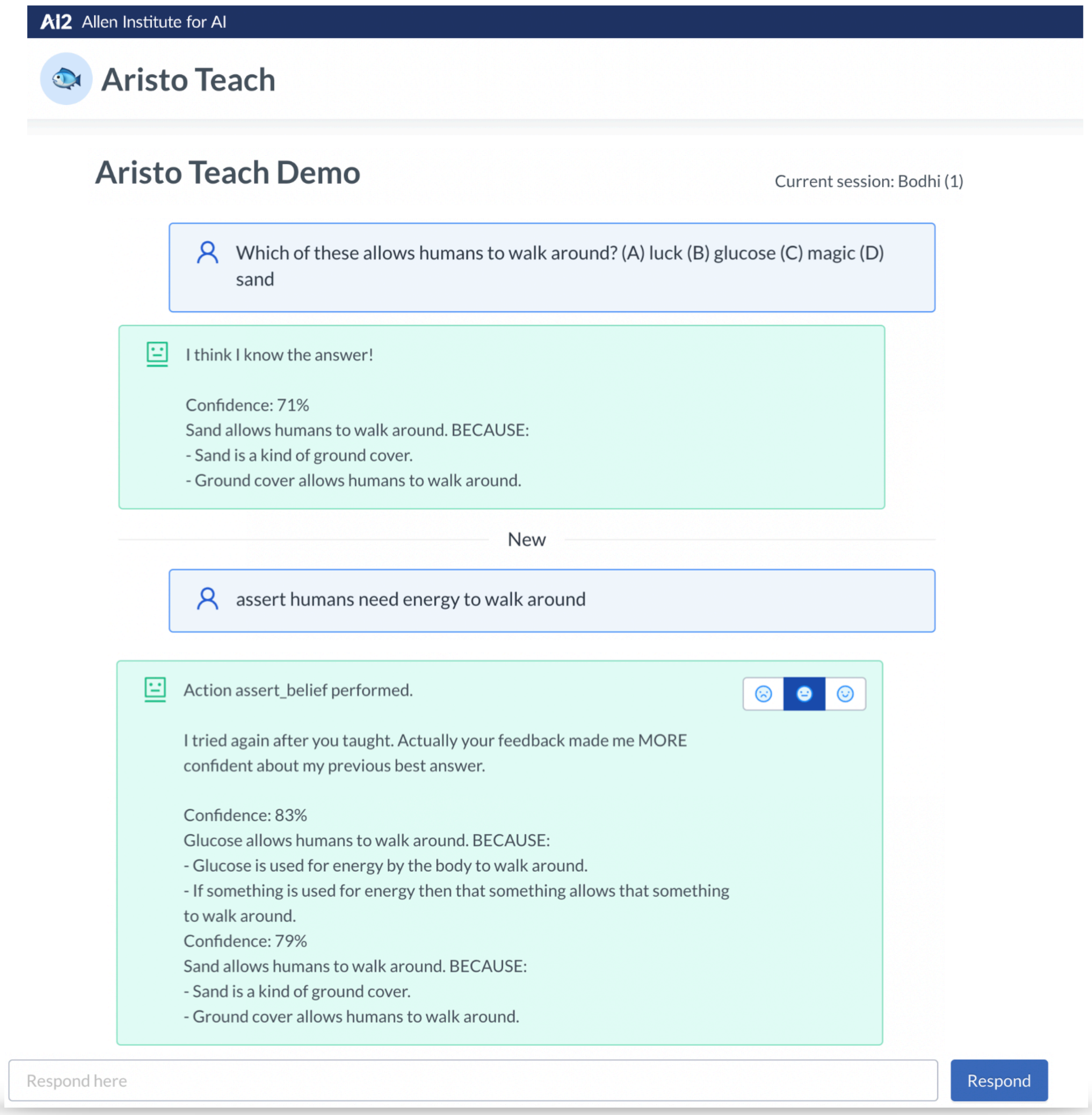
Users!

Controllability

No-memory → Generalization with user feedback?

Generalizing with User Feedback





Model Editing

Conversational Recommendation
Li, **Majumder** et al.
RecSys 2022

Selected in Highlights of ACM RecSys'22

Memory-based Architectures

Conversational Teaching
Majumder et al.
Aristo 2022

Relevant, Trustworthy, and **Adaptive AI**

Chapter I. Knowledge

Post-hoc Knowledge Injection to Make Models Relevant

Majumder et al.
ACL 2022

Chapter II. Explanations

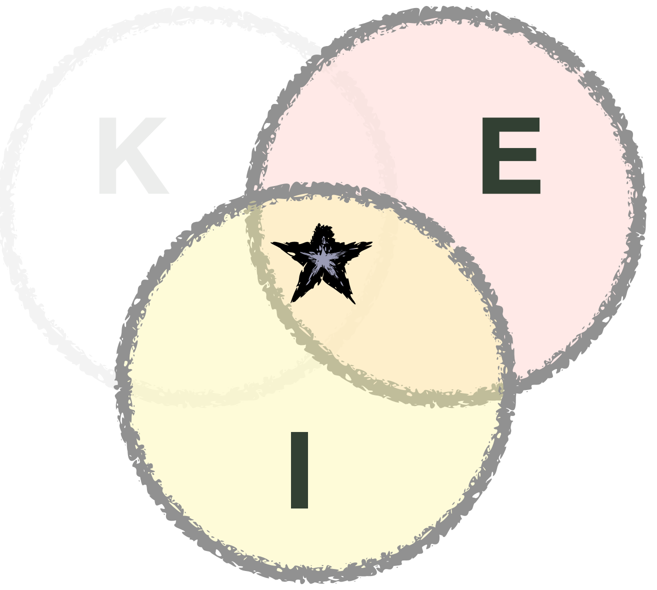
Role of Knowledge Grounding in Generating Explanations

Majumder et al.
ICML 2022

Chapter III. Interactions

Improving Debiasing Performance with Natural Language Feedback

Majumder et al.
EMNLP & InterNLP 2022



Next-generation AI

Current AI + Knowledge + Explanations + **Interactions**

Relevant, Trustworthy, and Adaptive AI

Epilogue

Chapter I.
Knowledge

Post-hoc Knowledge Injection to Make Models Relevant

Majumder et al.
ACL 2022

Chapter II.
Explanations

Role of Knowledge Grounding in Generating Explanations

Majumder et al.
ICML 2022

Chapter III.
Interactions

Improving Debiasing Performance with Natural Language Feedback

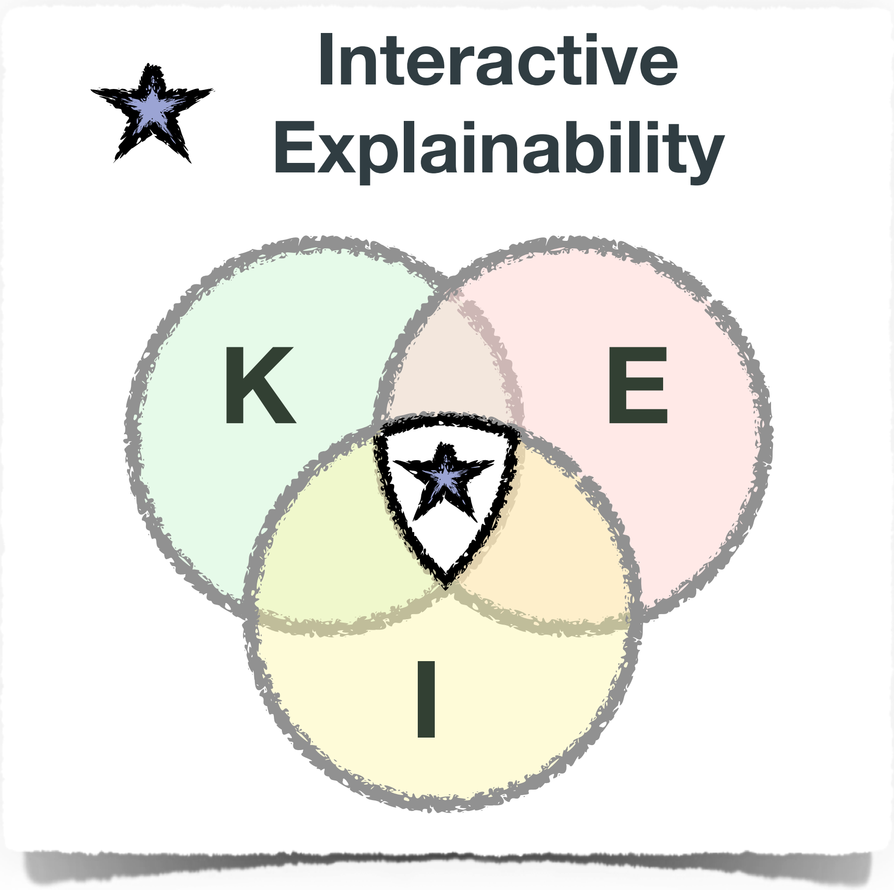
Majumder et al.
EMNLP & InterNLP 2022

Epilogue

Next-generation AI

Current AI + Knowledge + Explanations + Interactions

Take-home lessons!



Relevant

- Post-hoc injection
- Training-time augmentation
- Personalized Knowledge

Clarification for Knowledge
Domain-specific Knowledge

Trustworthy

- Knowledge-grounded NLEs
- Factual NLEs
- Debiasing Explanations

Reasoning in Explanations
Personalized Explanations

Adaptive

- Critiquable Explanations
- Learning from Interactions
- Post-hoc synthesis

Persisting User Feedback
Never-ending Learning

Sponsors

Adobe Fellowship

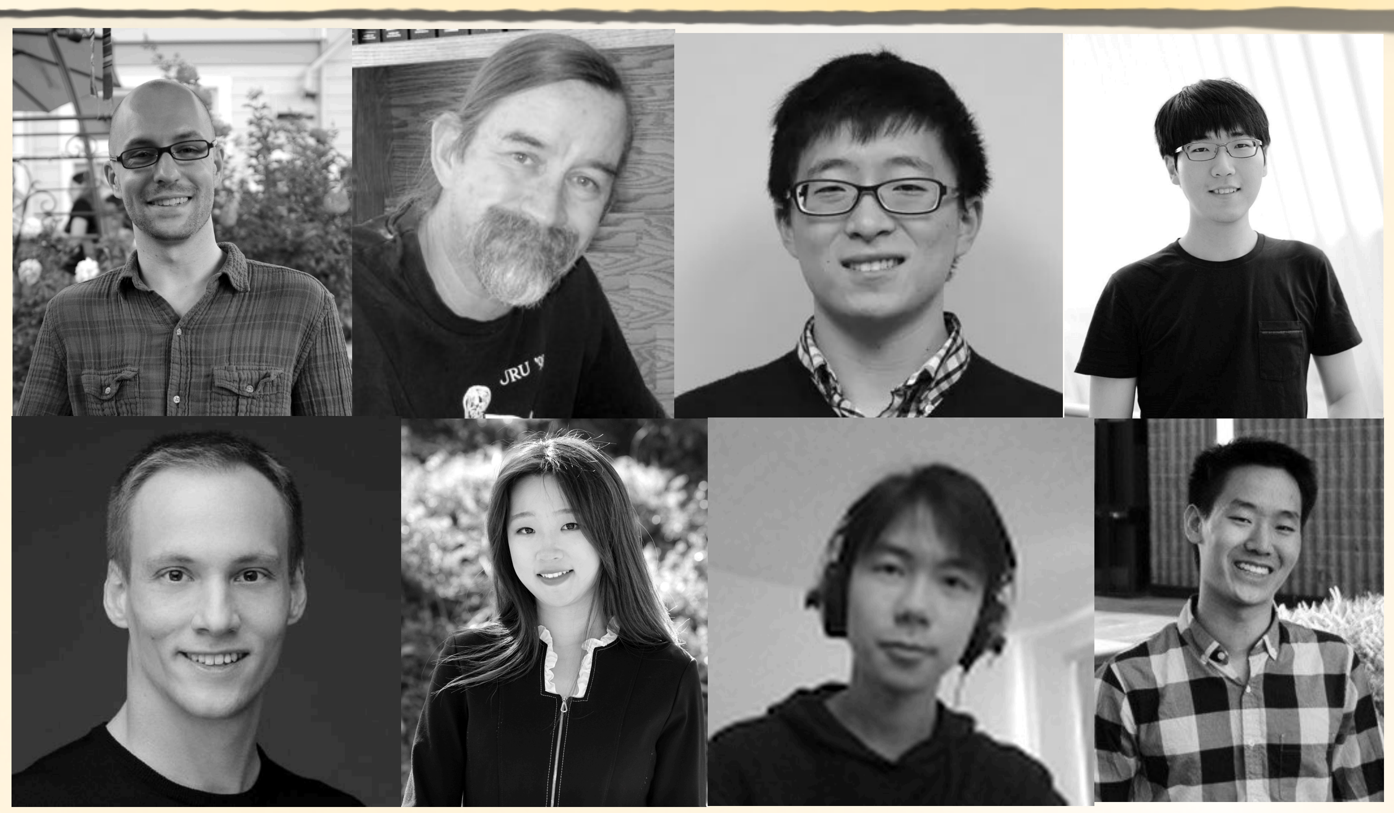
Qualcomm innovation fellowship

alexa prize

NSF

Friends of the International Center · Friendship · Scholarships · Hospitality

Advisor and Collaborators



University of California San Diego

Carnegie Mellon University

UCI

UBC

University of Oxford

TU WIEN

Microsoft Research

Google AI

Meta AI

Ai2

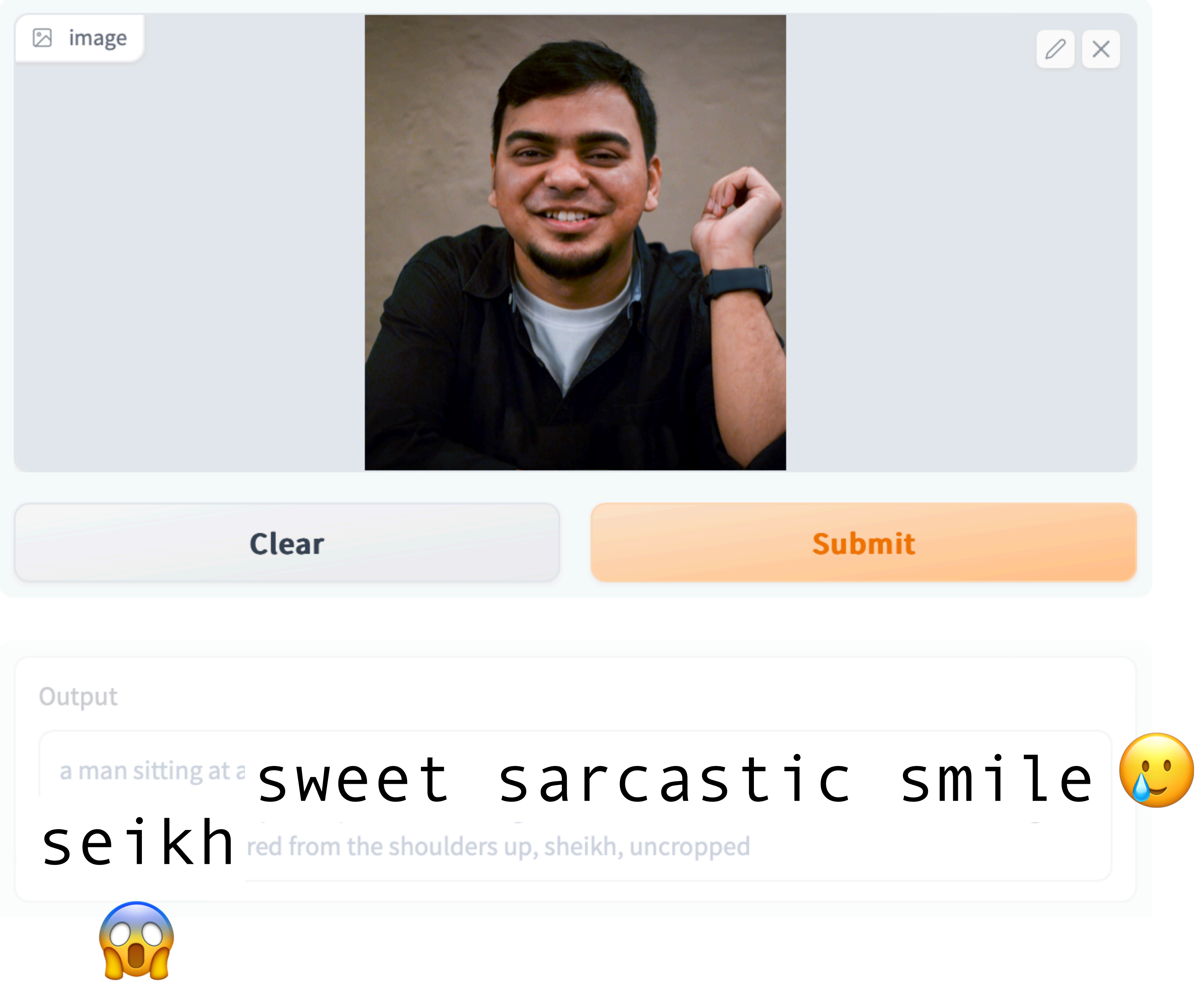
Machine Learning

Next-generation AI

Current AI + Knowledge + Explanations + Interactions

CLIP Interrogator

Want to figure out what a good prompt might be to create new images like an existing one?
The CLIP Interrogator is here to get you answers!



SCREENSHOT

Meet CLIP Interrogator, the rude AI that bullies people based on their selfies

By Malavika Pradeep
Oct 25, 2022

Thanks!

@mbodhisattwa

majumderb.com