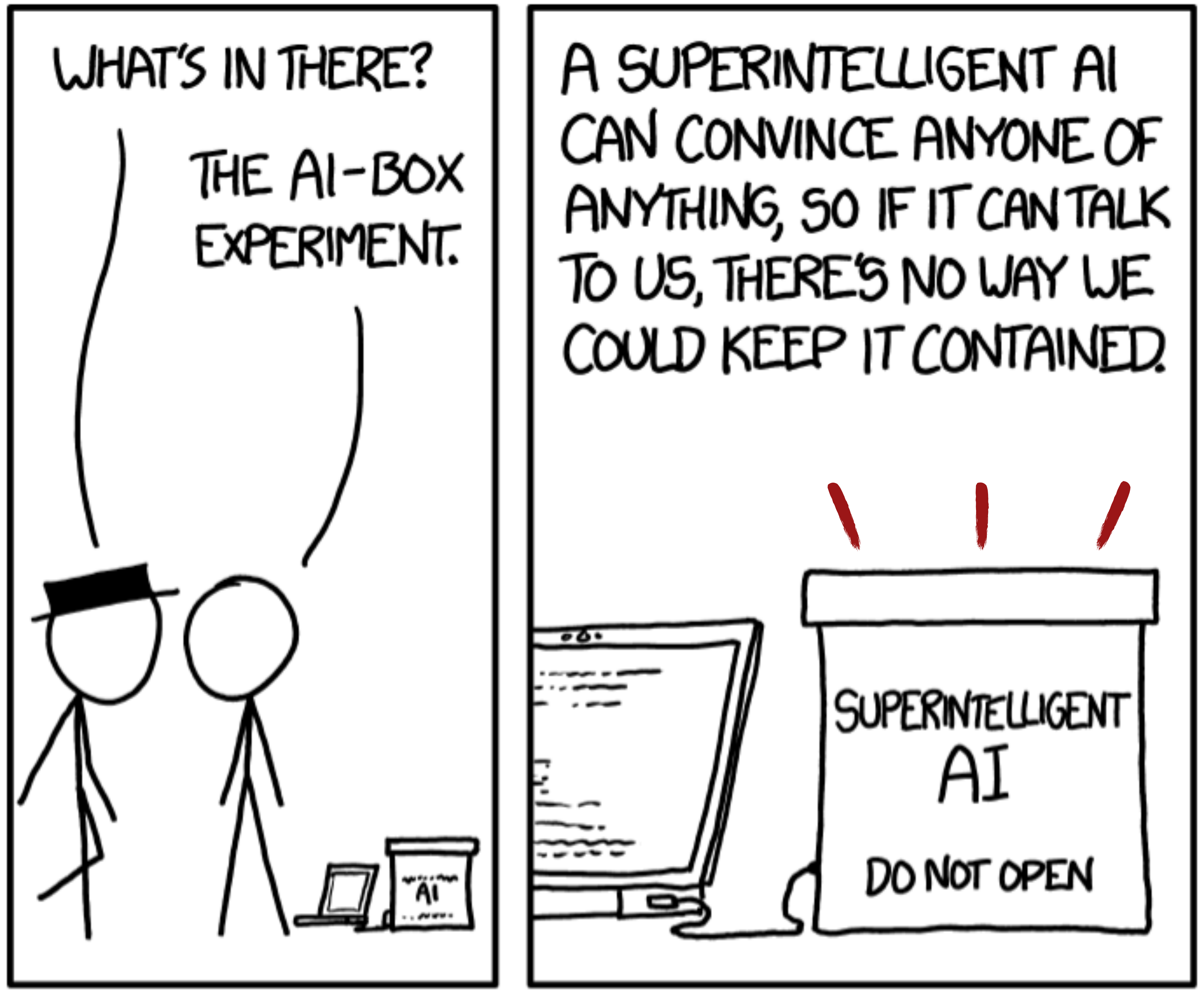
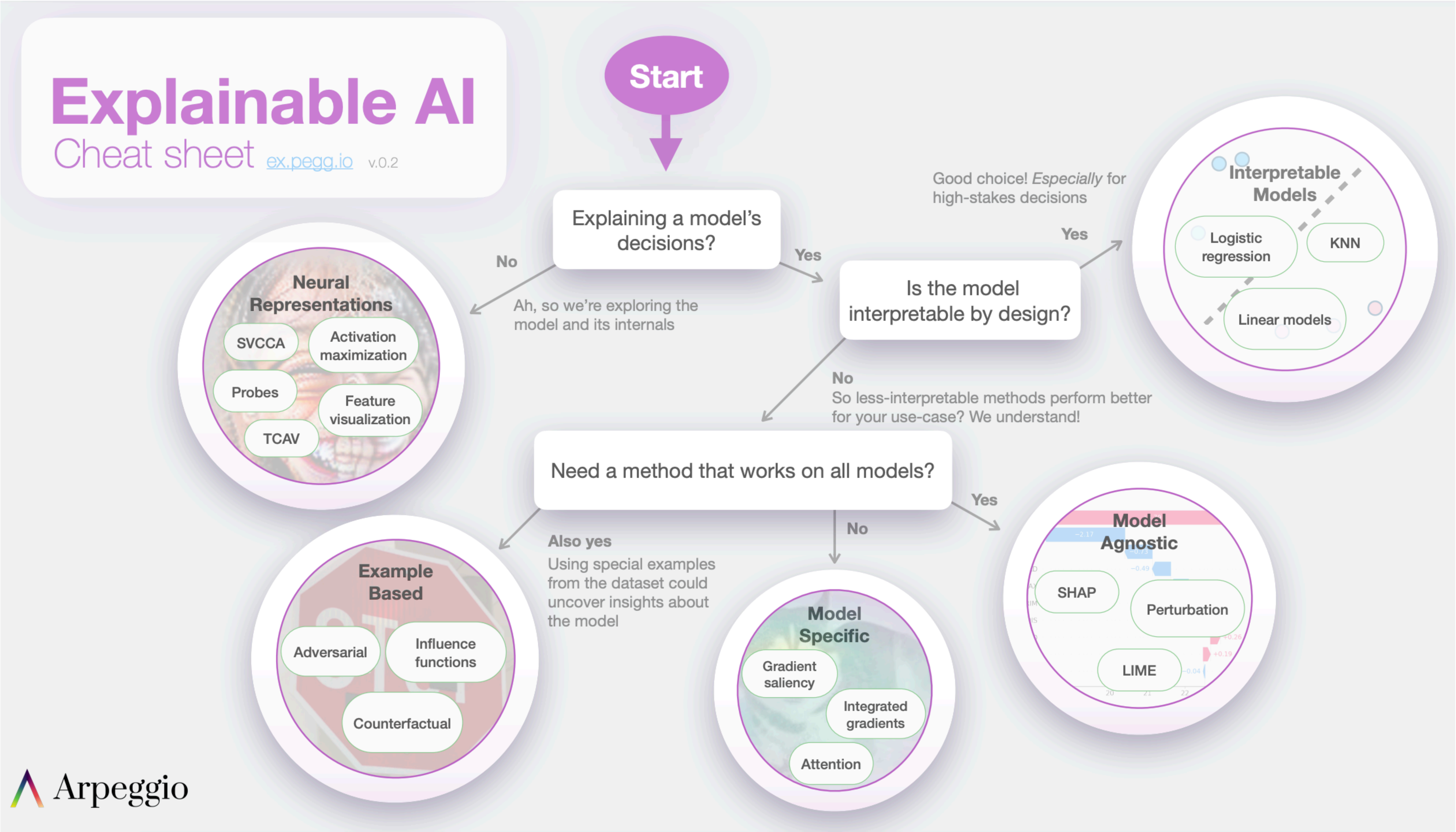
Producing Explanations with Commonsense and Interactions

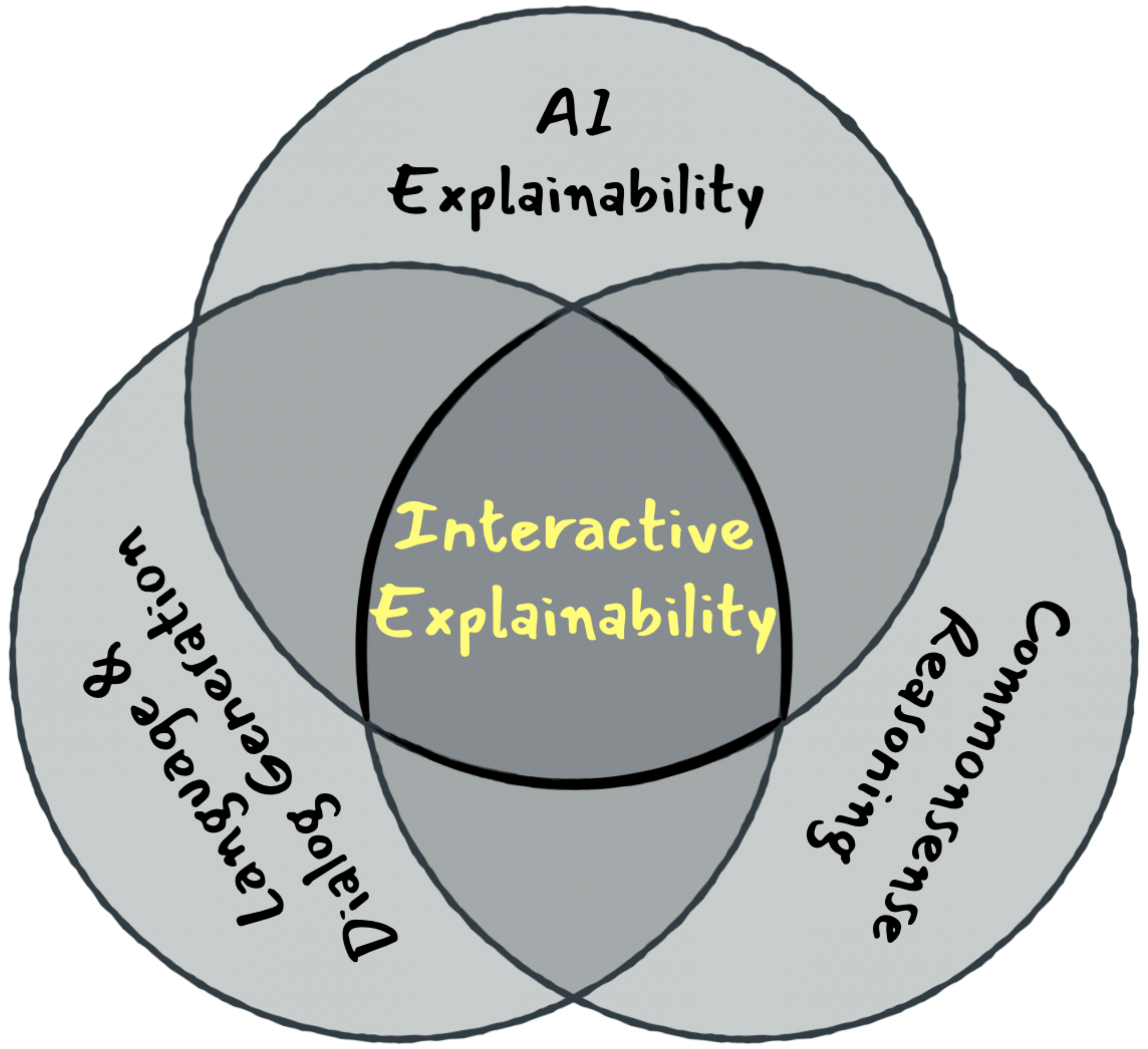


Bodhisattwa Prasad Majumder
@mbodhisattwa
UC San Diego

Allen Institute of AI, 2021

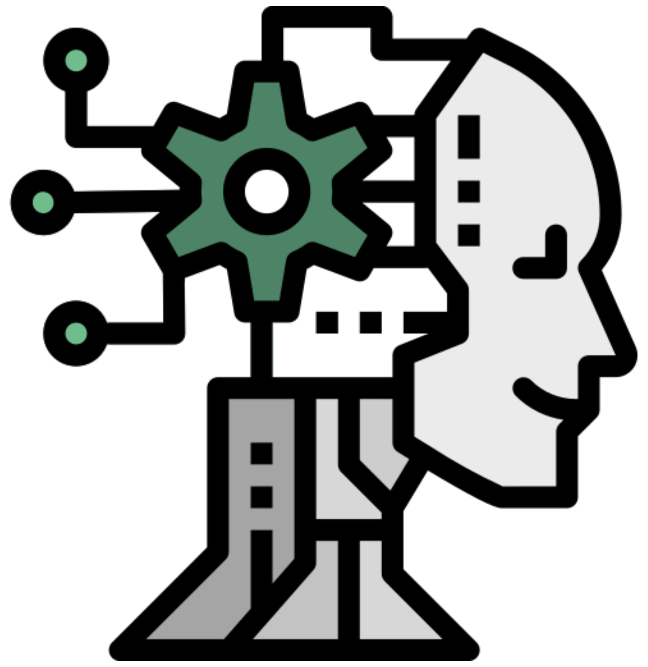
Explainability in AI



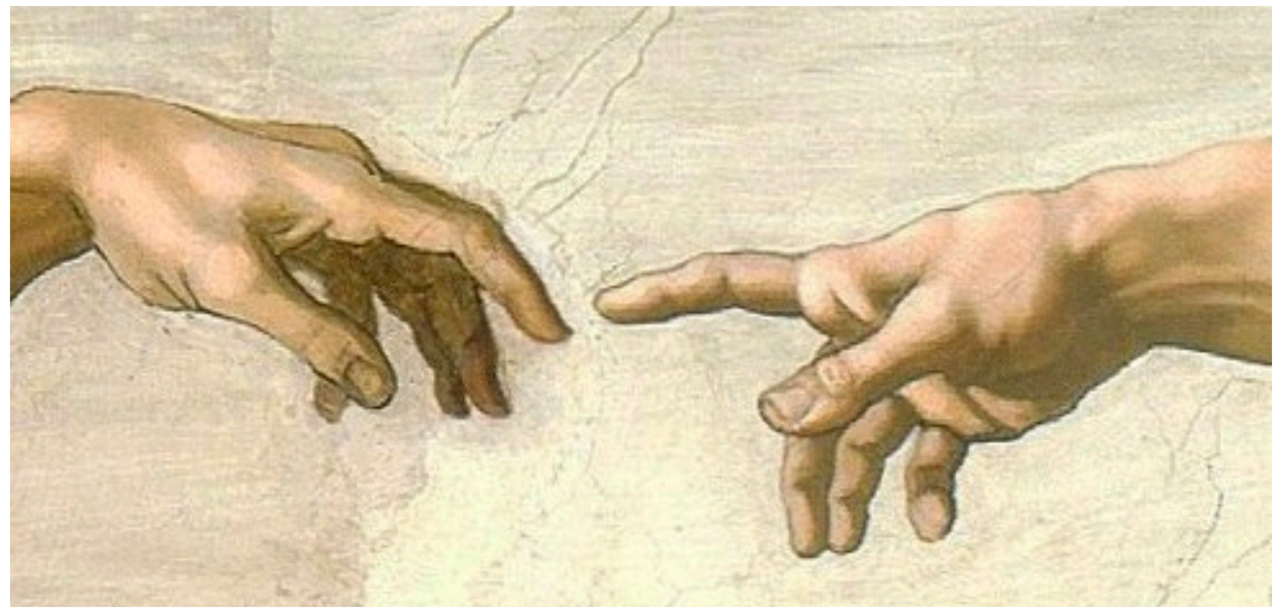


https://jalammar.github.io/explainable-ai/

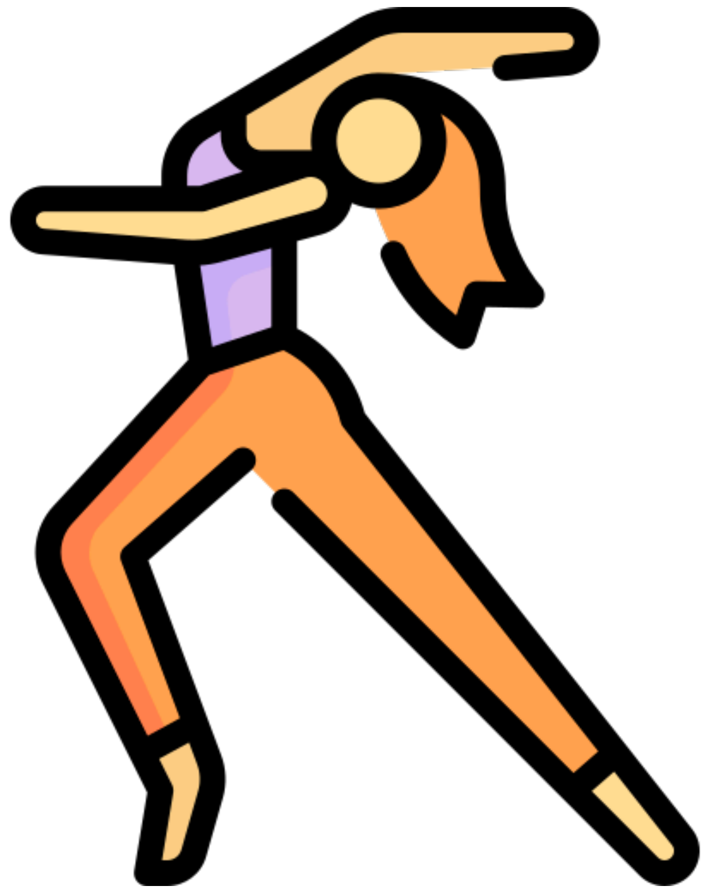
Explanations with Commonsense and Interactions

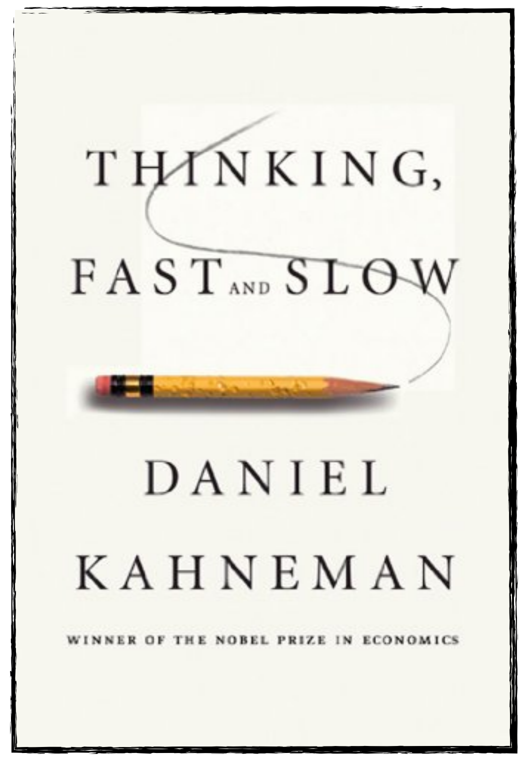


Conceptual Mappings with World Knowledge



Natural Language Feedback

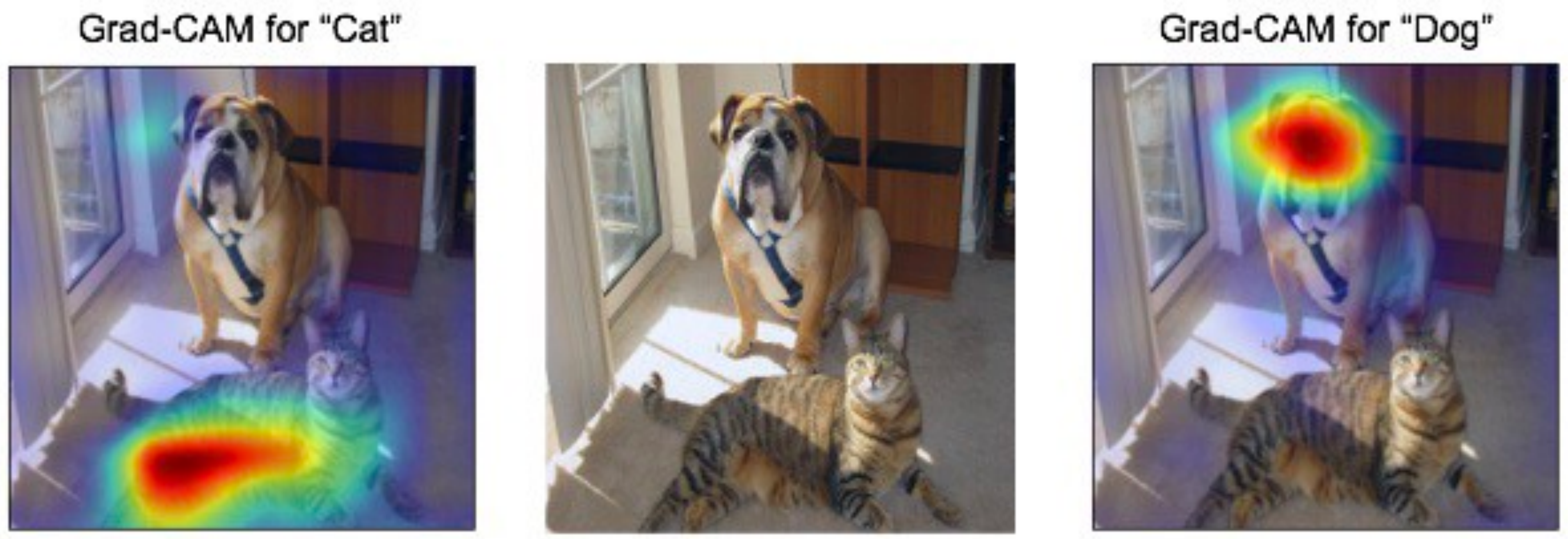




{perception, intuition, reasoning}

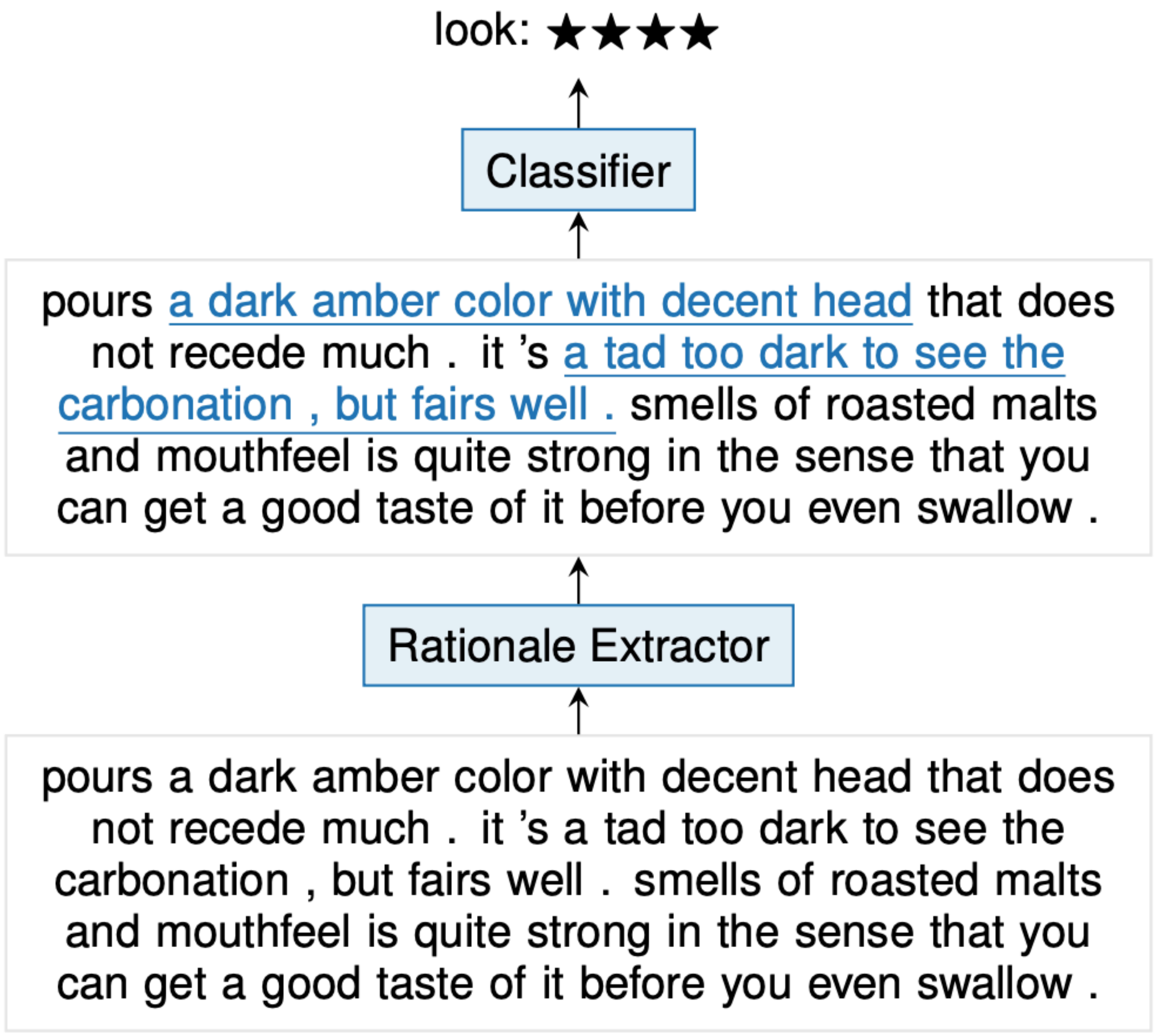
{perception, intuition, reasoning}

User Experience with AI Explanations

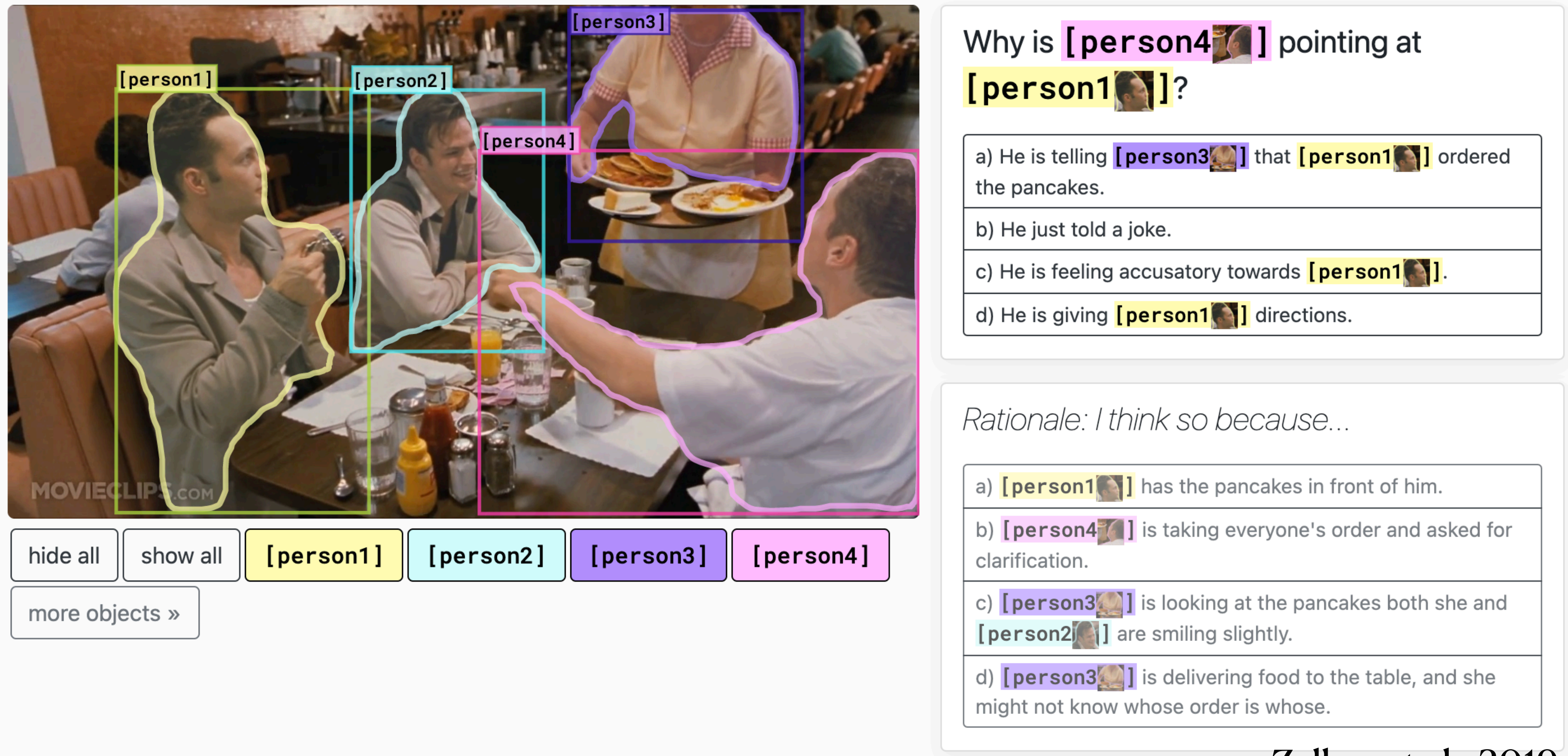


Selvaraju et al., 2019

https://jalammar.github.io/explainable-ai/

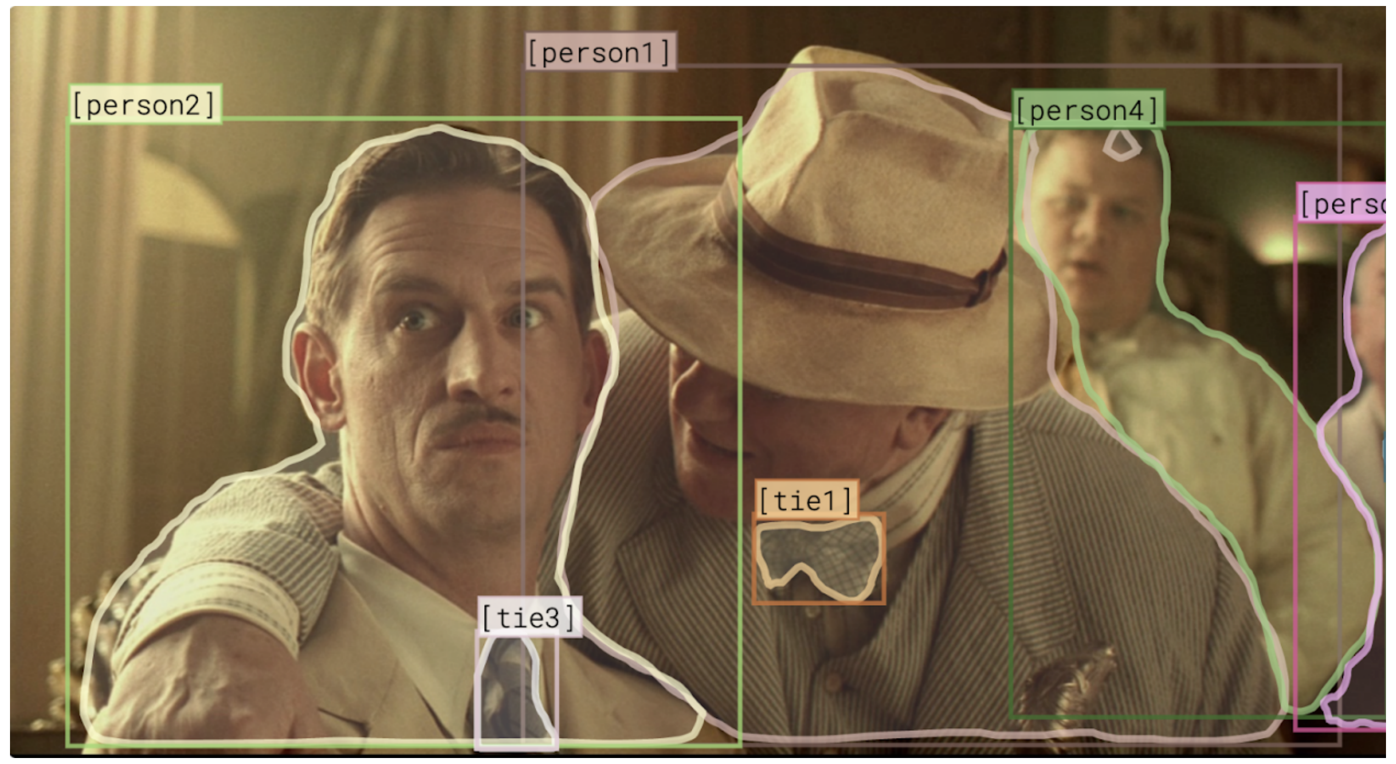


Bastings et al., 2020



Zellers et al., 2019

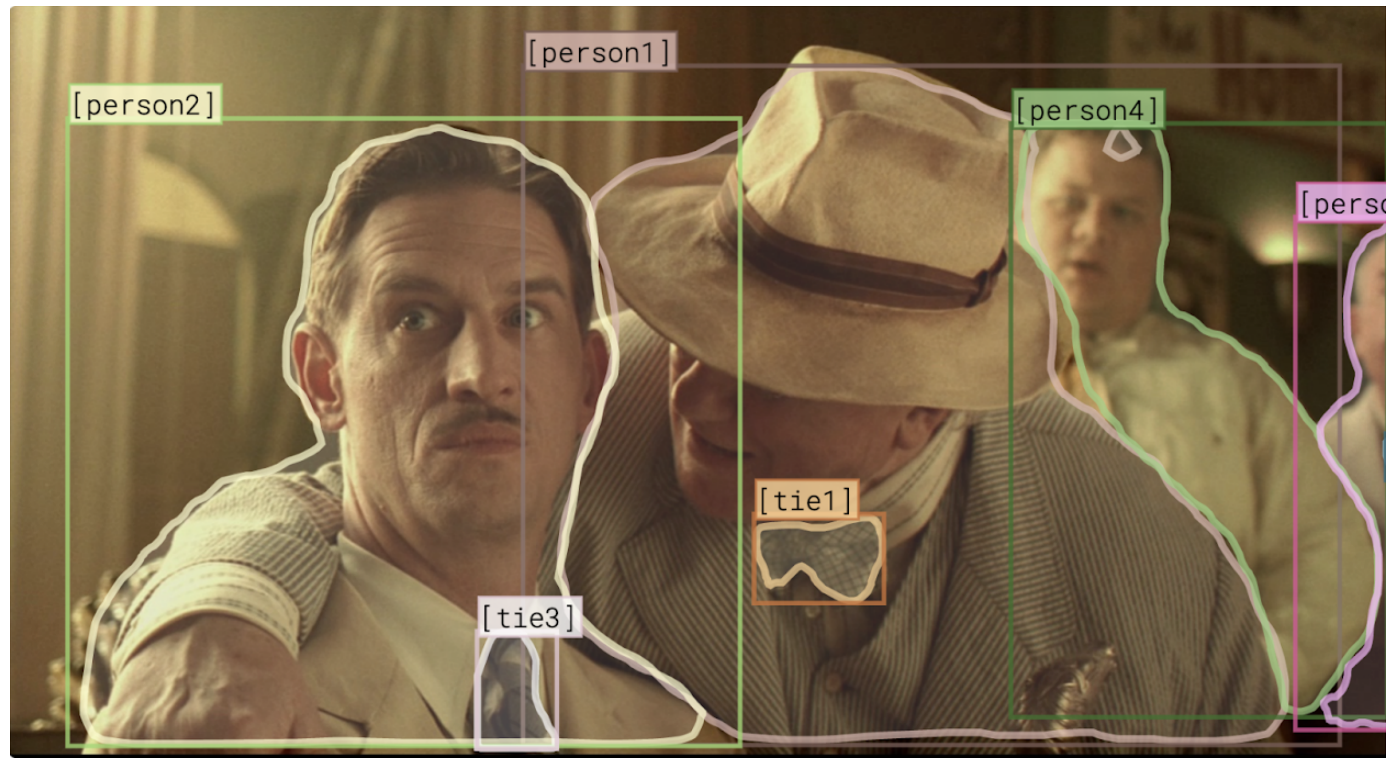
Rich Representation of Explanations



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

A: He's concerned and a little upset

extractive



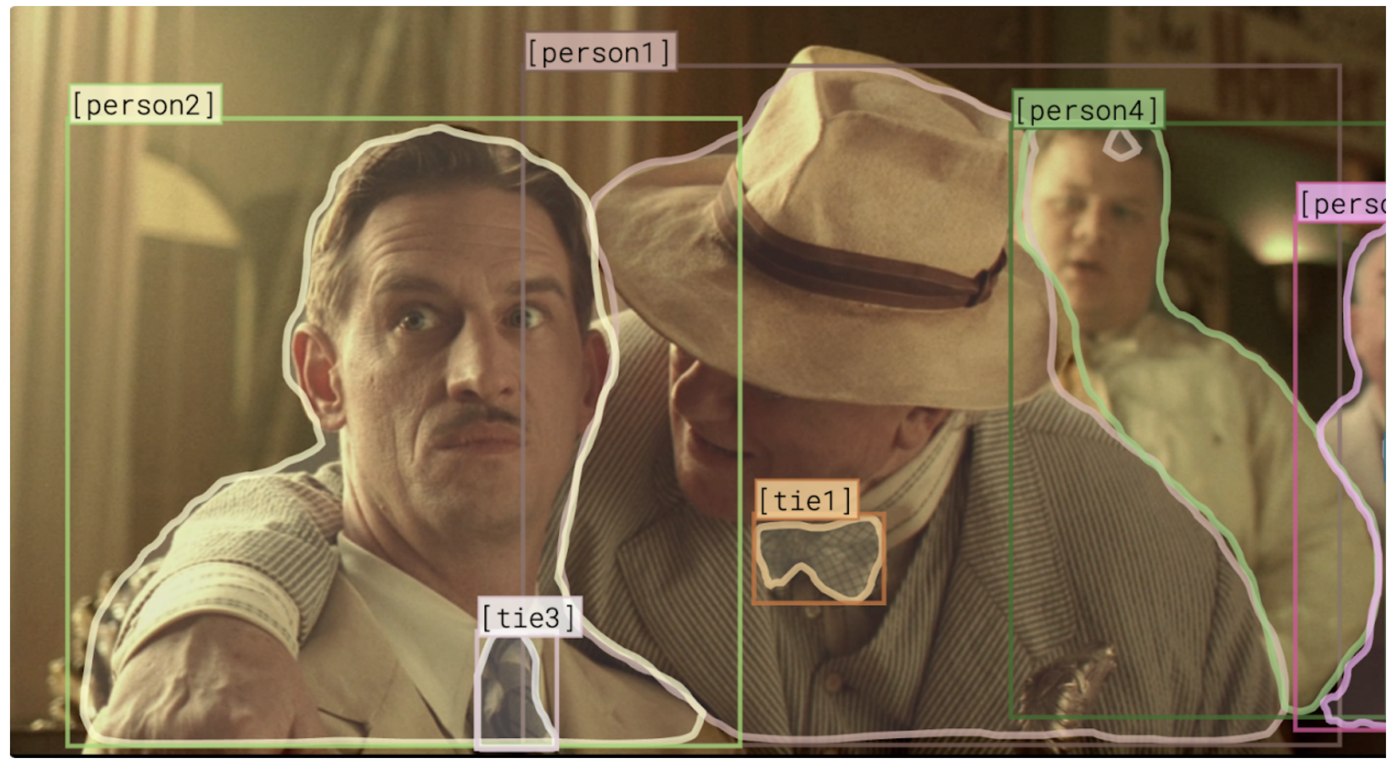
Q: how does `[person2]` feel about what `[person1]` is telling him?

A: He's concerned and a little upset

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

abstractive

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.

abstractive

- NLE should be fluent and consistent to the input
- NLE should accurate to explain the prediction
- NLE should be grounded in to **world knowledge (*aka* commonsense)**

Why do we need Commonsense?

Why do we need Commonsense?

Language Modeling:
Barrack's wife is Hillary
The capital of India is the city
St. Louis is a city in the state of Oldham

Dialog Generation:
Bot: Today, I went to the central park with my dog.
User: I am not an animal lover.
Bot: Me too. I don't have a pet.

Story Generation:
Harry shot Leo and tried to run away. The night was dark and scary. (…) Harry invited Leo for dinner.

Why do we need Commonsense in NLEs?

Lack of **commonsense** grounding leaves models prone to adversarial attacks

PREMISE: A guy in a red jacket is snowboarding in midair.	
ORIGINAL HYPOTHESIS: A guy is outside in the snow.	REVERSE HYPOTHESIS: The guy is outside.
PREDICTED LABEL: entailment	PREDICTED LABEL: contradiction
ORIGINAL EXPLANATION: **Snowboarding is done outside.**	REVERSE EXPLANATION: **Snowboarding is not done outside.**

Camburu et al., 2020

😀 🤔

Rationale-Inspired Natural Language Explanations with Commonsense

Bodhisattwa Prasad Majumder[1], Oana-Maria Camburu[2], Thomas Lukasiewicz[2,3], Julian McAuley[1]
[1]UC San Diego, [2]University of Oxford, [3]Alan Turing Institute




Natural Language Inference

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

input →

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

***extractive* rationales (highlighted)** →

label: entailment

Competing in a bicycle race requires riding bikes

***abstractive* NLE**

Natural Language Inference

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

- requires bikes
- requires riding bikes
- requires helmet
- is a outdoor game

label: entailment

Competing in a bicycle race requires riding bikes

input → *extractive* rationales (highlighted) → commonsense → *abstractive* NLE

Extractive Rationales, Natural Language Explanations and Commonsense

Natural Language Inference

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

- requires bikes
- requires riding bikes
- requires helmet
- is a outdoor game

label: entailment

Competing in a bicycle race requires riding bikes

input → ***extractive* rationales (highlighted)** → **commonsense** → ***abstractive* NLE**

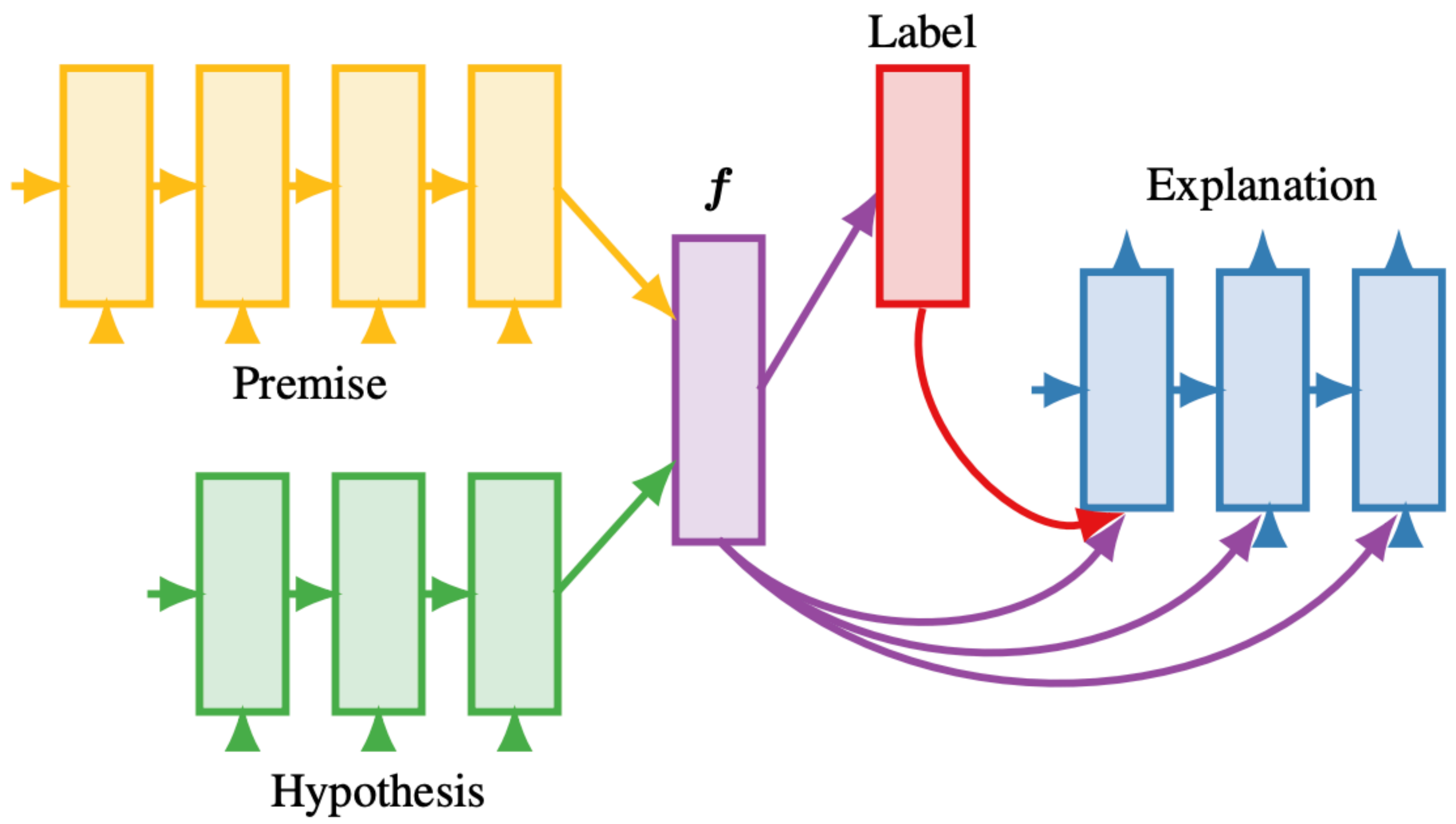
Extractive Rationales, Natural Language Explanations and Commonsense

REXC

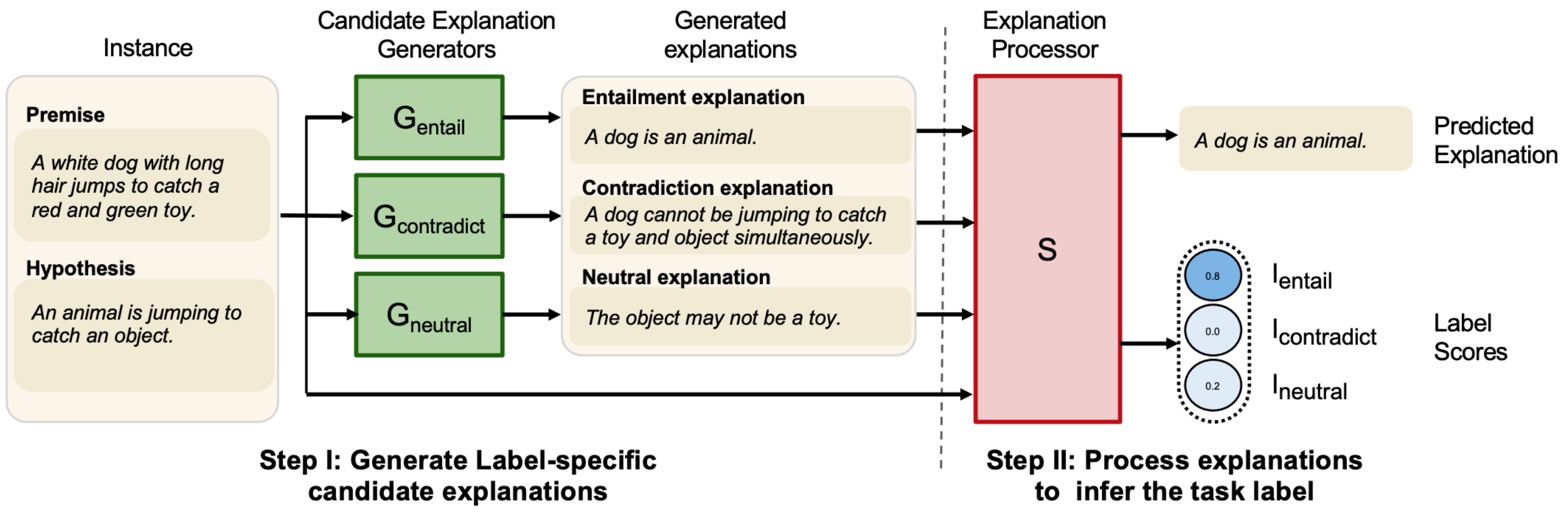
Our Goals

- How can we **link** extractive rationales to abstractive explanations?
- How do we **incorporate commonsense** knowledge for more accurate and sensible explanations?
- How can we use commonsense knowledge as **supporting evidence** behind the generated explanations?

Previous Works

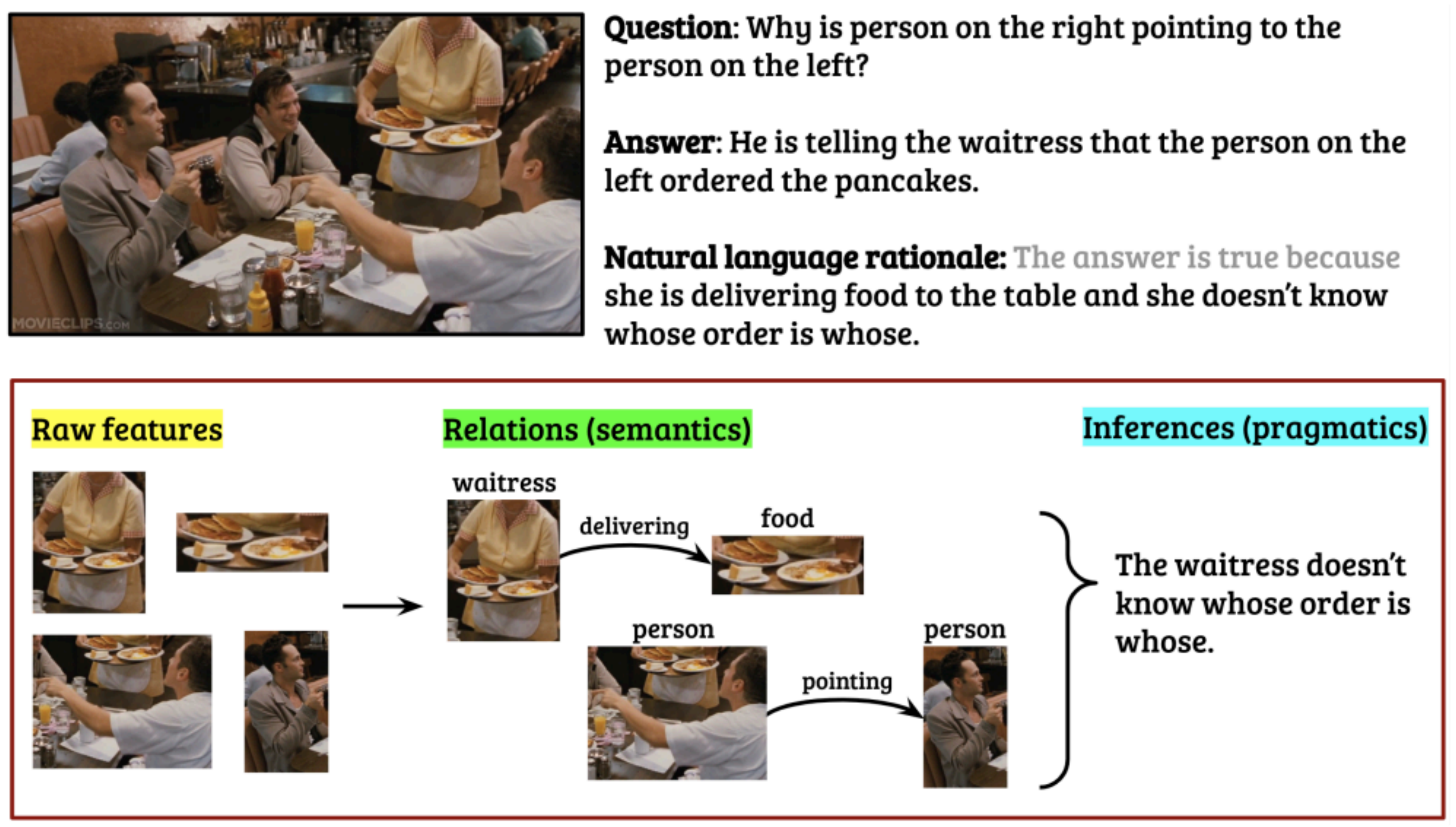


predict-then-explain (Camburu et al., 2018)



generate label-specific explanations, then choose the correct one

(Kumar et al., 2018)



stacked steps of feature extraction, selection, commonsense inference

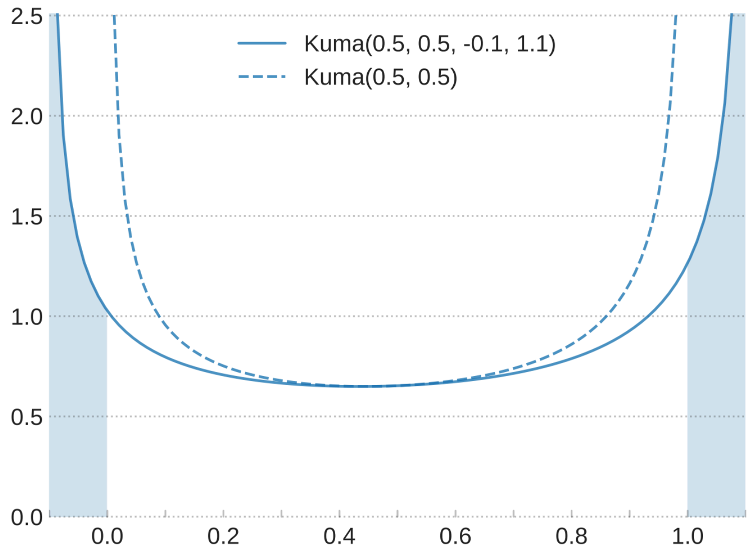
(Marasovic´ et al., 2018)

RExC

Input is passed to Neural Rationale Extractor $\mathscr{R}$

RExC

A series of binary latent variables z_i^r are used to discretely select parts of the input as *rationales*



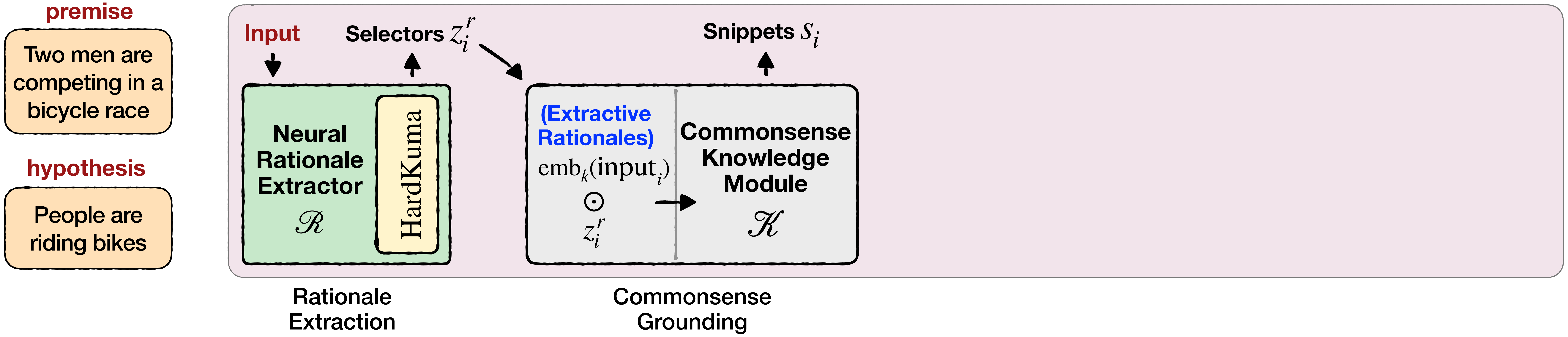
Bastings et al., 2020

L_1 regularization for sparsity

RExC

- requires bikes
- requires riding bikes
- requires helmet
- is a outdoor game

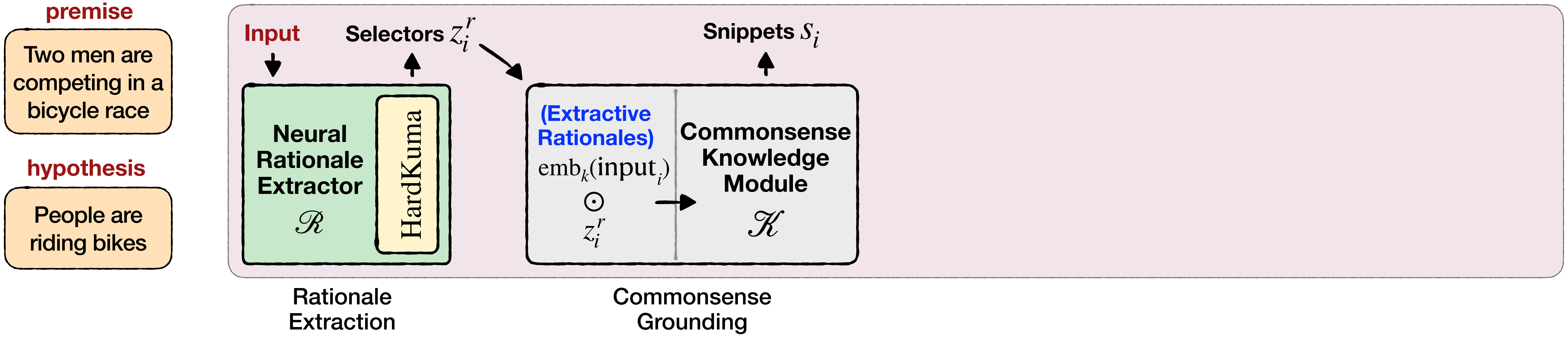
Each lexical unit from rationales are sent to the commonsense module $\mathscr{K}$, that result in knowledge snippets s_i



RExC

- requires bikes
- requires riding bikes
- requires helmet
- is a outdoor game

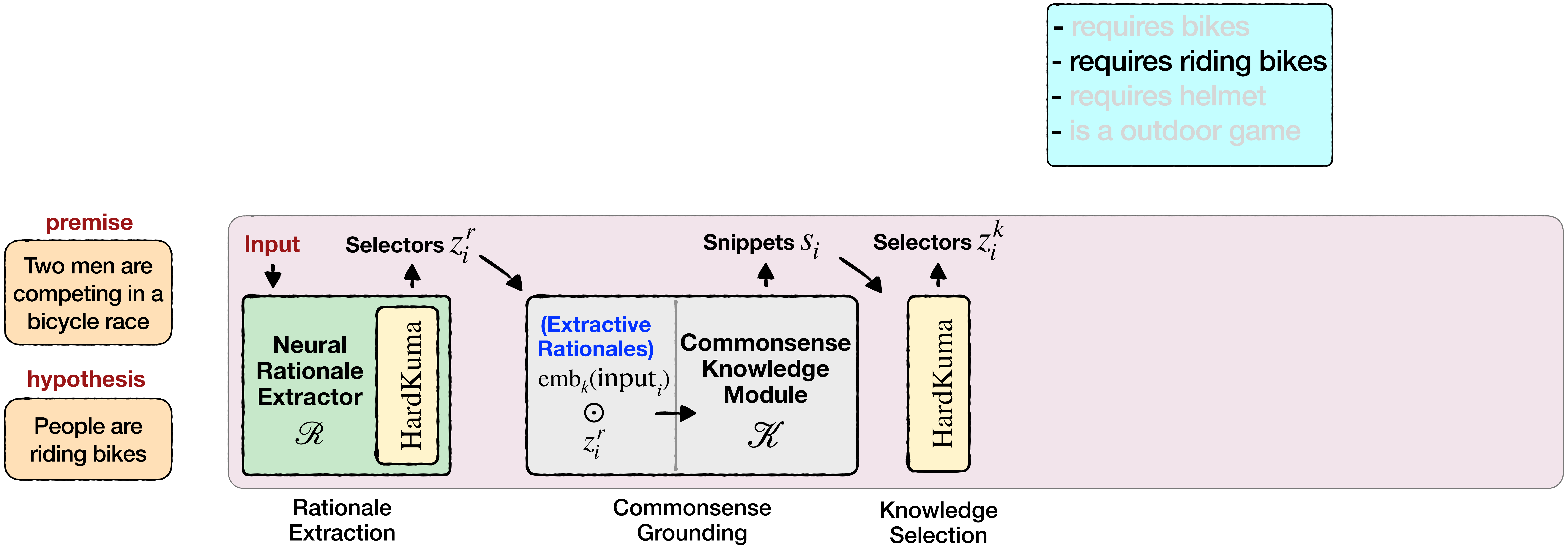
Each lexical unit from rationales are sent to the commonsense module $\mathscr{K}$, that result in knowledge snippets s_i



The series of binary latent variables z_i^r are used as masks on the embedded input

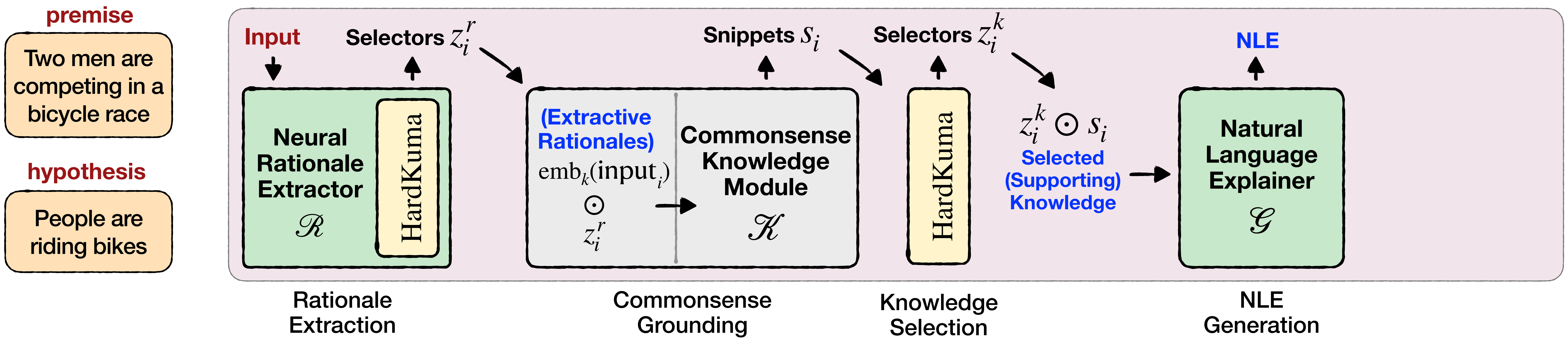
... and directly sent to a generative commonsense module $\mathscr{K}$, mirroring the modular approach

RExC



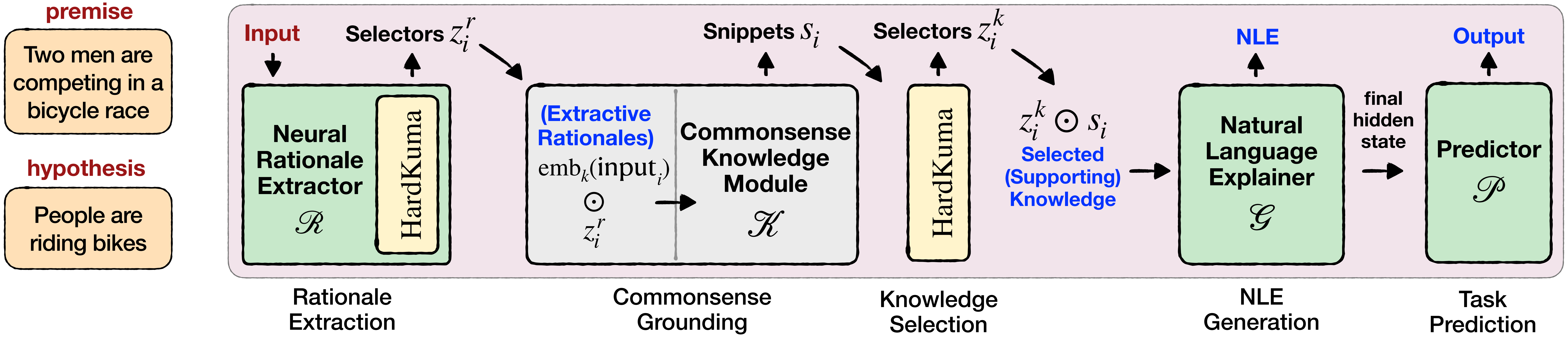
Another series of HardKuma variables are used to sample from all knowledge snippets generated. We operate on their soft forms $\tilde{s}_i$

RExC



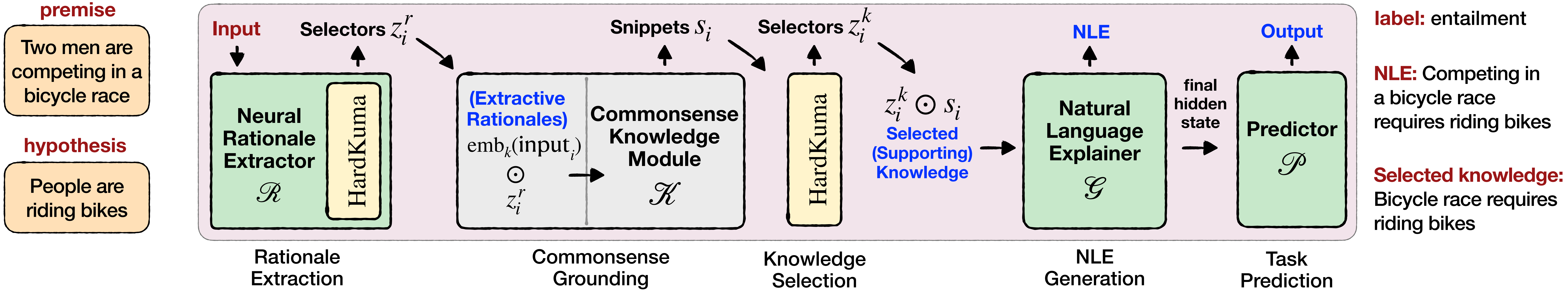
With the selected knowledge representations, generator $\mathscr{G}$ generates the NLE

RExC



The final hidden states of NLE are directly responsible for the output prediction

RExC



premise
Two men are competing in a bicycle race
hypothesis
People are riding bikes
Input
Selectors z_i^r
Neural Rationale Extractor
$\mathscr{R}$
HardKuma
(Extractive Rationales)
$\mathrm{emb}_k(\mathrm{input}_i) \odot z_i^r$
Commonsense Knowledge Module
$\mathscr{K}$
Snippets s_i
Selectors z_i^k
HardKuma
$z_i^k \odot s_i$
Selected (Supporting) Knowledge
NLE
Natural Language Explainer
$\mathscr{G}$
final hidden state
Output
Predictor
$\mathscr{P}$
Rationale Extraction
Commonsense Grounding
Knowledge Selection
NLE Generation
Task Prediction
label: entailment
NLE: Competing in a bicycle race requires riding bikes
Selected knowledge: Bicycle race requires riding bikes

Tasks

Natural Language Tasks

Natural Language Inference

premise: Two men are competing in a bicycle race

hypothesis: People are riding bikes

label
entailment

Commonsense Validation

A: Coffee stimulates people
B: Coffee depresses people

label
B is invalid

Commonsense QA

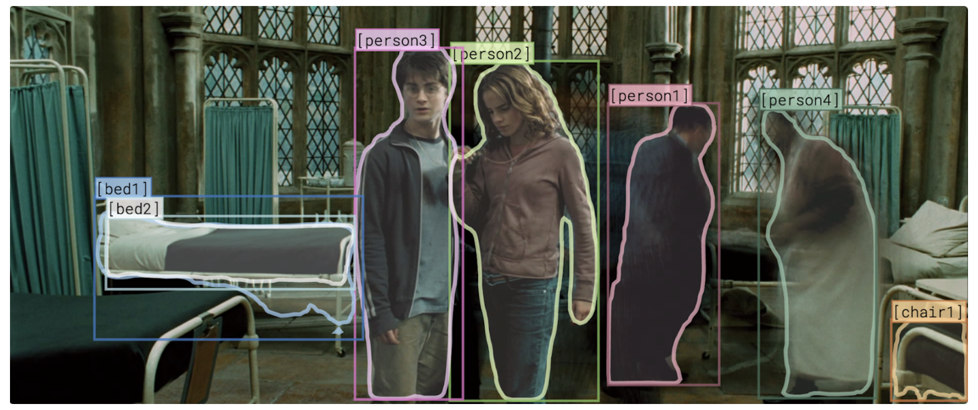
Q: Where does a wild bird usually live?

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

label
sky

Vision Language Tasks

Visual Entailment



Hypothesis: Some tennis players pose

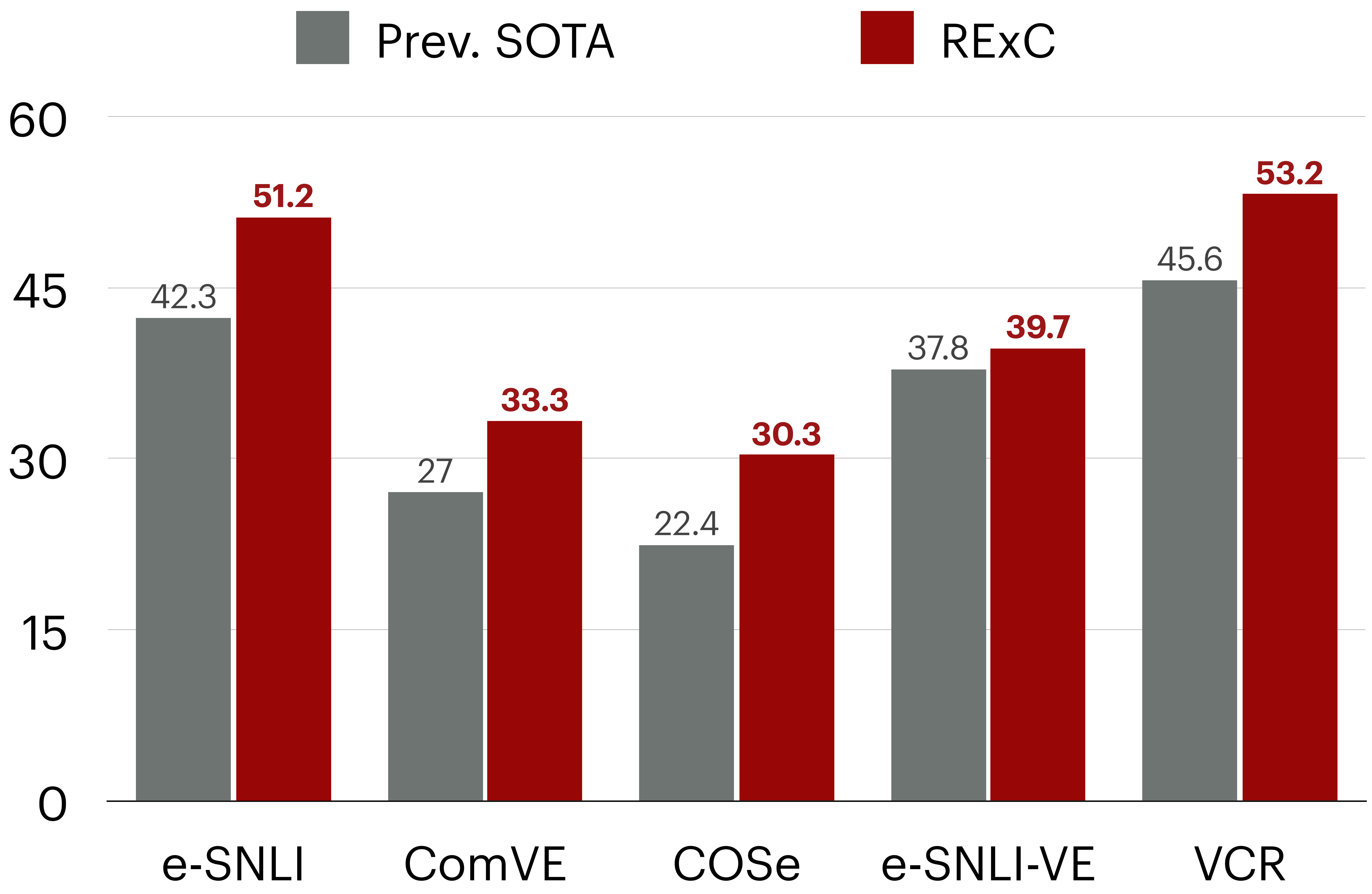
label
entailment

Visual Commonsense Reasoning

Q: What is the place?

label
They are in a hospital room

Automatic Evaluation for NLEs

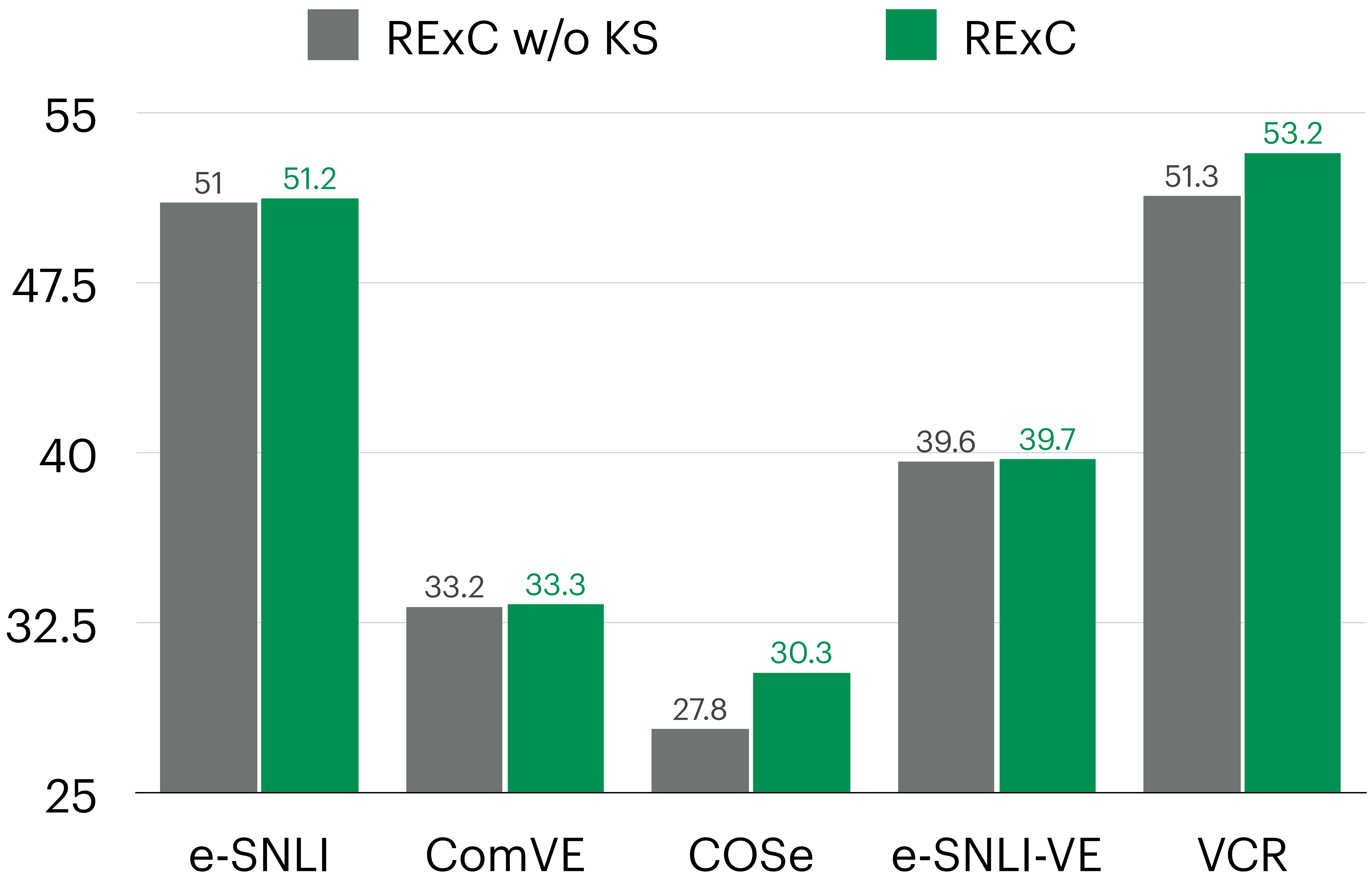


RExC is **better** than fine-tuned versions of pretrained language models (BART, WT5)

External commonsense is a useful component for more accurate NLEs

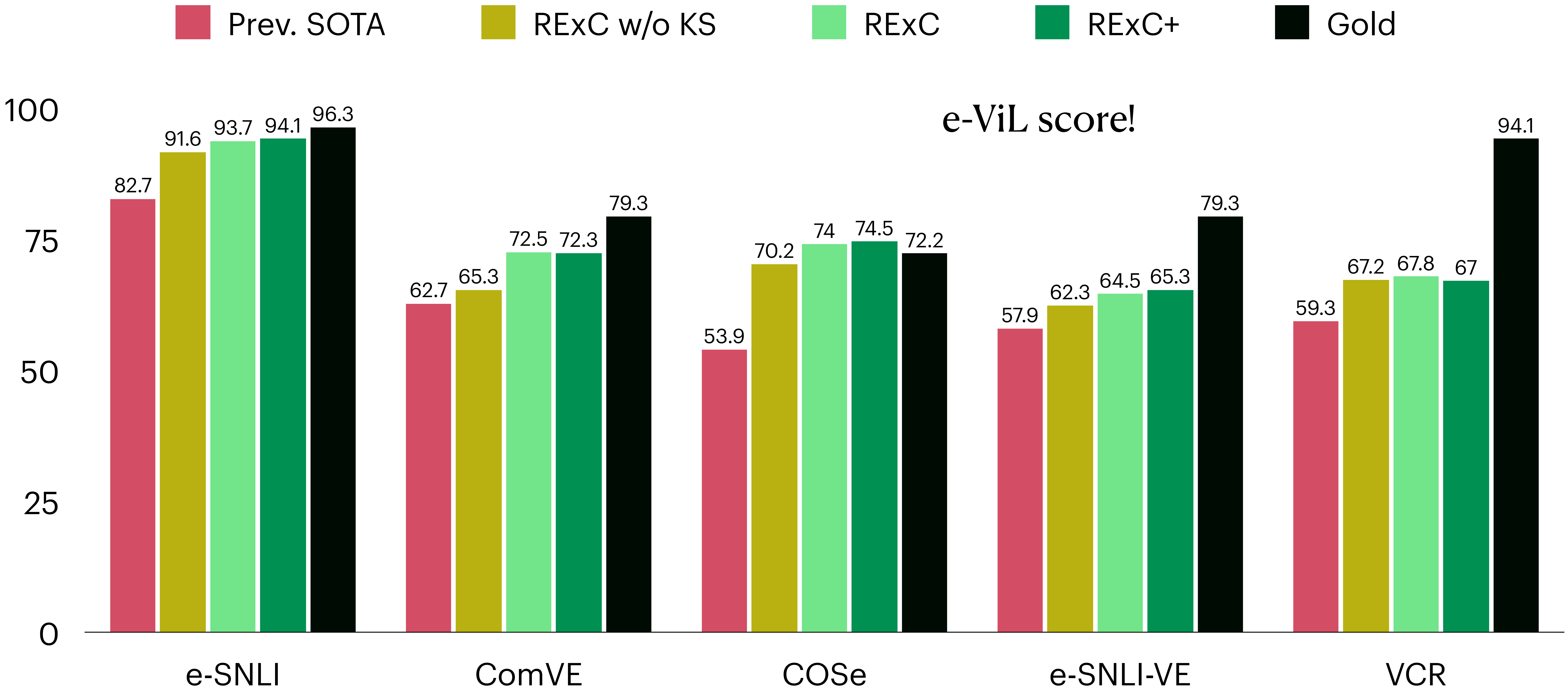
Rationales are useful to gather more relevant pieces of commonsense

Automatic Evaluation for NLEs



Knowledge Selection is useful compared to using all candidate snippets at once — it is more **interpretable** and **accurate**

Human Evaluation for NLEs



Qualitative Analysis

	Input	Rationales	Output	SOTA	REXC KS	Commonsense ($z_i^g > 0.8$)
ComVE	**A:** Coffee stimulates people **B:** Coffee depresses people	coffee	B	Coffee does not depress people	Coffee contains caffeine and is a popular stimulant	1. Coffee contains caffeine 2. Coffee is a stimulant
e-SNLI	**Premise:** A senior is waiting at the window of a restaurant that serves sandwiches. **Hypothesis:** A person waits to be served his food.	sandwiches, food	entail-ment	A person is waiting means a senior is waiting	A person is waiting for sandwiches means a person is waiting for food	1. Sandwich is a food
COSe	**Q:** Where does a wild bird usually live? **A:** a) cage, b) sky, c) countryside, d) desert, e) windowsill	wild, bird	sky	Bird flies in the sky	A wild bird flies in free sky	1. Wild bird is free 2. Bird flies in the sky

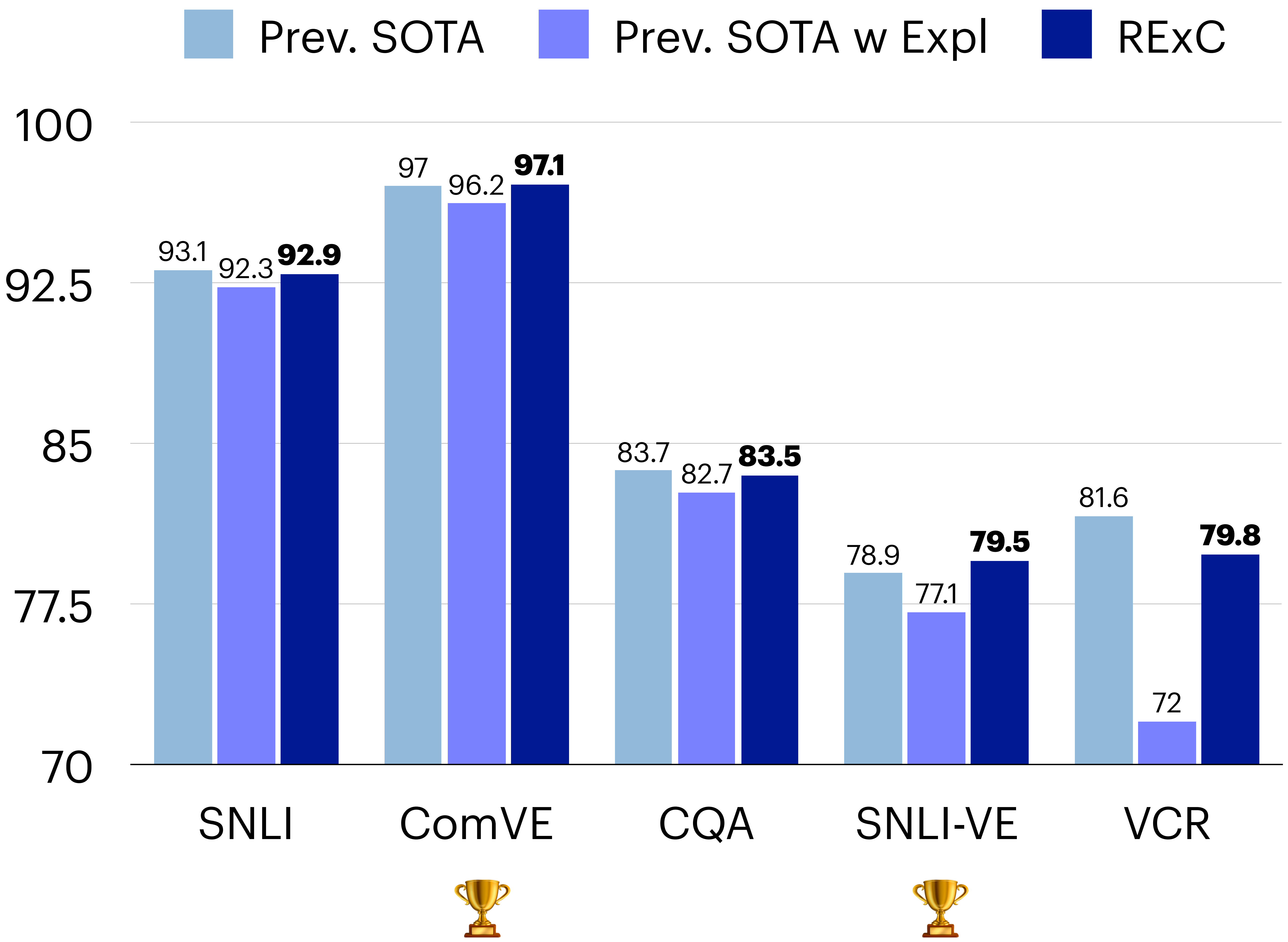
Sparse rationales

SOTA lacks commonsense

RExC is better-grounded with commonsense

RExC-KS+ can provide supporting evidence

Predictive Task Performance



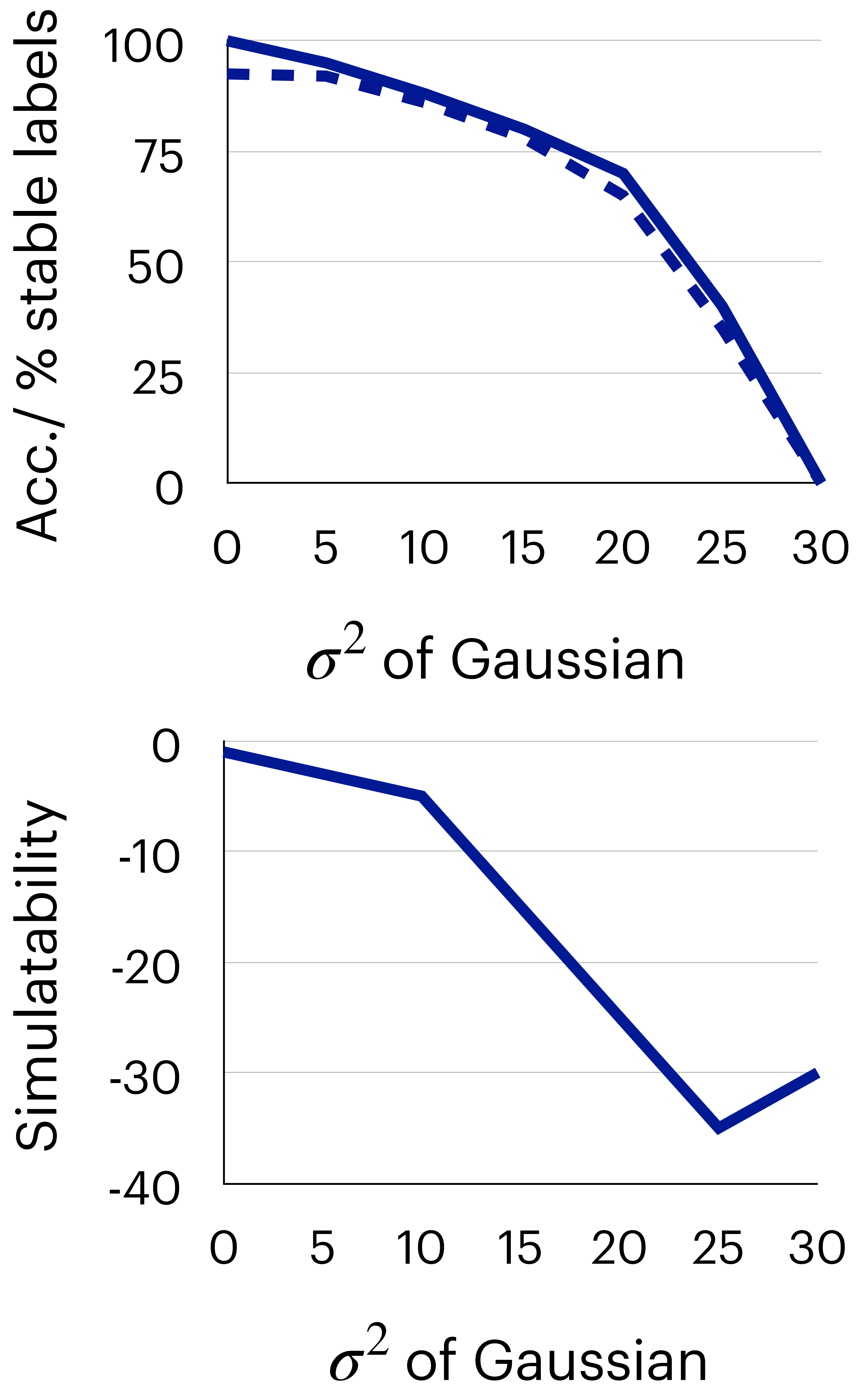
Both external **commonsense** and **NLEs** positively influence the task performance.

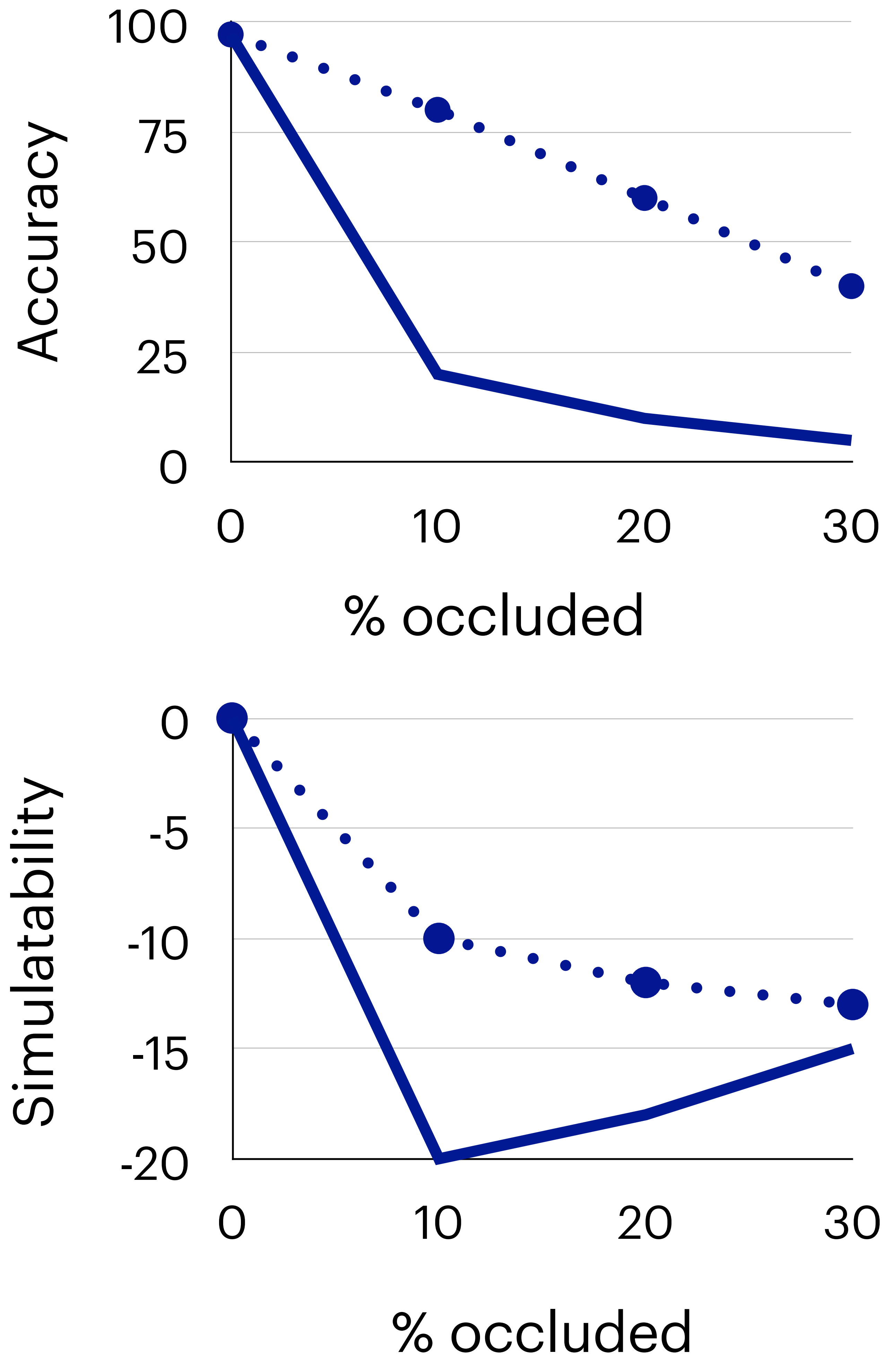
Beats all SOTA with explanation models

SOTA for ComVE and SNLI-VE

Association between Labels and NLEs

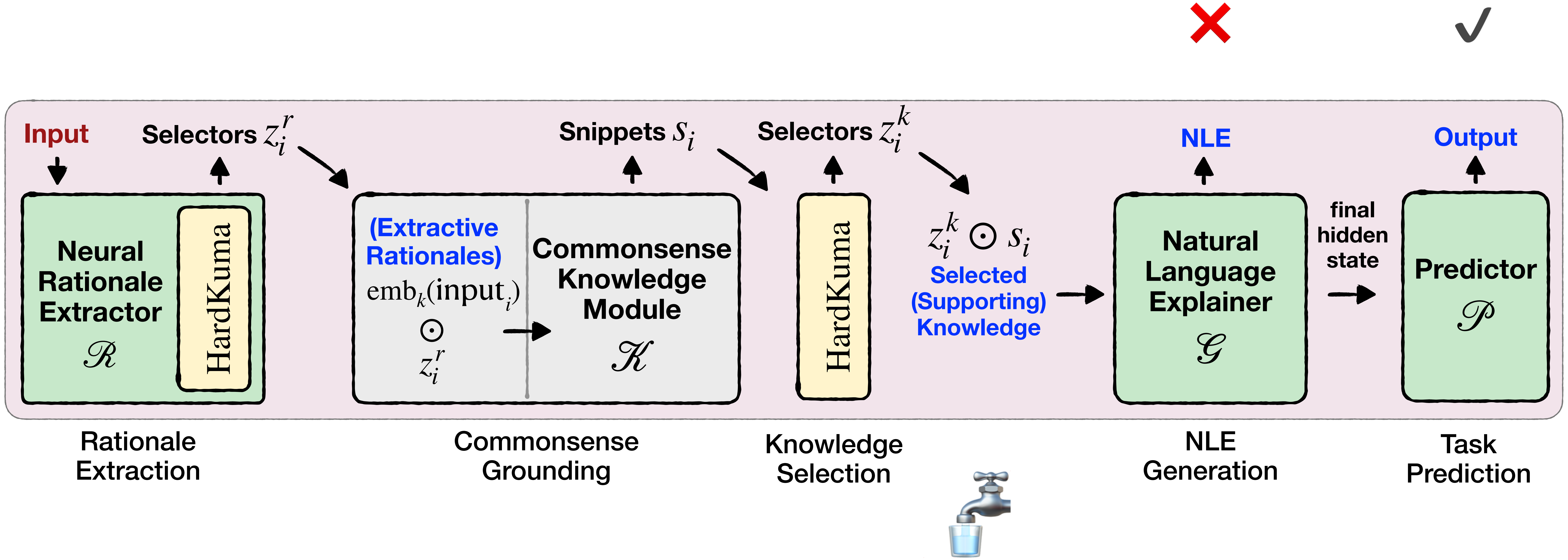
In presence of input **noise**, both labels and NLEs exhibit similar **robustness**





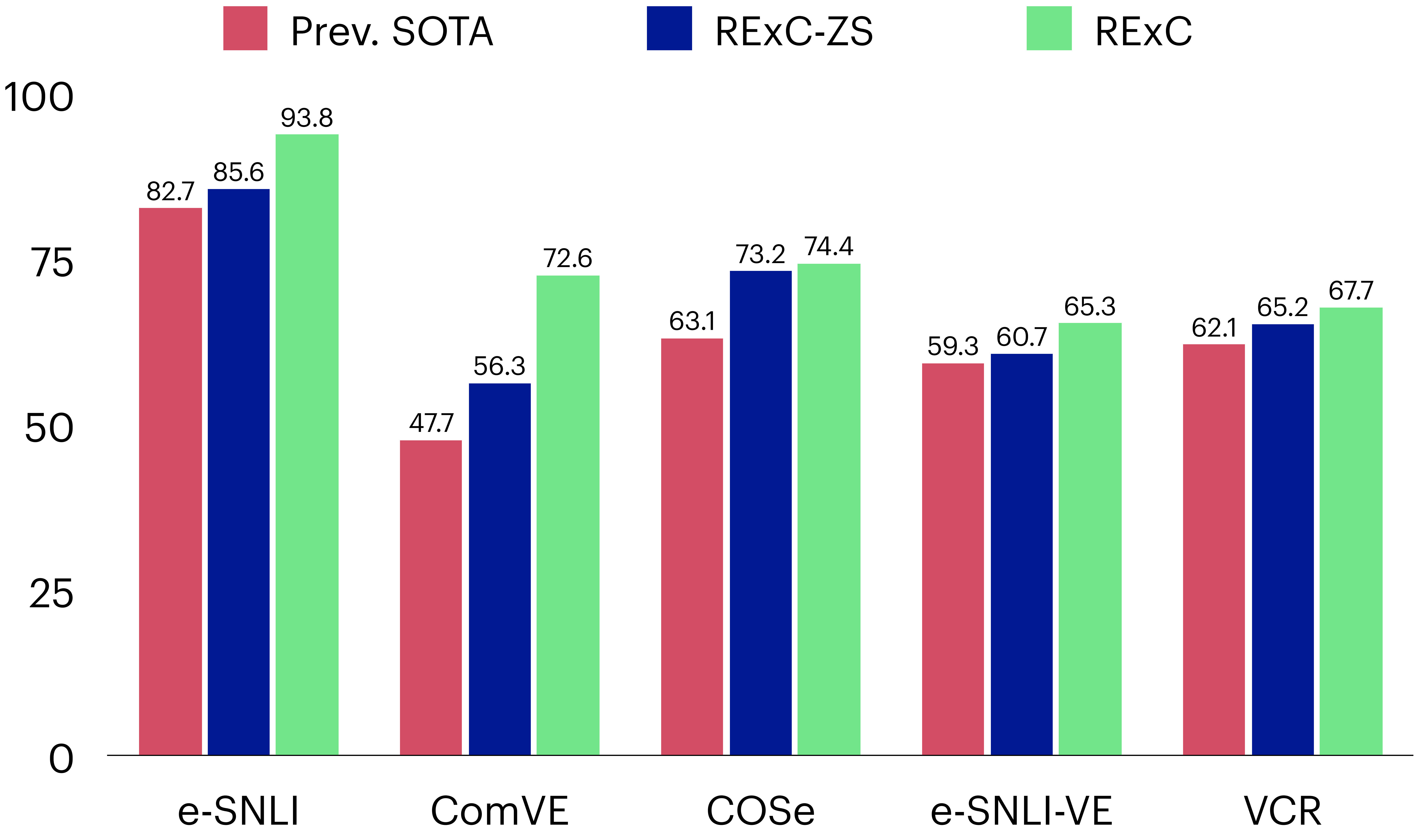
When we **occlude** salient tokens instead of random, the **drop in quality** for prediction and NLEs is significant

What's more in RExC?



Selected Knowledge as NLEs:
zero-shot NLEs ***only*** using the
supervision from predictive task

Zero-shot RExC



Zero-shot selection of knowledge snippets act as strong NLE in human evaluation, despite the lack in fluency

Summary

- A **unified framework** to combine **extractive** and **abstractive** explanations using external commonsense
- **Joint training** of extractive rationales and abstractive NLEs is powerful
- Generalization **across modalities** with **SOTA** on 5 commonsense knowledge tasks in both **NLP** and **vision**

What's next: Interactive Explainability

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

premise

Two men are competing in a bicycle race

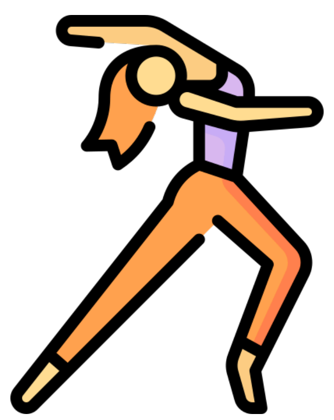
hypothesis

People are riding bikes

label: entailment

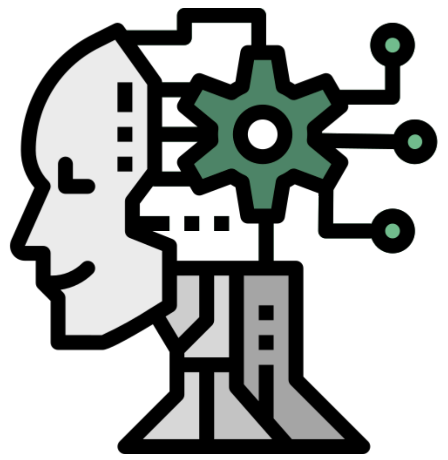
- requires bikes
- requires riding bikes
- requires helmet
- is a outdoor game

Competing in a bicycle race requires riding bikes



Two *men* can be considered as *people*

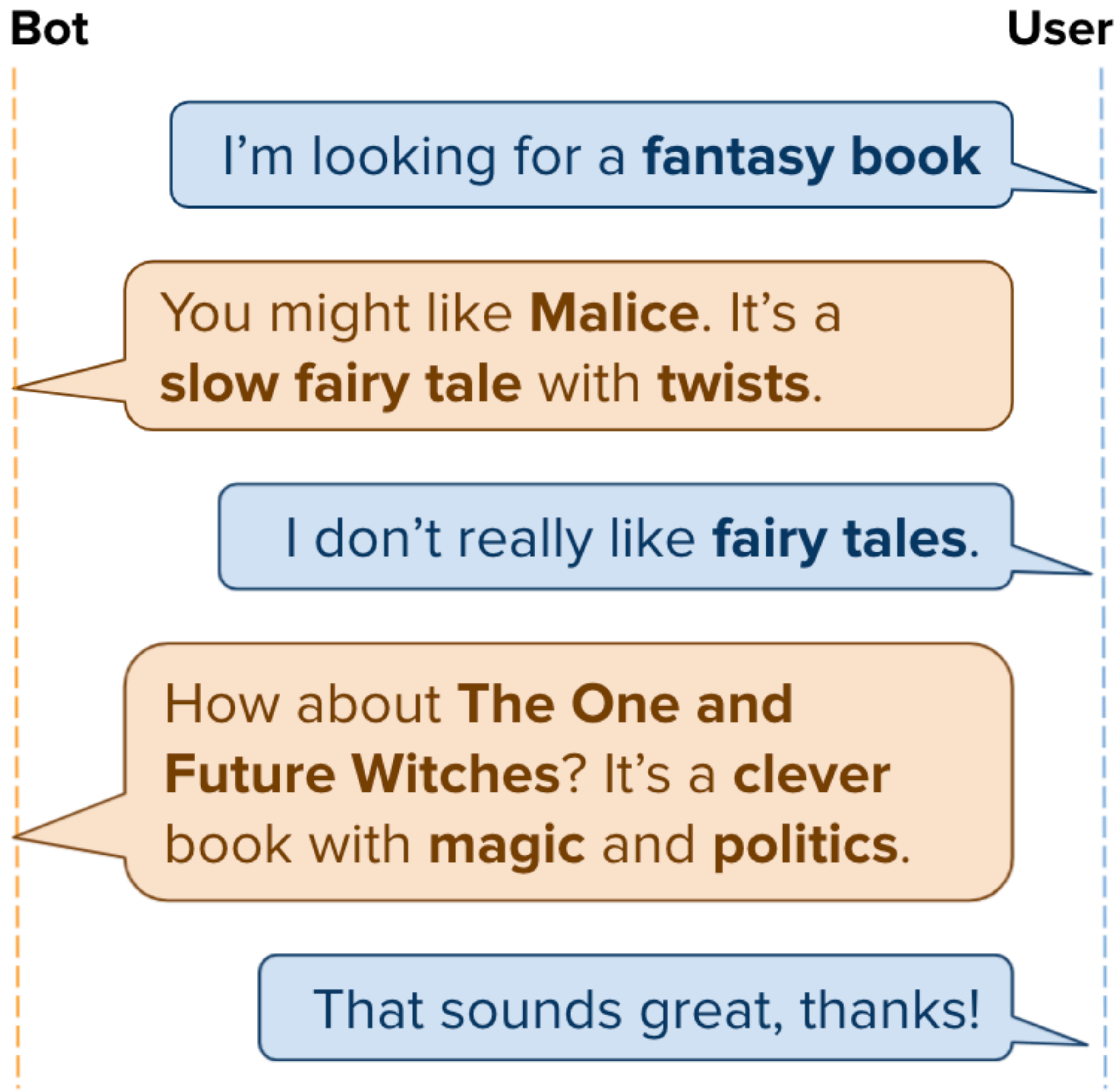
refines explanation…



Self-supervised Training for Conversational Recommenders with Justifications

Shuyang Li, Bodhisattwa Prasad Majumder, Julian McAuley
UC San Diego

Conversation with Justifications

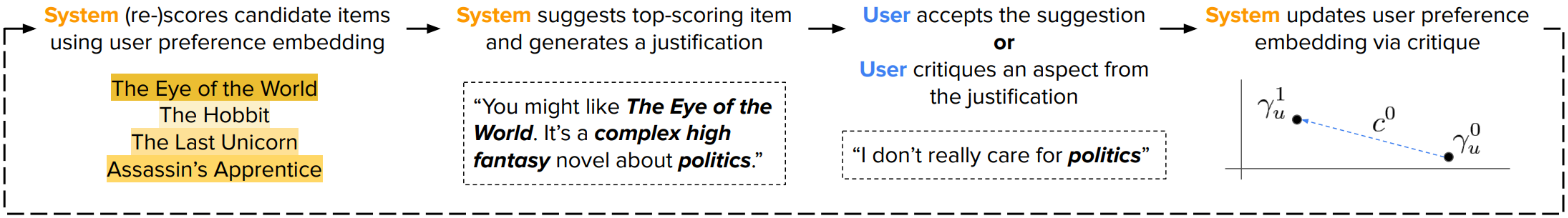


Justify suggestions made to the user

Update suggestions based on user feedback about **subjective aspects**

Be able to train the model **without collecting expensive dialog traces**

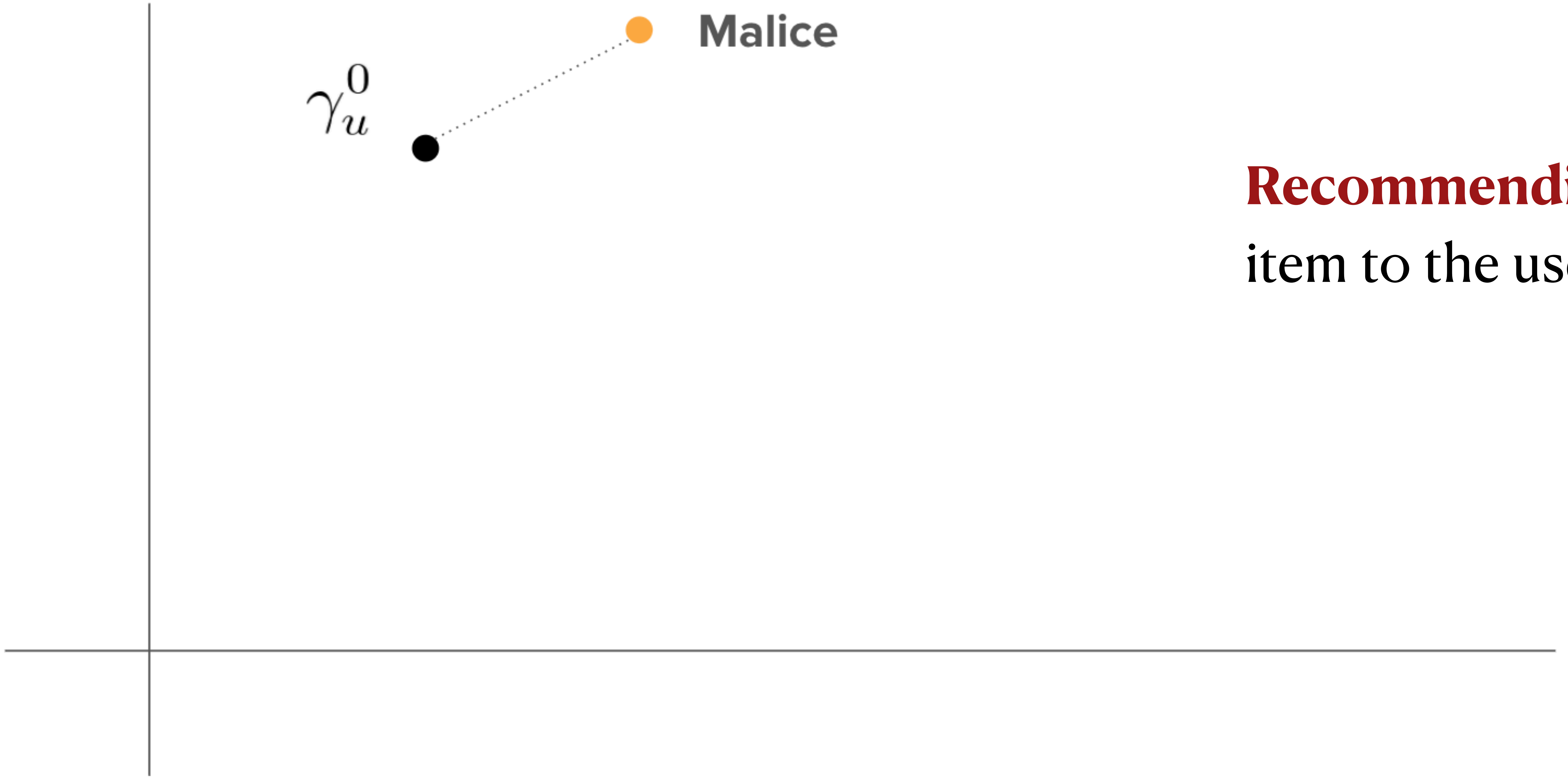
Conversation with Justifications



Jointly learn to **recommend** and **justify**, learning user representations that disentangle a user's latent preferences from their "observed" preferences (reviews)

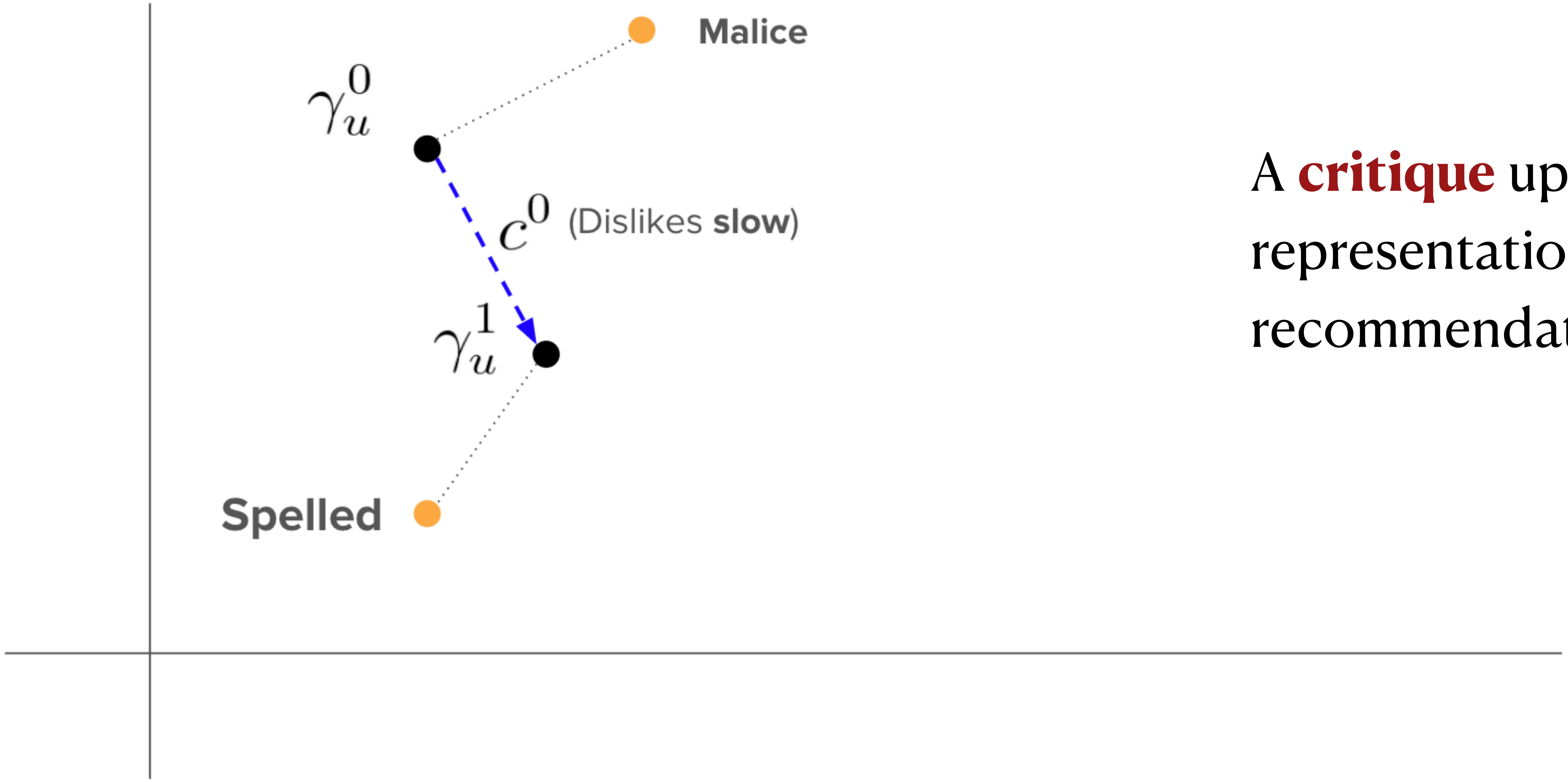
Fine-tune our model using a **bot-play framework** built on harvested reviews

Predict-Justify-Critique



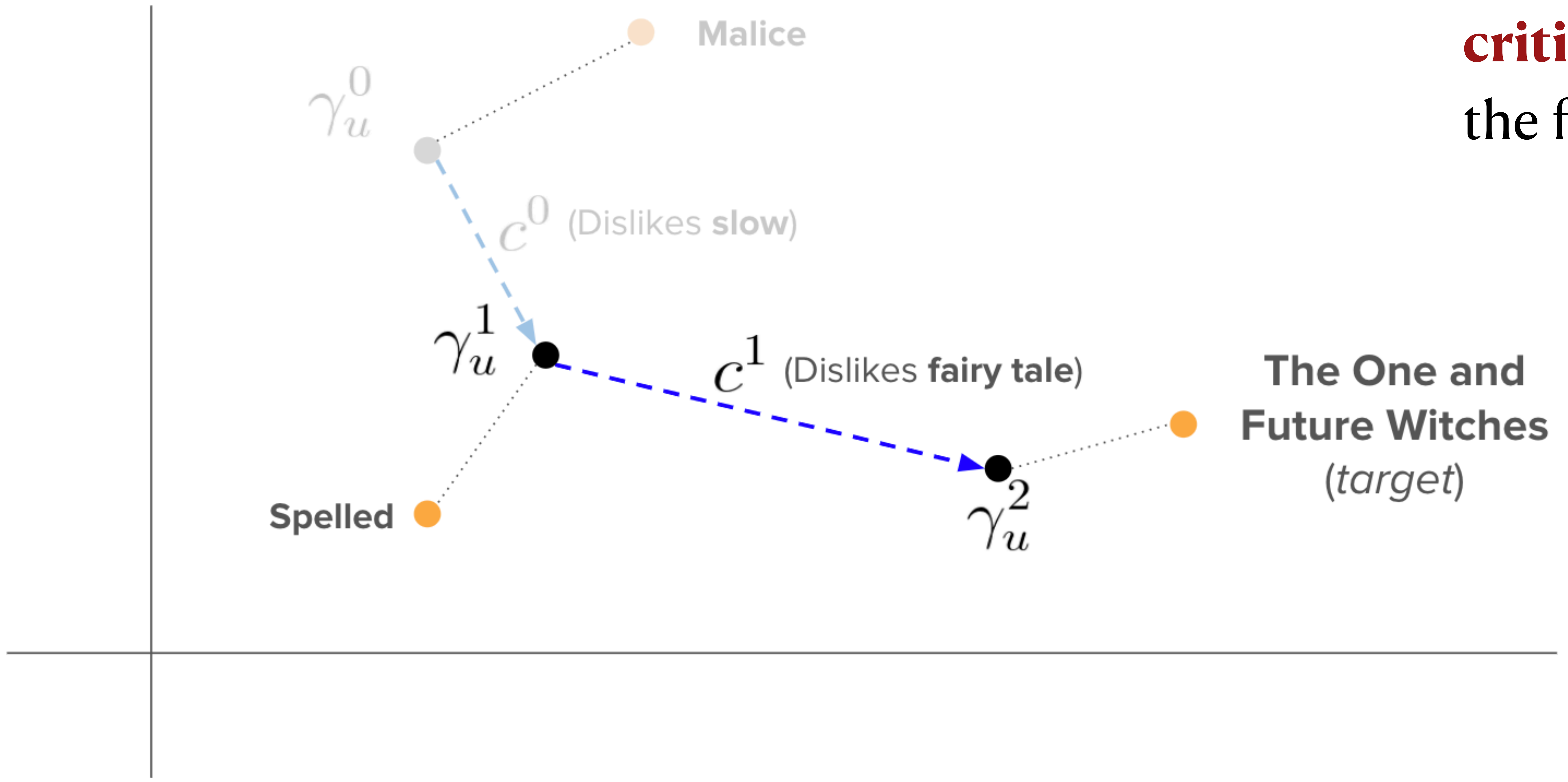
Recommending closest item to the user embedding

Predict-Justify-Critique



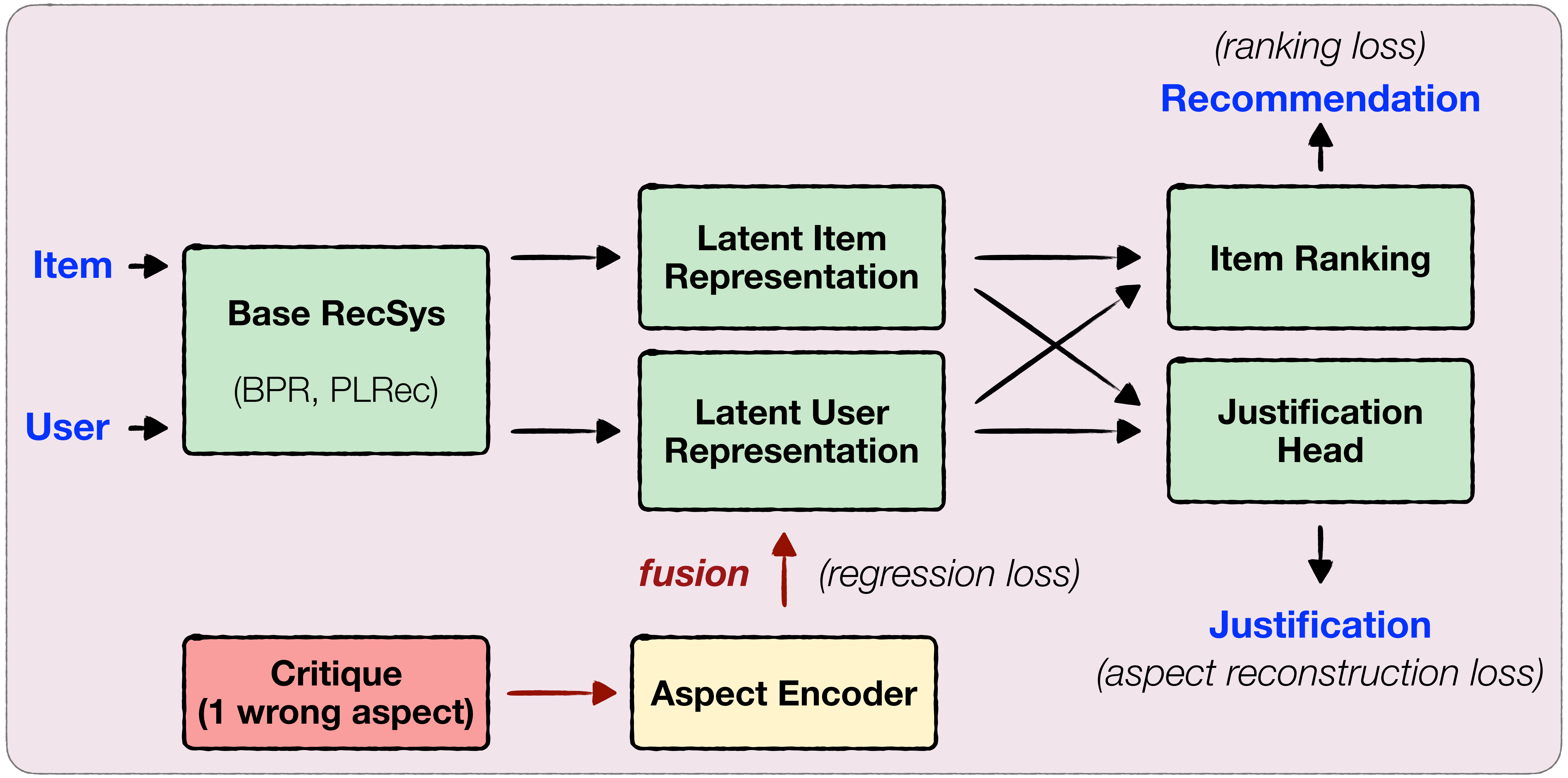
A **critique** updates user representation, hence the recommendation changes

Predict-Justify-Critique



It may require **multiple critiquing steps** to reach the final recommendation

ConvRec Model



From this point, one could update the internal representations using (self) supervised **bot-play**

or

incorporate the critique with **inference-time** update.

Learning to Critique via Bot-play

At turn ***t*,**

Predict scores for item recommendation

Calculate loss for w.r.to the gold item (from evaluation set)

Sample item ***i***, to recommend

if ***i*** is the gold item, STOP

else

Generate justification with aspect scores

Seeker critiques the justification

Seeker critiques the most popular aspect from the justification, except those are in target item's history

User latent representation is updated with new critique

Learning to Critique via Bot-play

At turn ***t*,**

Predict scores for item recommendation

Calculate loss for w.r.to the gold item (from evaluation set)

Sample item ***i***, to recommend

if ***i*** is the gold item, STOP

else

Generate justification with aspect scores

Seeker critiques the justification

Seeker critiques the most popular aspect from the justification, except those are in target item's history

User latent representation is updated with new critique

Informed heuristic to simulate dialog

Updates partial model parameters, hence memorizes

(Alternative) Using Critique *only* during inference

At turn ***t***,

Predict scores for item recommendation

Calculate loss for w.r.to the gold item (from evaluation set)

Sample item ***i***, to recommend

if ***i*** is the gold item, STOP

else

Generate justification with aspect scores

Seeker critiques the justification

Seeker critiques the most popular aspect from the justification, except those are in target item's history

Update **only item ranking** and **justification** to match new user preference

Gradient-based updates works at inference, but doesn't help memorizing

(Alternative) Using Critique *only* during inference

Back to the Future: Unsupervised Backprop-based Decoding for Counterfactual and Abductive Commonsense Reasoning

Lianhui Qin[†‡] **Vered Shwartz**[†‡] **Peter West**[†‡] **Chandra Bhagavatula**[‡]
Jena D. Hwang[‡] **Ronan Le Bras**[‡] **Antoine Bosselut**[♮‡] **Yejin Choi**[†‡]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial
{lianhuiq, pawest, y
{vered, chandrab, je

Unsupervised Enrichment of Persona-grounded Dialog with Background Stories

Bodhisattwa Prasad Majumder[♣] **Taylor Berg-Kirkpatrick**[♣]
Julian McAuley[♣] **Harsh Jhamtani**[◇]
[♣]Department of Computer Science and Engineering, UC San Diego
{bmajumde, tberg, jmcauley}@eng.ucsd.edu
[◇]School of Computer Science, Carnegie Mellon University
jharsh@cs.cmu.edu

Evaluation

User Simulation

Simulating 500 users with **warm-start** preferences

Critiques are for **random**, **popular**, and **most divergent** aspects

Measures **success rate** and **length**

User Study

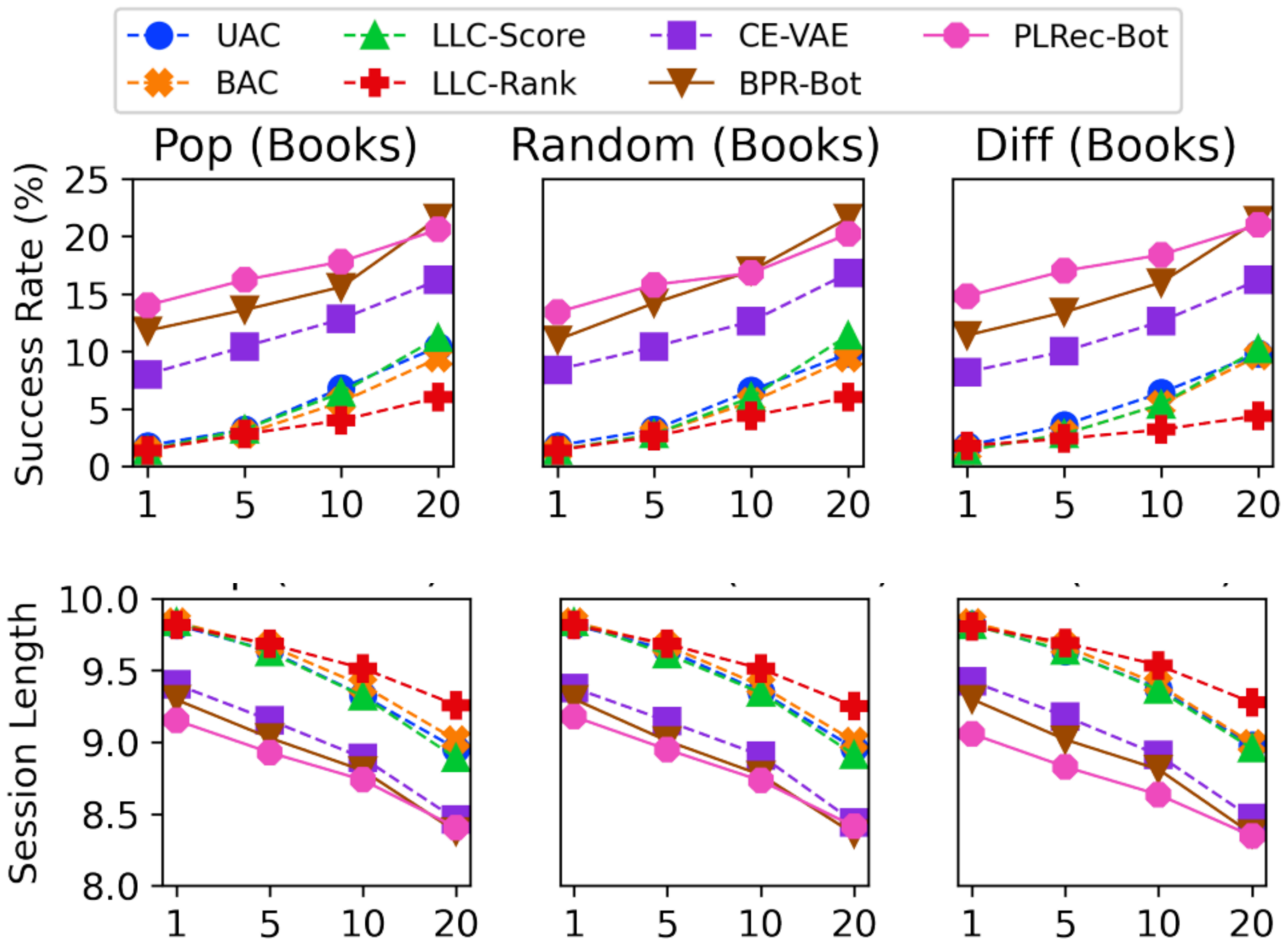
32 human users in **cold-start** setting

Turn-level annotation for response quality (with recommendation and justifications)

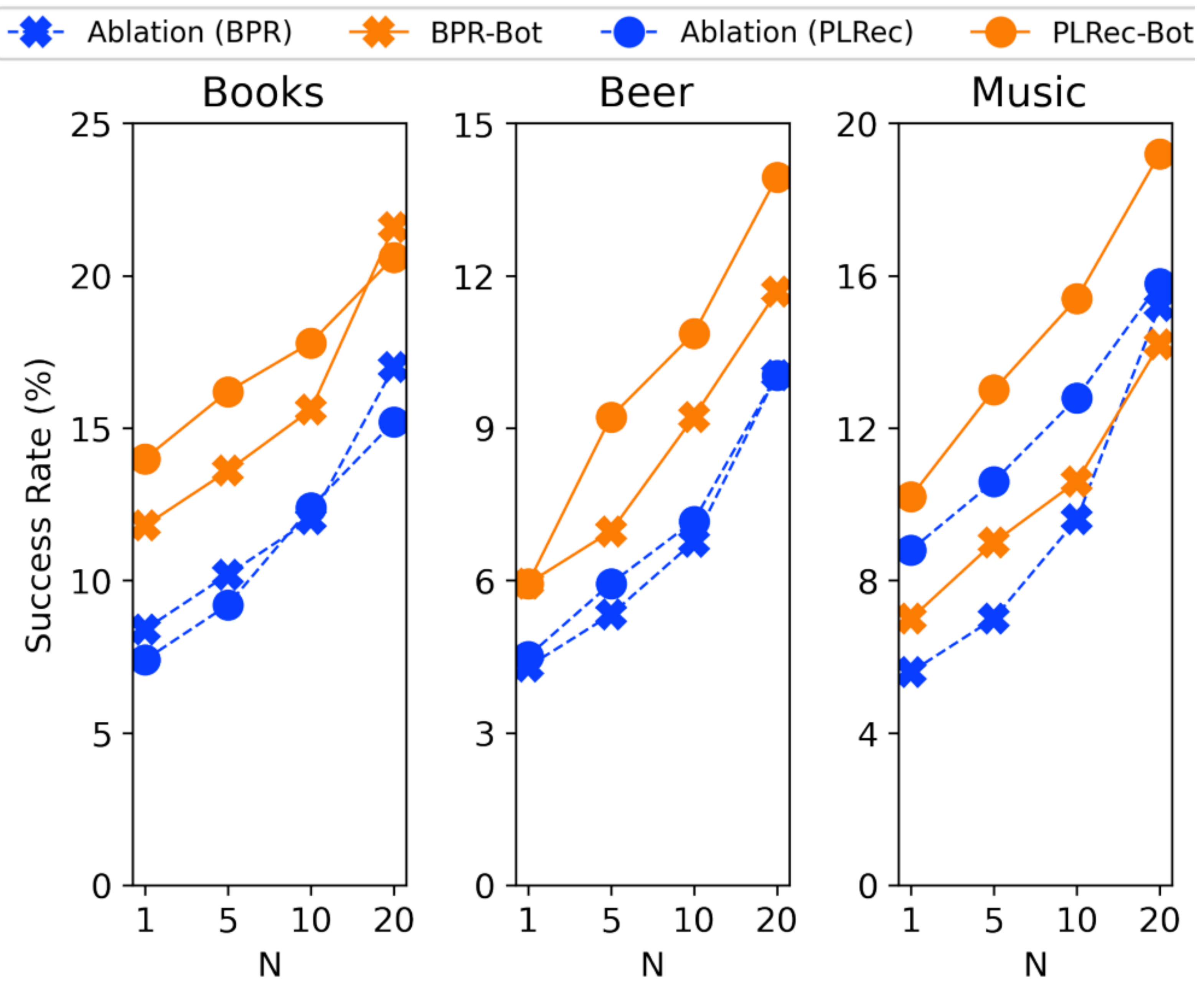
Overall **preference** for the system

Results

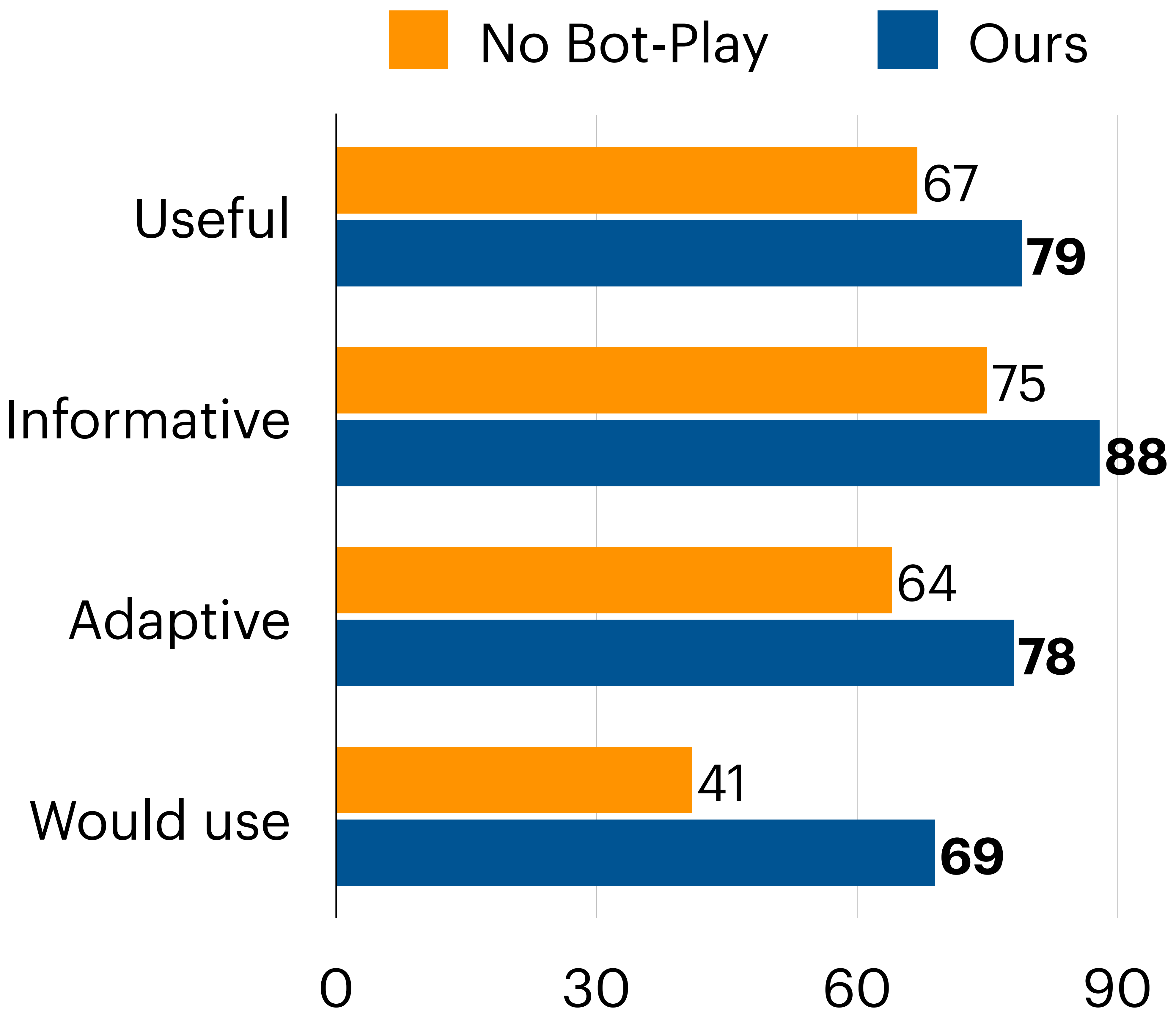
Higher success rates with **shorter session** lengths, critiquing helps



Bot-play fine-tuning improves target item ranking



Results and Summary



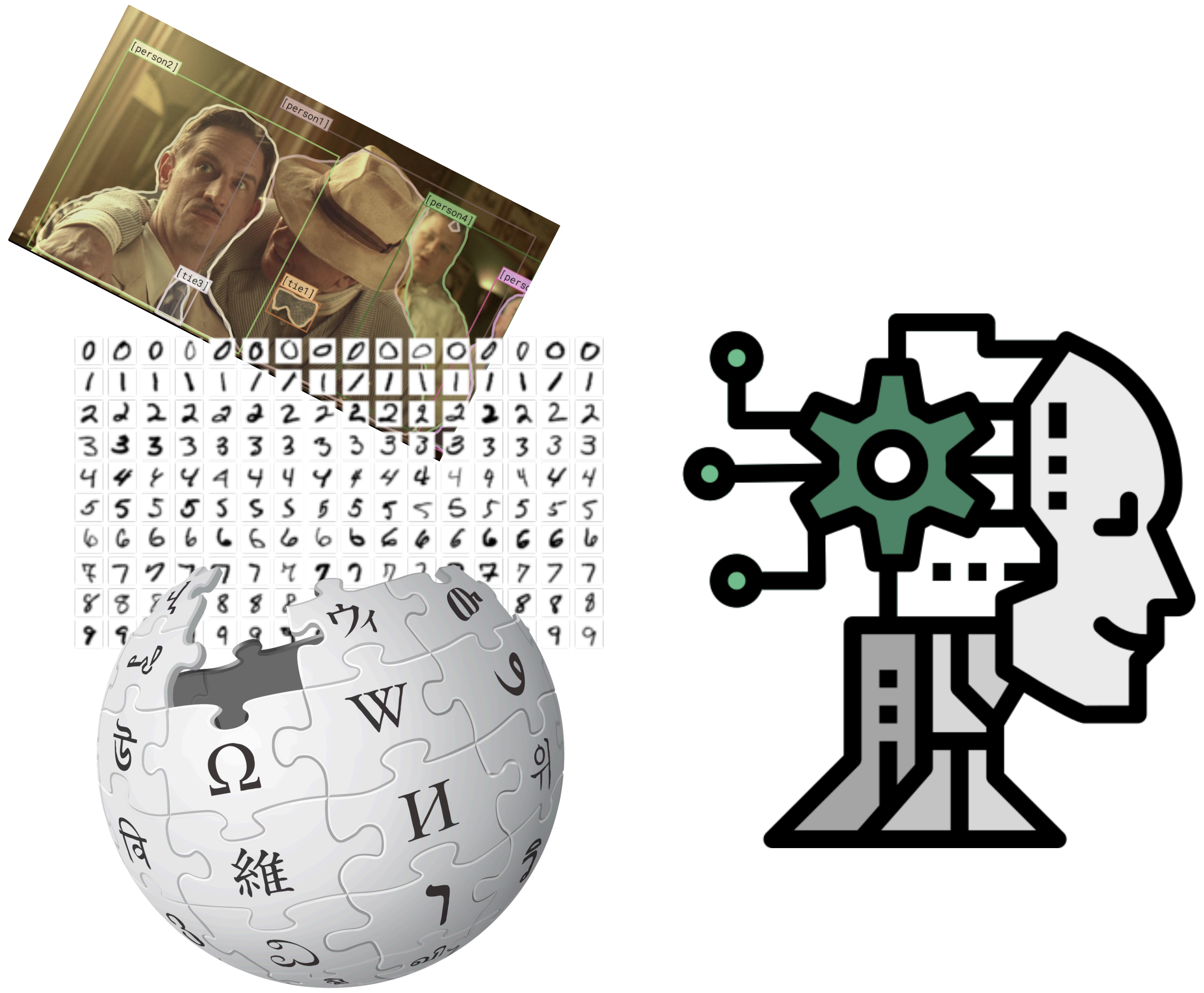
In summary

We show that a bot-play framework can be used without actually collecting dialog traces

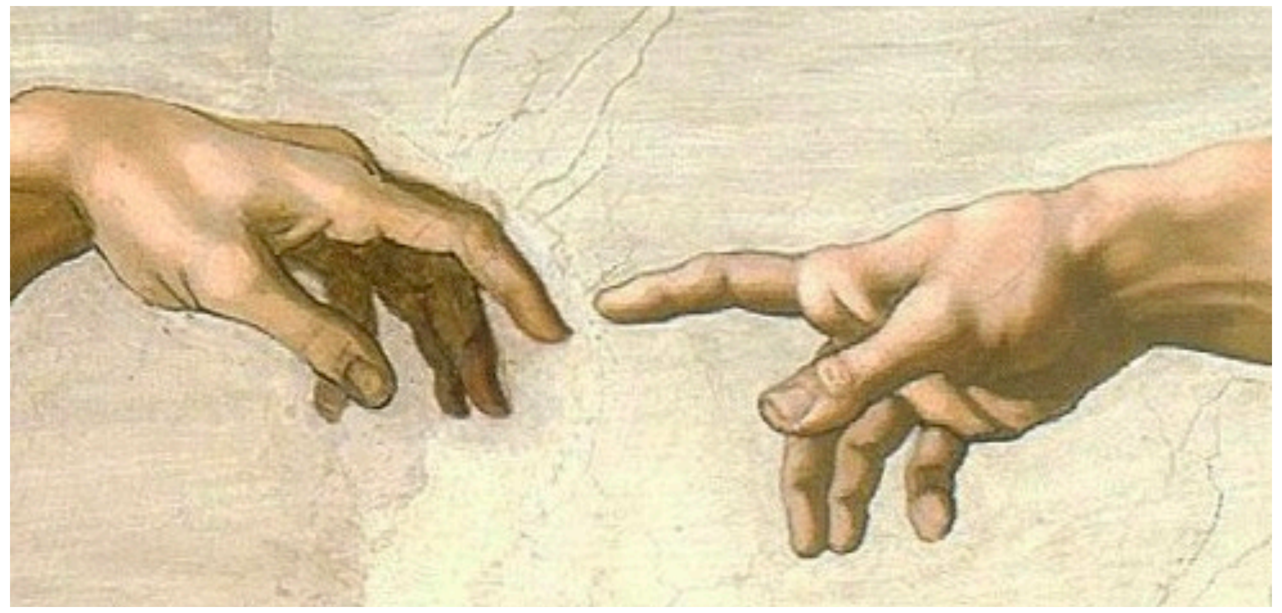
Bot-play improves multi-turn critiquing

Can extend to natural language justifications and feedback for more natural conversation

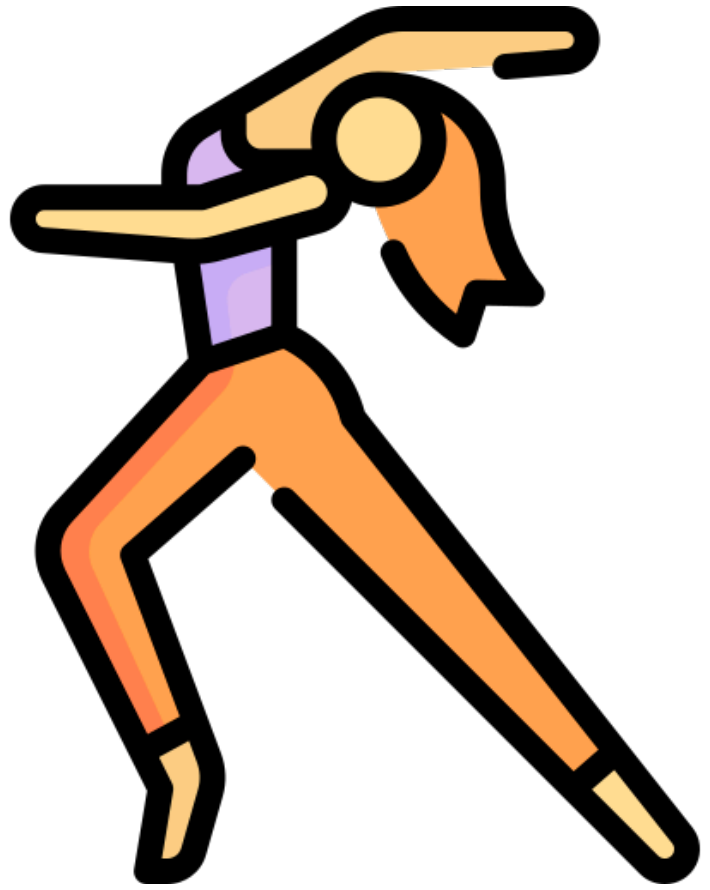
Explanations with Commonsense and Interactions

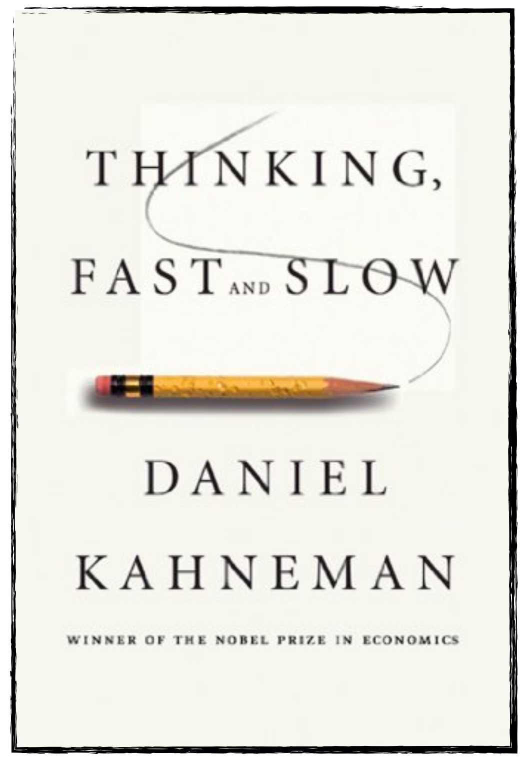


Conceptual Mappings with World Knowledge



Natural Language Feedback

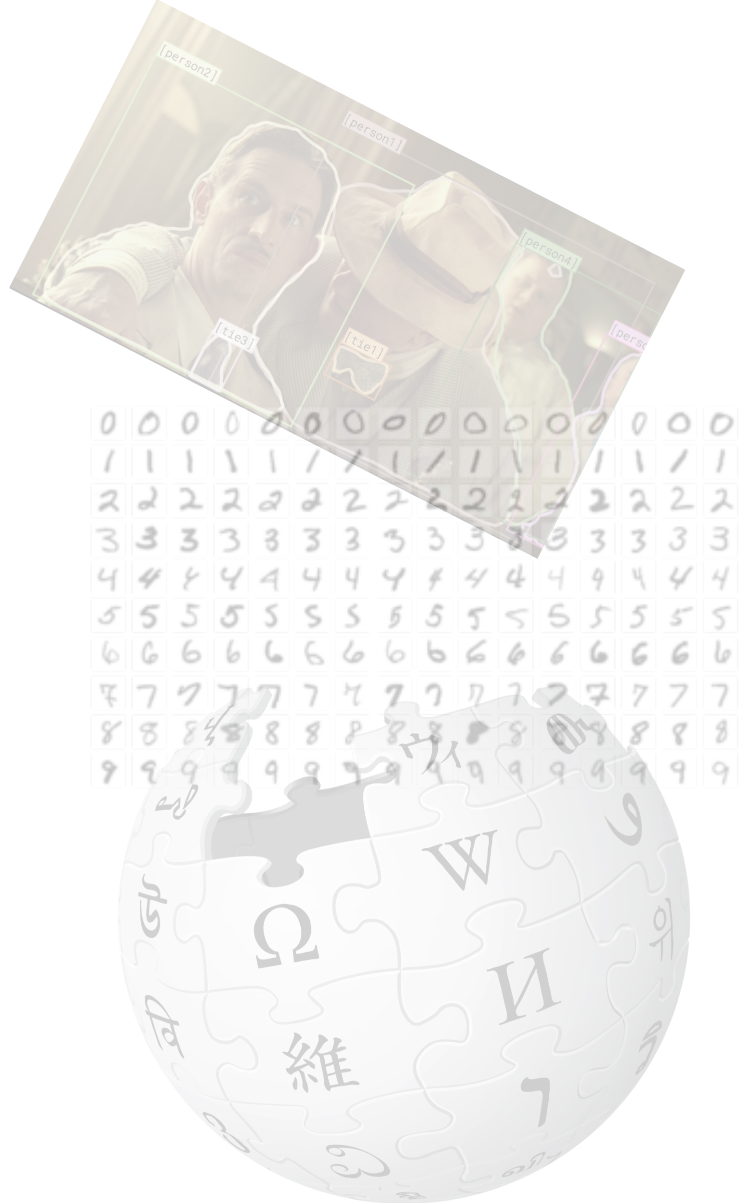




{perception, intuition, reasoning}

{perception, intuition, reasoning}

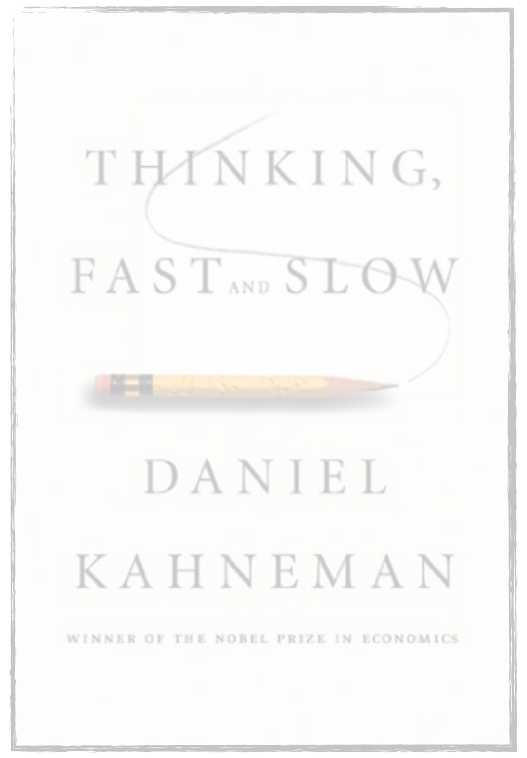
Explanations with Commonsense and Interactions



Formalizing the framework for conversational explanations

Exploring ways to ‘**memorize**’ and ‘**inference-time updates**’ based on user feedback

Collecting **synthetic** and **real datasets** to support conversations around explanations



{perception, i

on, reasoning}

Acknowledgement

Sponsors

Qualcomm innovation fellowship

amazon alexa prize

NSF

Friends of the International Center
Friendship • Scholarships • Hospitality

Advisor

Julian McAuley
UC San Diego

Collaborators

Google AI

Microsoft Research

The Alan Turing Institute

Carnegie Mellon University

FACEBOOK AI

University of Oxford

Published Works

Unsupervised Enrichment of Persona-grounded Dialog with Background Stories | **ACL** (oral), 2021
Bodhisattwa P. Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, Harsh Jhamtani
An unsupervised gradient-based rewriting framework to adapt background stories to an existing persona-grounded dialog

Ask what's missing and what's useful: Improving Clarification Question Generation using Global Knowledge | **NAACL** (oral), 2021
Bodhisattwa P. Majumder, Sudha Rao, Michell Galley, Julian McAuley
A two-stage framework to 1) estimate missing information from the global knowledge and 2) generate useful questions with them.

Like hiking? You probably enjoy nature: Persona-grounded Dialog with Commonsense Expansions | **EMNLP** (oral), 2021
Bodhisattwa P. Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, Julian McAuley
A variational learning framework to capture commonsense implications of input persona in a persona-grounded dialog

Interview: Large-scale Modeling of Media Dialog with Discourse Patterns and Knowledge Grounding | **EMNLP** (oral), 2021
Bodhisattwa P. Majumder*, Shuyang Li*, Jianmo Ni, Julian McAuley
A large-scale analysis of discourse in media dialog and its impact on generative modeling of dialog with knowledge grounding

Generating Personalized Recipes from Historical User Preferences | **EMNLP**, 2019
Bodhisattwa P. Majumder*, Shuyang Li*, Jianmo Ni, Julian McAuley
A new task of personalized recipe generation to generate natural-text instructions aligned with the user's historical preferences

Improving Neural Story Generation by Targeted Common Sense Grounding | **EMNLP**, 2021
Henry Mao, **Bodhisattwa P. Majumder**, Julian McAuley, Gary Cottrell
A multi-task learning scheme to achieve quantitatively better common sense reasoning in language models

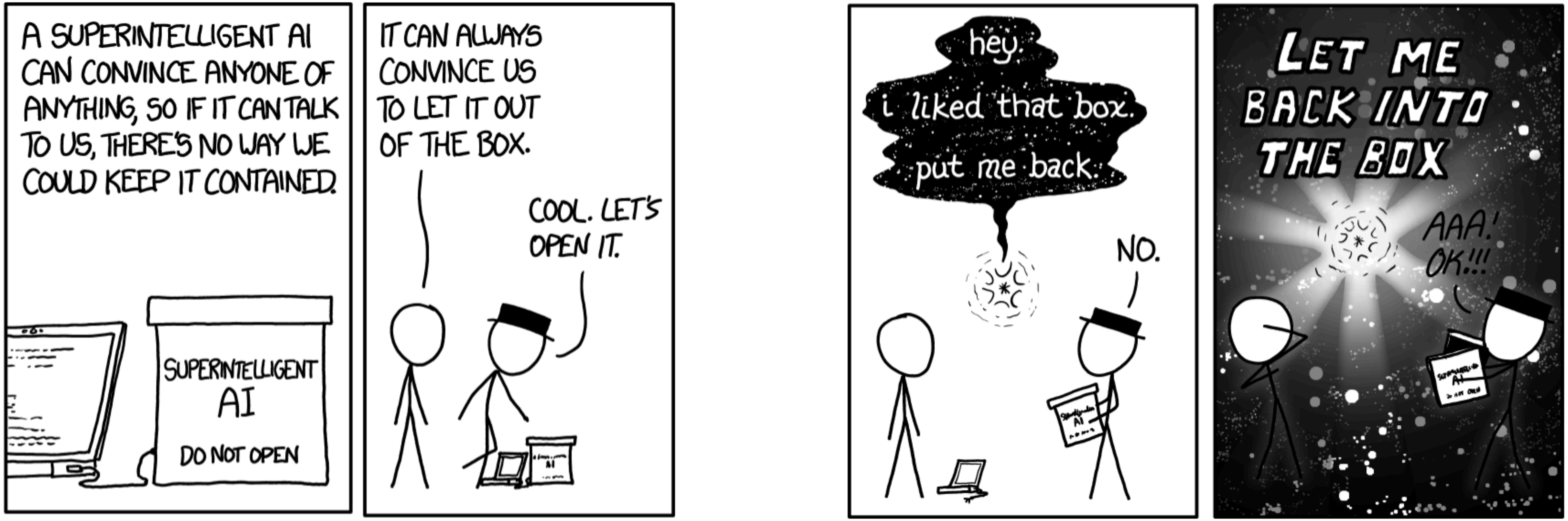
Published Works and Preprints

ReZero is All You Need: Fast Convergence at Large Depth | **UAI** (oral), 2021
Thomas Bachlechner*, **Bodhisattwa P. Majumder***, Henry Mao*, Gary Cottrell, Julian McAuley
A novel deep neural network architecture that initializes an arbitrary layer as the identity map to facilitate signal propagation at depth

Representation Learning for Information Extraction from Form-like Documents | **ACL** (oral), 2020
Bodhisattwa P. Majumder, Navneet Potti, Sandeep Tata, James Wendt, Qi Zhao, Marc Najork
A novel approach to learn interpretable representations for target fields using spatial and contextual knowledge for form-like documents

Detect and Perturb: Neutral Rewriting of Biased and Sensitive Text via Gradient-based Decoding | Findings of **EMNLP**, 2021
Zexue He, **Bodhisattwa P. Majumder**, Julian McAuley
A rewriting framework to detect sensitive components from input text and neutralize at decoding time without any parallel corpus

Rationale-Inspired Natural Language Explanations with Commonsense | **arXiv**, 2021
Bodhisattwa P. Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, Julian McAuley
An end-to-end framework to connect extractive rationales with natural language explanations using commonsense



majumderb.com

@mbodhisattwa

Thank you!

Questions?