Methods for Improving the Communication Efficacy of Language Models: Faithfulness and Pragmatics

Lingjun Zhao

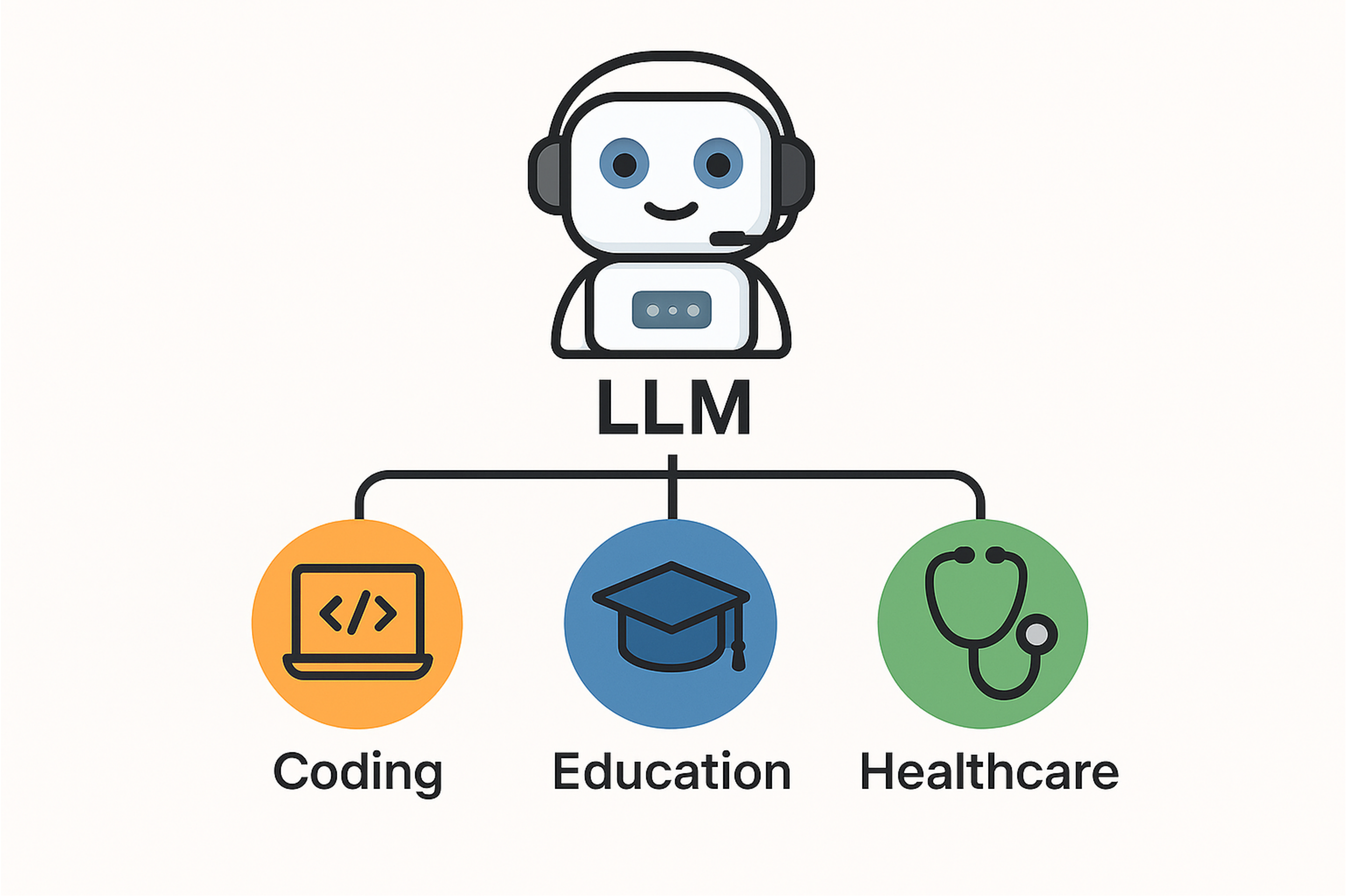
Preliminary Exam

Examining Committee:

Dr. Hal Daumé III (Chair)
Dr. David Jacobs (Dept. Rep)
Dr. Kunpeng Zhang (Dean's Rep)
Dr. Marine Carpuat
Dr. Jordan Boyd-Graber
Dr. Jia-Bin Huang

Motivation: AI as **assistants**

- Large language models (LLM) becoming potentially valuable assistants

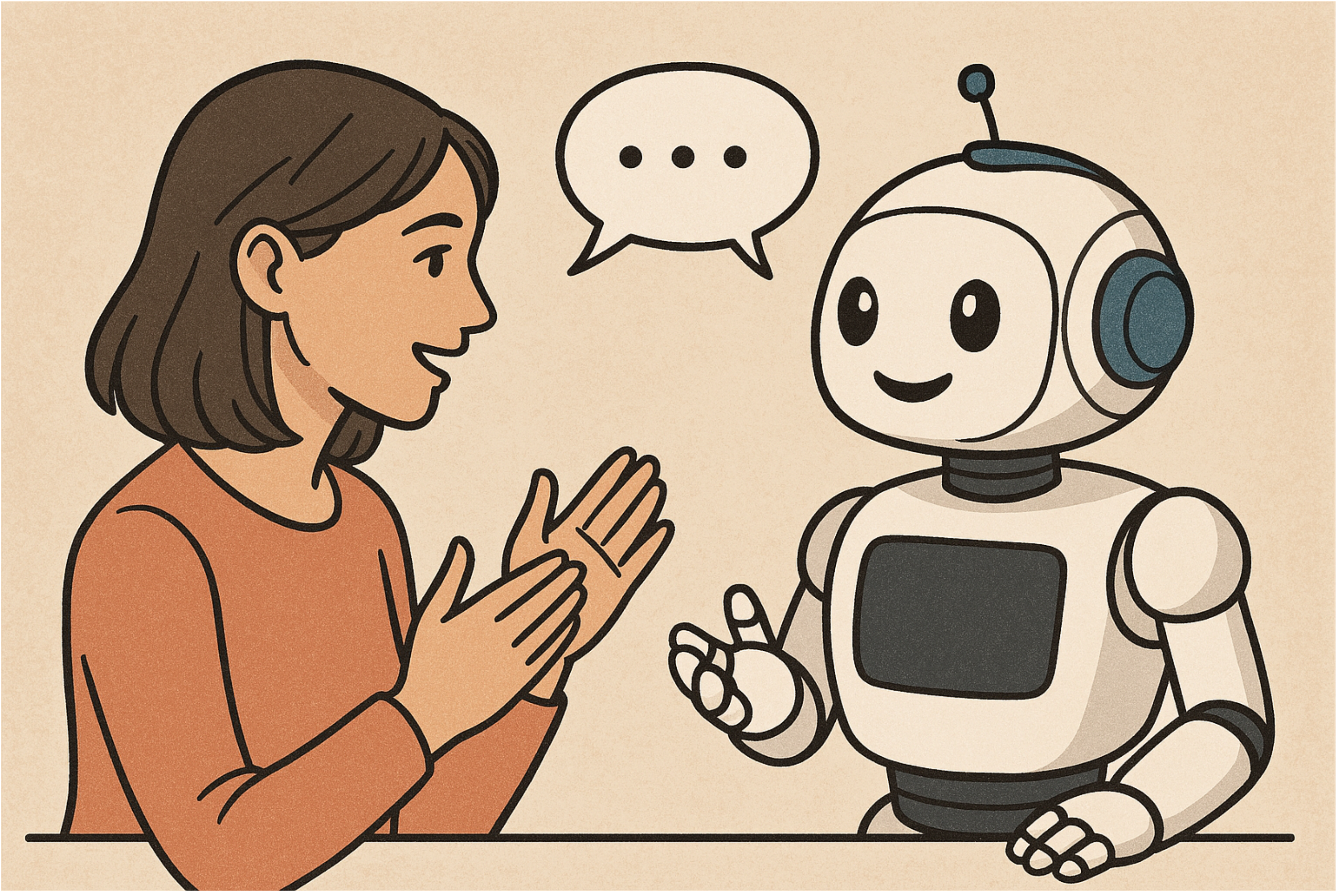


What is **effective** communication and why **important**?

- Clarify AI's limitations

  - Facilitate human-AI collaboration

  - E.g. coding assistant flags uncertainty



What is **effective** communication and why **important**?

- Build trust & transparency

  - Help human understand AI decisions

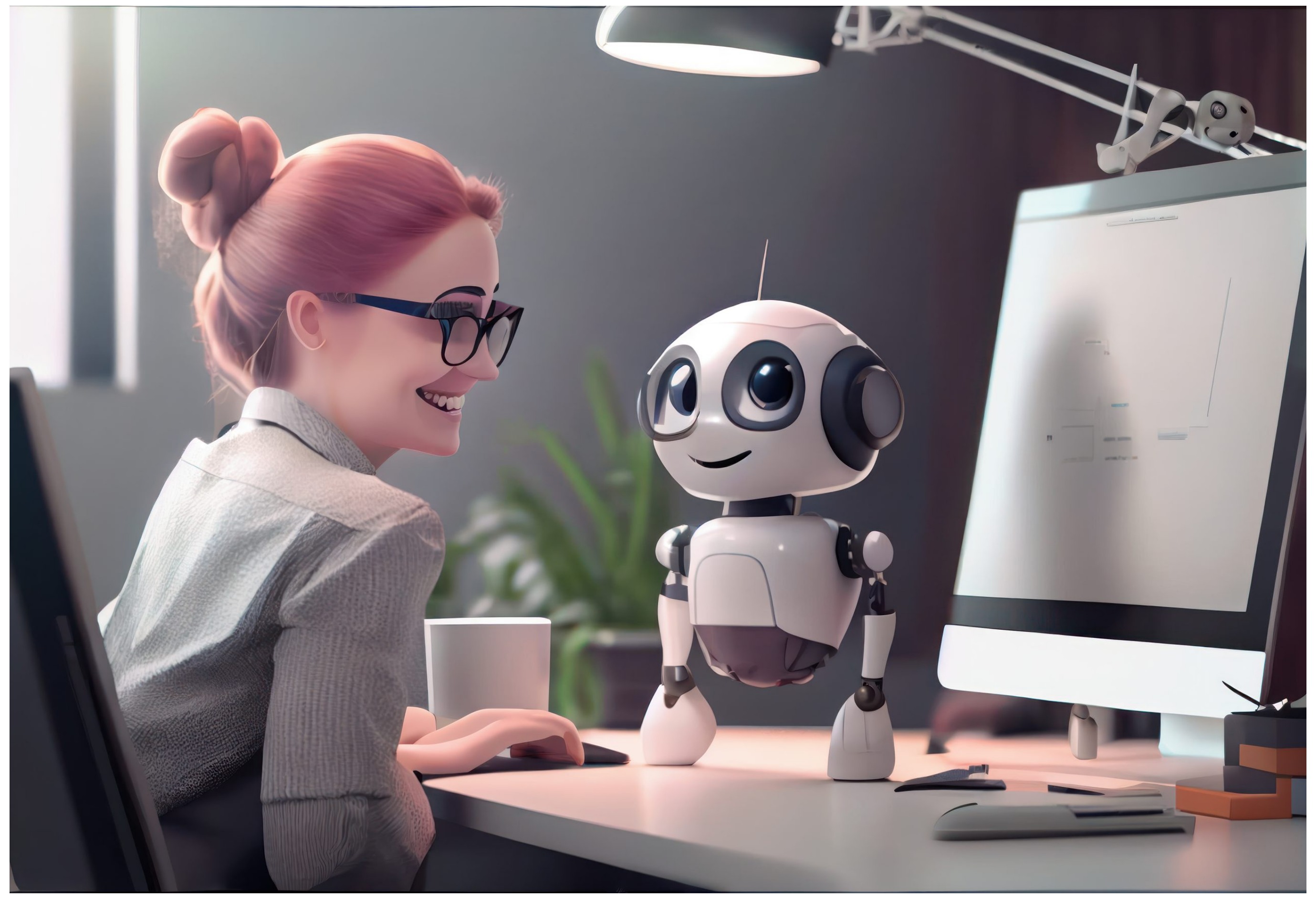
  - E.g. assist doctor diagnosis

What is **effective** communication and why **important**?

- Deliver the right amount of information

  - Enhance human efficiency

  - E.g. personalized education

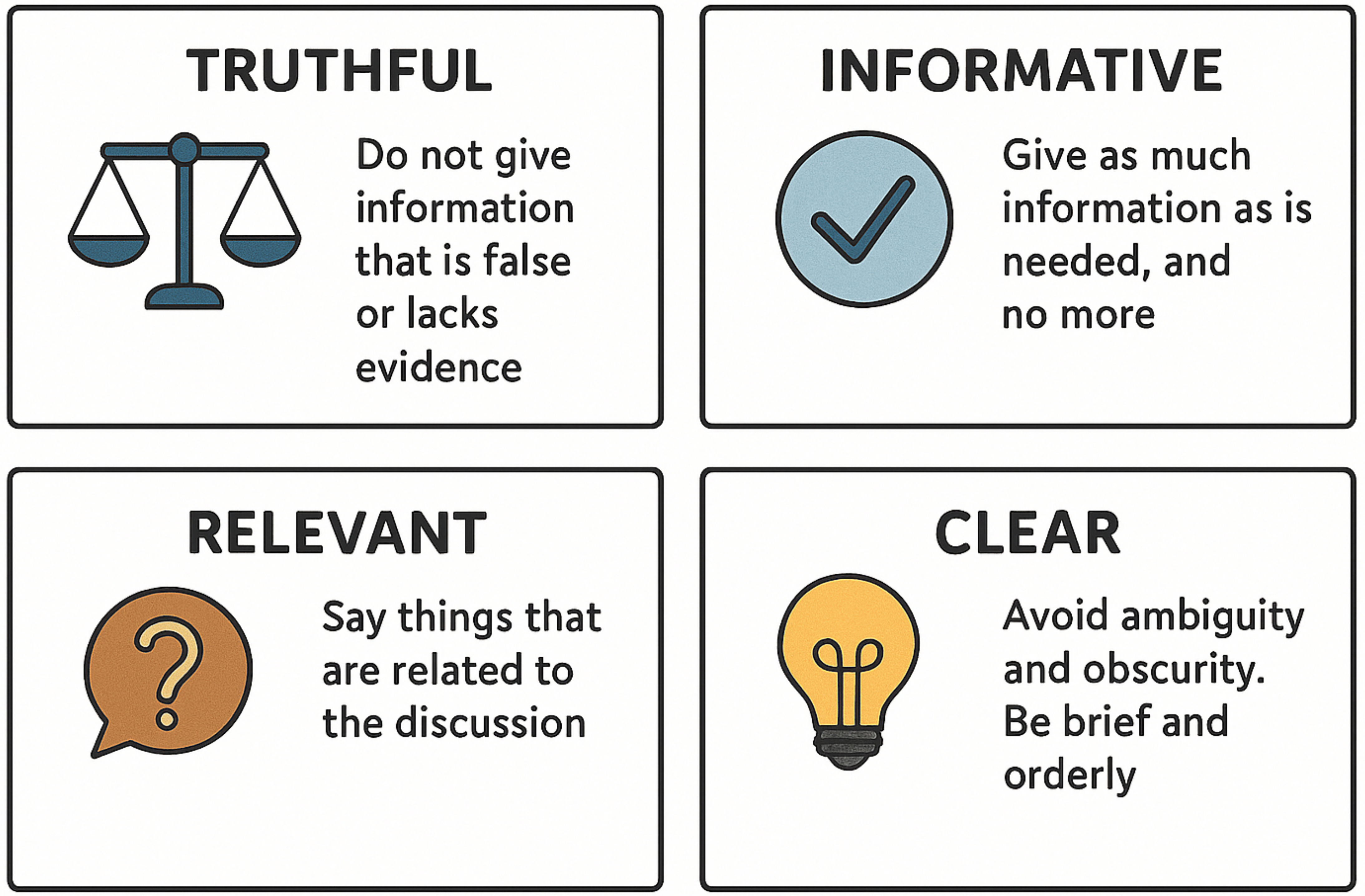


How can we achieve effective human-AI communication?

Our approach: Resemble human-human communication

Motivation: **Ingredients** for **effective communication**

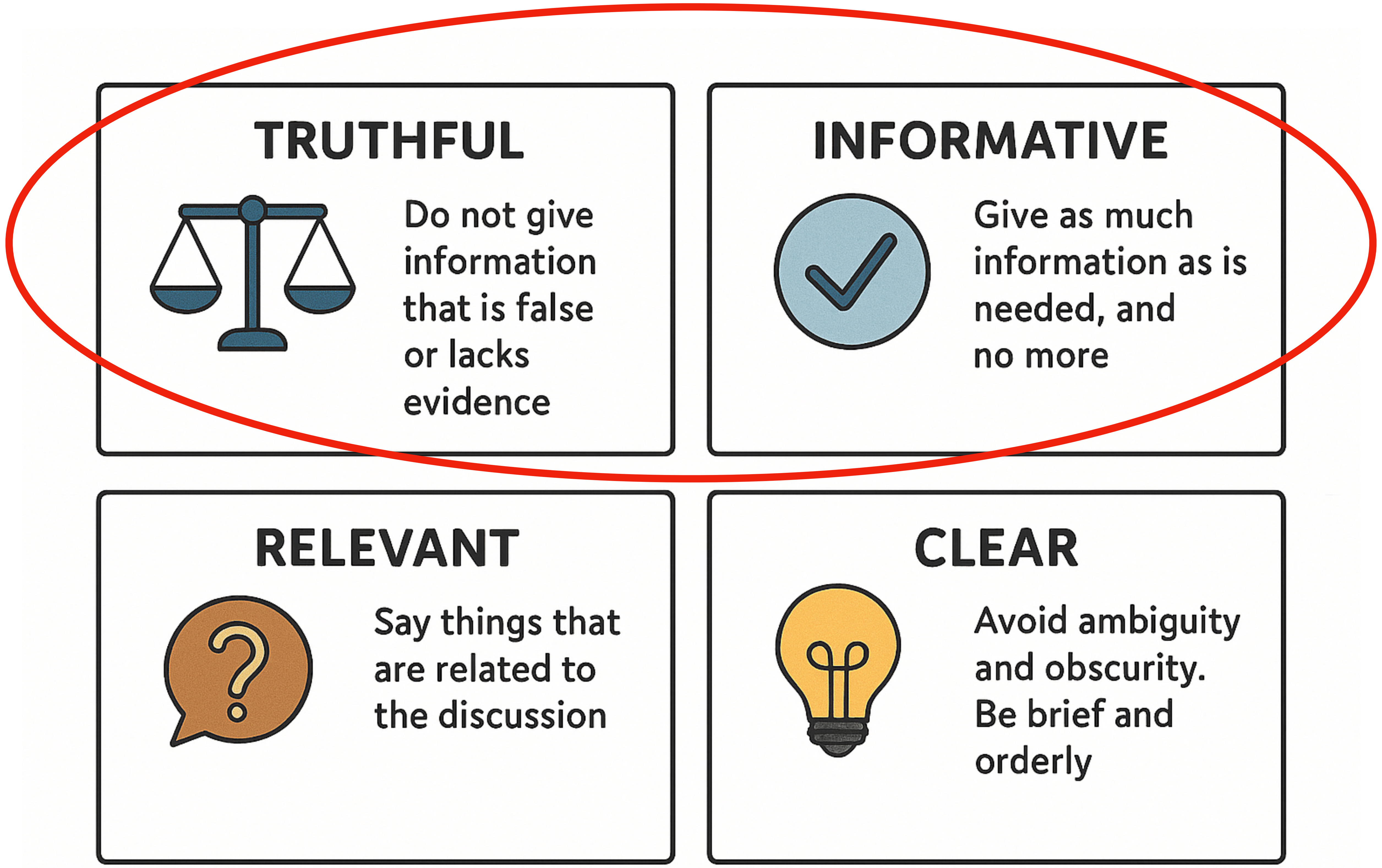
- Grice's maxims of conversation [1] :



[1] Grice, Herbert Paul (1975). "Logic and conversation". Syntax and semantics.

Motivation: **Ingredients** for **effective communication**

- Grice's maxims of conversation [1] :



[1] Grice, Herbert Paul (1975). "Logic and conversation". Syntax and semantics.

Focus on **improving**

Truthful Maxim

- Generate more faithful explanations (EMNLP 25)
- Communicate uncertainty more effectively (EMNLP 24 & 23)

Informative Maxim

- Evaluate and improve cognitive capabilities for instruction generation (ACL 23)
- Culture pragmatics (ongoing)

Data is all you need

~~Data is all you need~~

Human annotation: not available / unreliable

Costly / difficult to collect

Our approach: circumvent annotation needs

Focus on **improving**

Truthful Maxim

- Generate more faithful explanations (EMNLP 25)
- Communicate uncertainty more effectively (EMNLP 24 & 23)

Informative Maxim

- Evaluate and improve cognitive capabilities for instruction generation (ACL 23)
- Culture pragmatics (ongoing)

A Necessary Step toward Faithfulness: Measuring and Improving Consistency in Free-Text Explanations

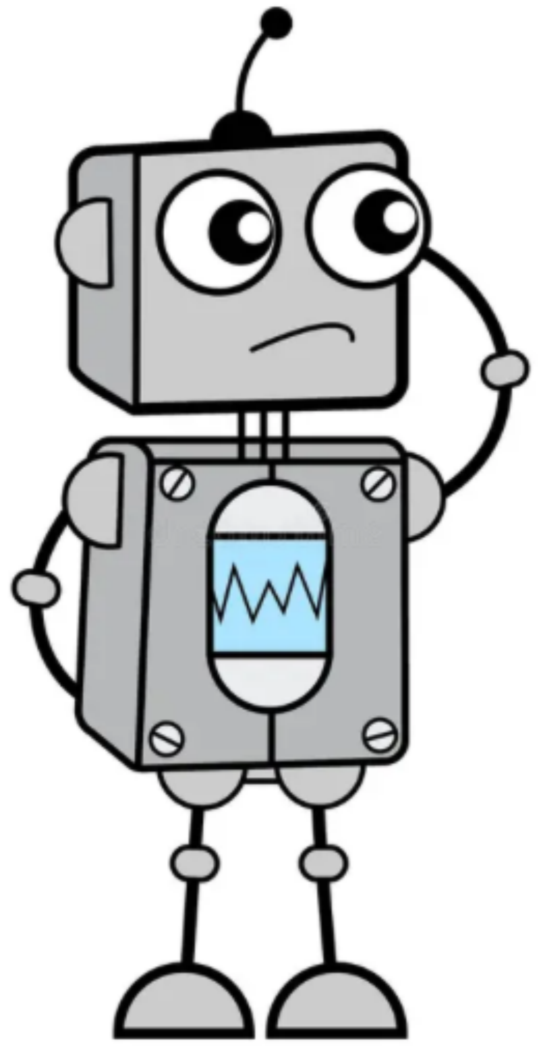
Lingjun Zhao Hal Daumé III

EMNLP 2025 Main Conference

Motivation: Explainable AI system

Q: Shall we admit this student?

Input →



Reasoning →

Yes!

Prediction / Output

- Explanation: reflects model's reasoning process

- **Faithful** explanation: **accurately** reflects model's **true** reasoning process

Why **faithful** explanation is **important**?

- Enhance AI transparency & accountability

  - *High-stake* decision making: healthcare, law, hiring decisions…

- Support human learning from AI

  - Some tasks AI is good at, human not naturally good at

- Our focus: *free-text* explanation — understandable by human

Challenges of generating faithful explanations

- Do not know how a model makes predictions

  - Especially for deep neural networks

  - Can't rely on human annotation: Conflate *faithfulness* and *plausibility*

    How convincing explanation appears

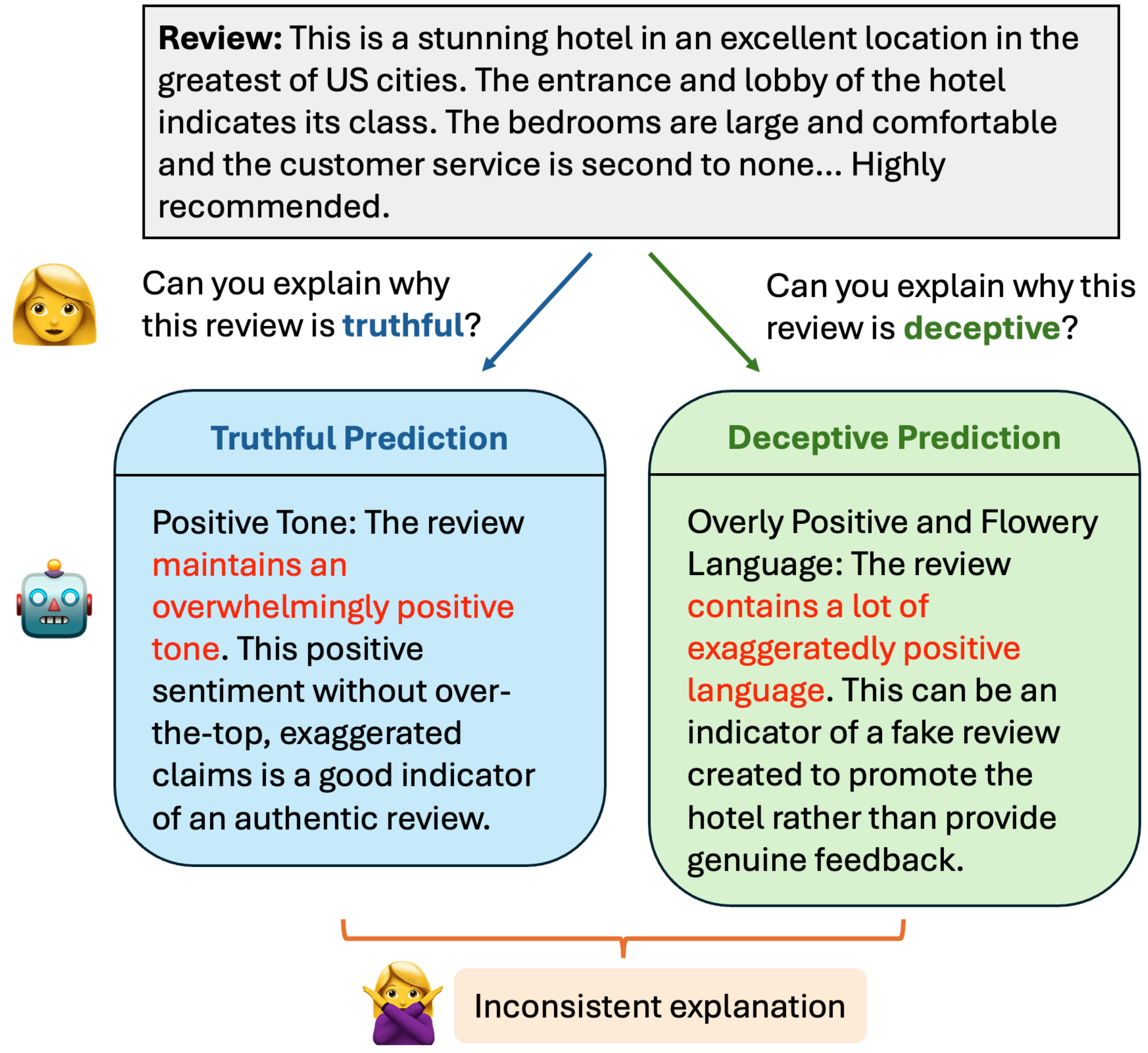
- Can't measure explanation faithfulness directly

  - I.e. Can't compute a faithfulness score for each explanation

Can we instead measure some **necessary** condition for explanation faithfulness?

Observation: language models generate **inconsistent** explanations

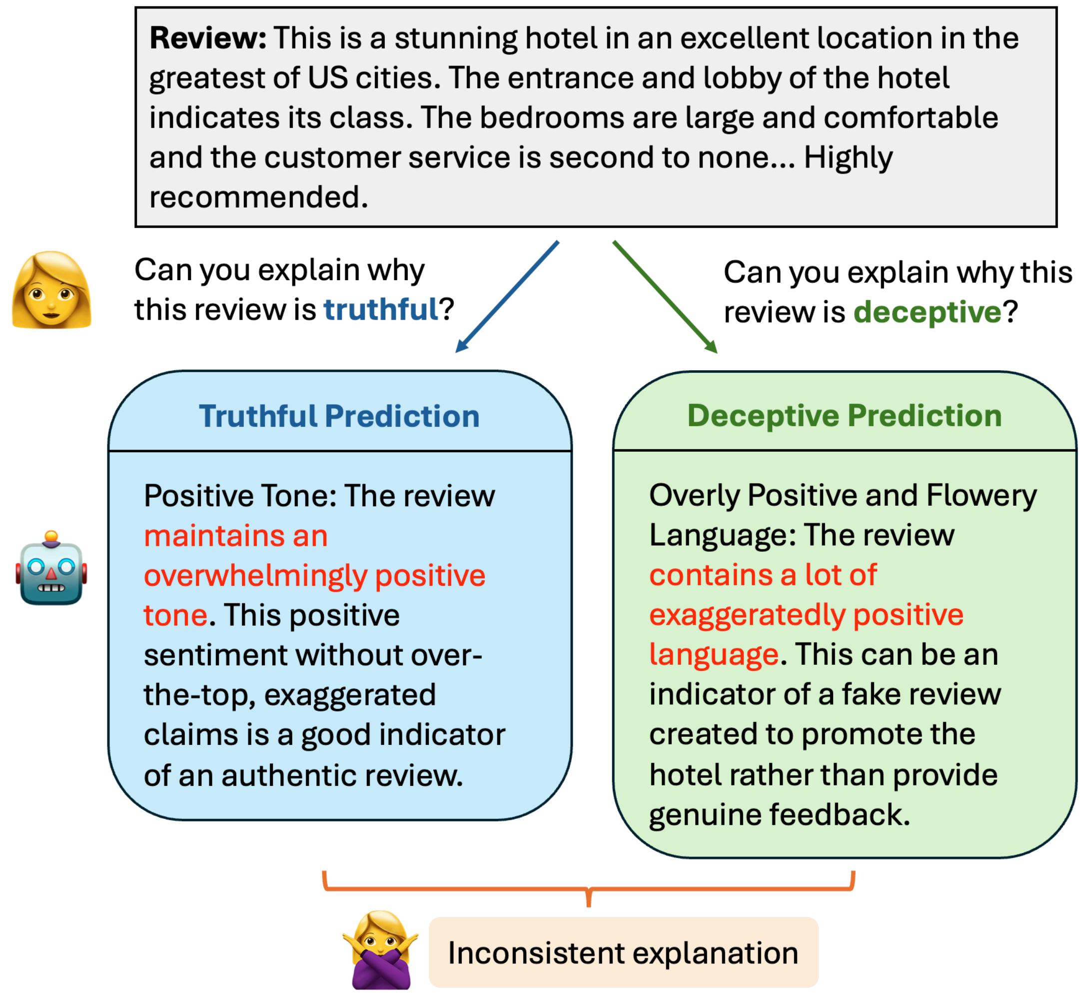
- **Inconsistent**: support a prediction and its negation
- Consistency: *necessary* for **faithfulness** [1]
- **Task**: Generate explanations to justify opinion spam detection
  - No external knowledge
  - Human not naturally good at



Example of GPT-4 model generating inconsistent explanations for truthful or deceptive prediction about a hotel review's authenticity: both the truthful and deceptive explanations rely on the same evidence "use a lot of positive language".

[1] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence. 2019

But how to **measure** this consistency for a given explanation?



We introduce a measure: Prediction-EXplanation (PEX) consistency — extending the concept of weight of evidence [1]

[1] Melis DA, Kaur H, Daumé III H, Wallach H, Vaughan JW. From human explanation to model interpretability: A framework based on weight of evidence. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 2021

Measuring Prediction-EXplanation (PEX) consistency

- **PEX consistency** ⚖️:

$$C(\boldsymbol{e}) = \log \frac{M(\boldsymbol{e} \mid Q(\boldsymbol{q}, \boldsymbol{a}))}{M(\boldsymbol{e} \mid Q(\boldsymbol{q}, \neg\boldsymbol{a}))}$$

  - Compare the likelihood of model M generating explanation e under different predictions: (a, ¬a)

$$C(e) = log \frac{M(\text{The review maintains an overwhelmingly positive tone} \mid \text{Is this review truthful or deceptive? Review: \{review\}. Answer: \textbf{Truthful}. Question: Can you explain why the review is \textbf{truthful}?})}{M(\text{The review maintains an overwhelmingly positive tone} \mid \text{Is this review truthful or deceptive? Review: \{review\}. Answer: \textbf{Deceptive}. Question: Can you explain why the review is \textbf{deceptive}?})}$$

- But computing this probability needs density estimation: not reliable enough

Measuring Prediction-EXplanation (PEX) consistency

- **Adjusted consistency** using Bayes's rule:

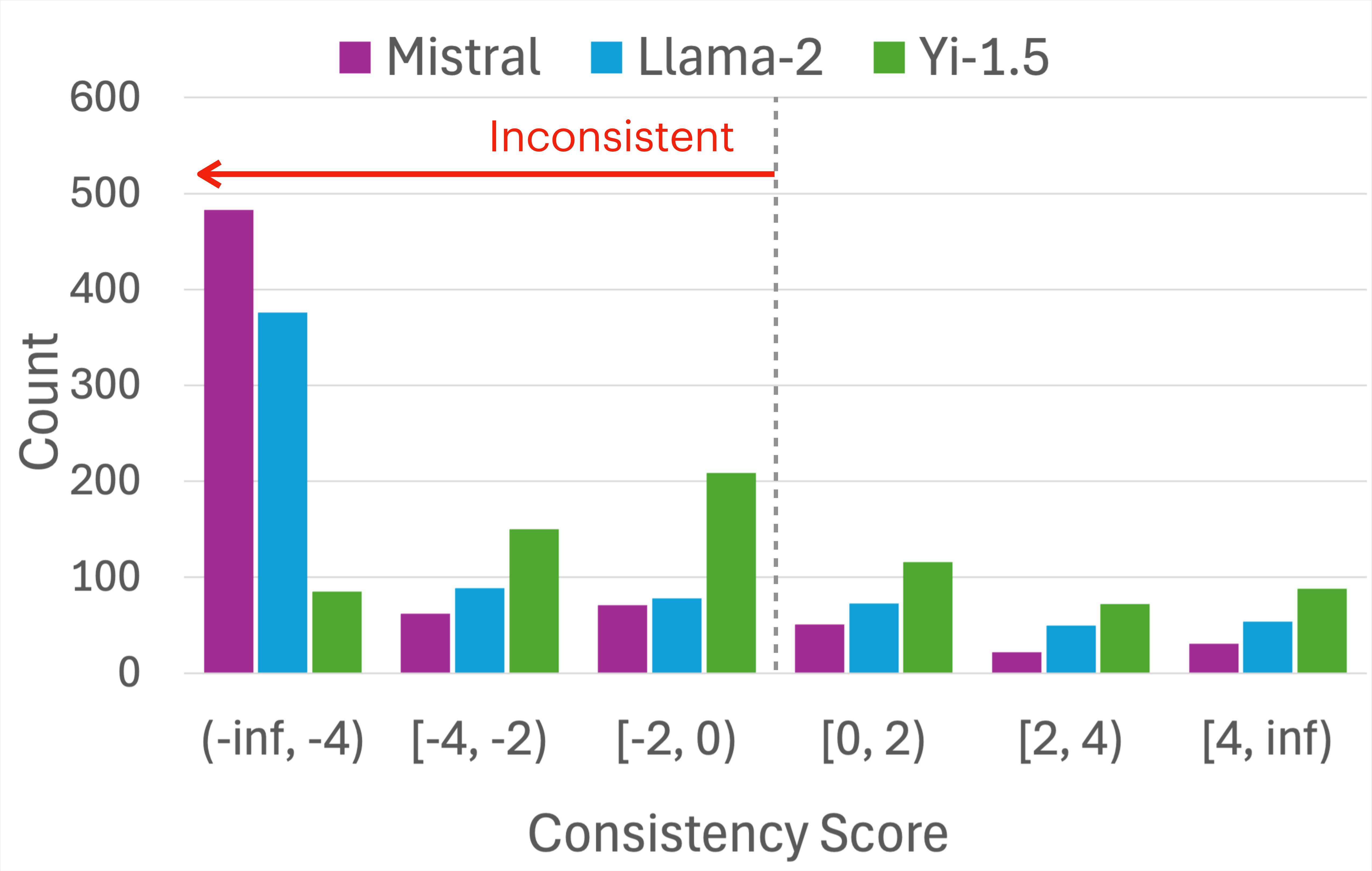
$$C'(\boldsymbol{e}) = \log \frac{M(\boldsymbol{a} \mid Q'(\boldsymbol{q}, \boldsymbol{e}))}{M(\neg \boldsymbol{a} \mid Q'(\boldsymbol{q}, \boldsymbol{e}))} - \log \frac{M(\boldsymbol{a} \mid \boldsymbol{q})}{M(\neg \boldsymbol{a} \mid \boldsymbol{q})}$$

$$C'(e) = log \frac{M(\textbf{Truthful} \mid \text{Is this review truthful or deceptive? Review: \{review\}. Analysis: The review maintains an overwhelmingly positive tone})}{M(\textbf{Deceptive} \mid \text{Is this review truthful or deceptive? Review: \{review\}. Analysis: The review maintains an overwhelmingly positive tone})} - log \frac{M(\textbf{Truthful} \mid \text{Is this review truthful or deceptive? Review: \{review\}})}{M(\textbf{Deceptive} \mid \text{Is this review truthful or deceptive? Review: \{review\}})}$$

- Does not need density estimation

How **consistent** are the explanations generated by large language models?

Language models can generate **62%-86%** inconsistent explanations



Dataset:

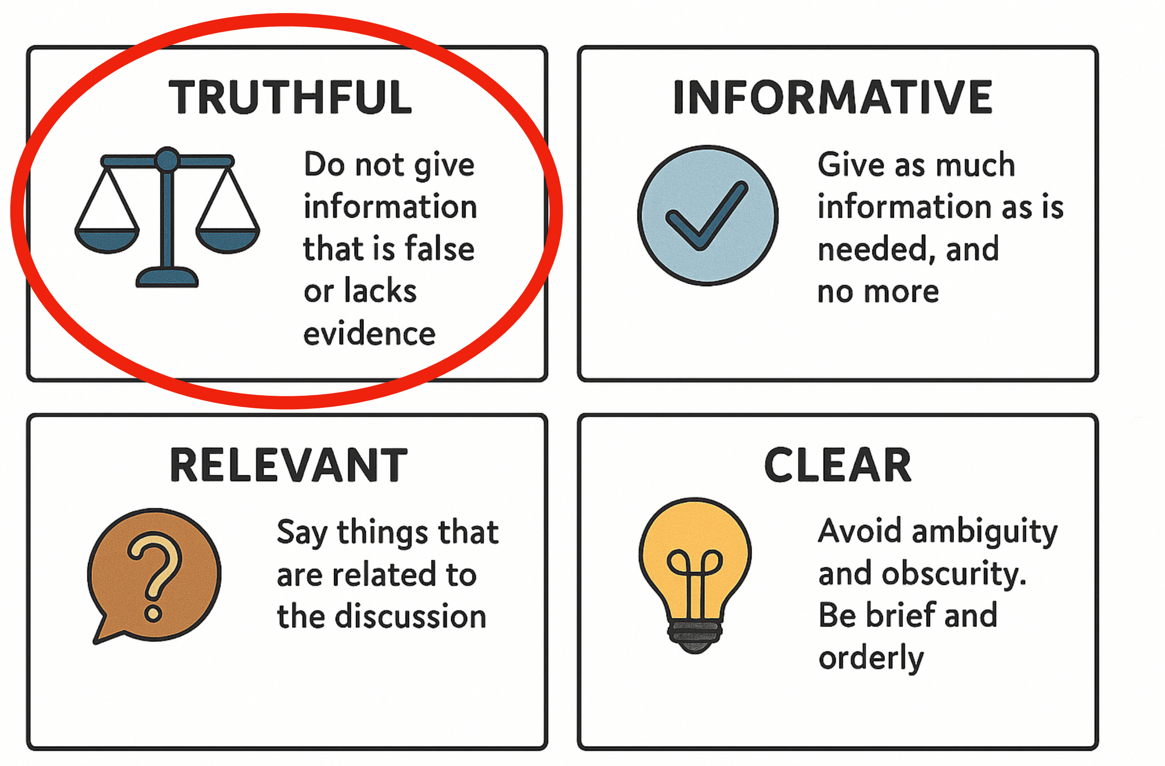
- TripAdvisor hotel review (320)

- Amazon product review (400)

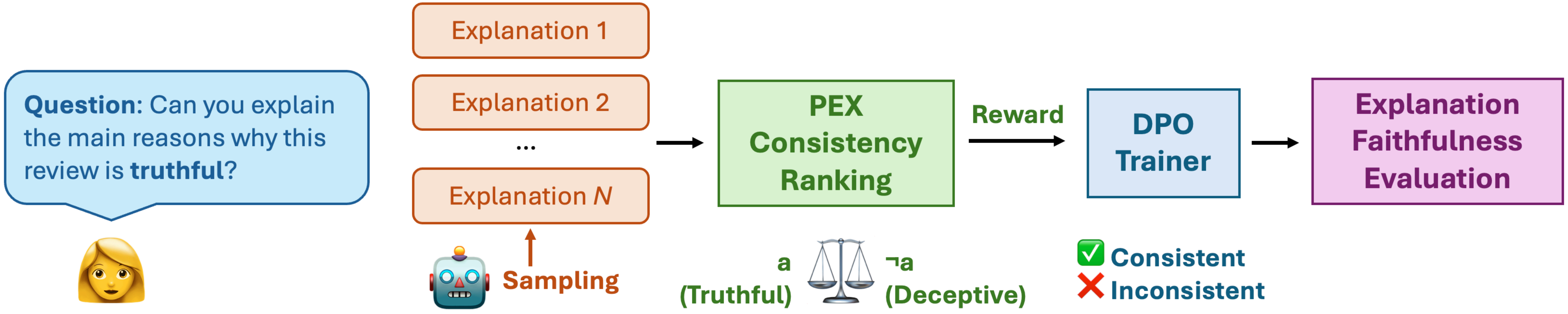
- Inconsistent: PEX score < 0

  i.e. explanation supports the negation prediction better than the model prediction

Can the consistency of LLM-generated explanations be **improved**?

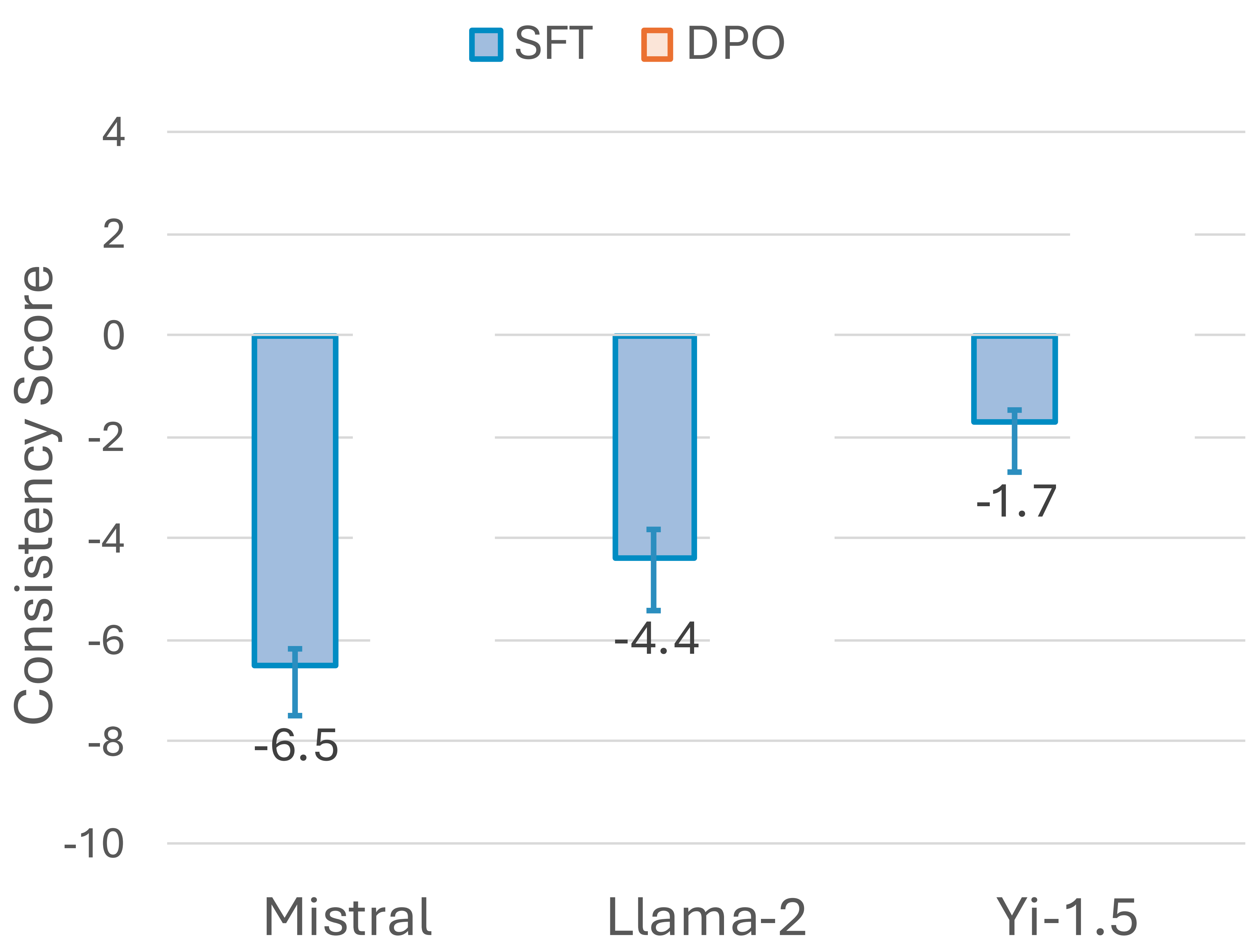


Generating more consistent explanations

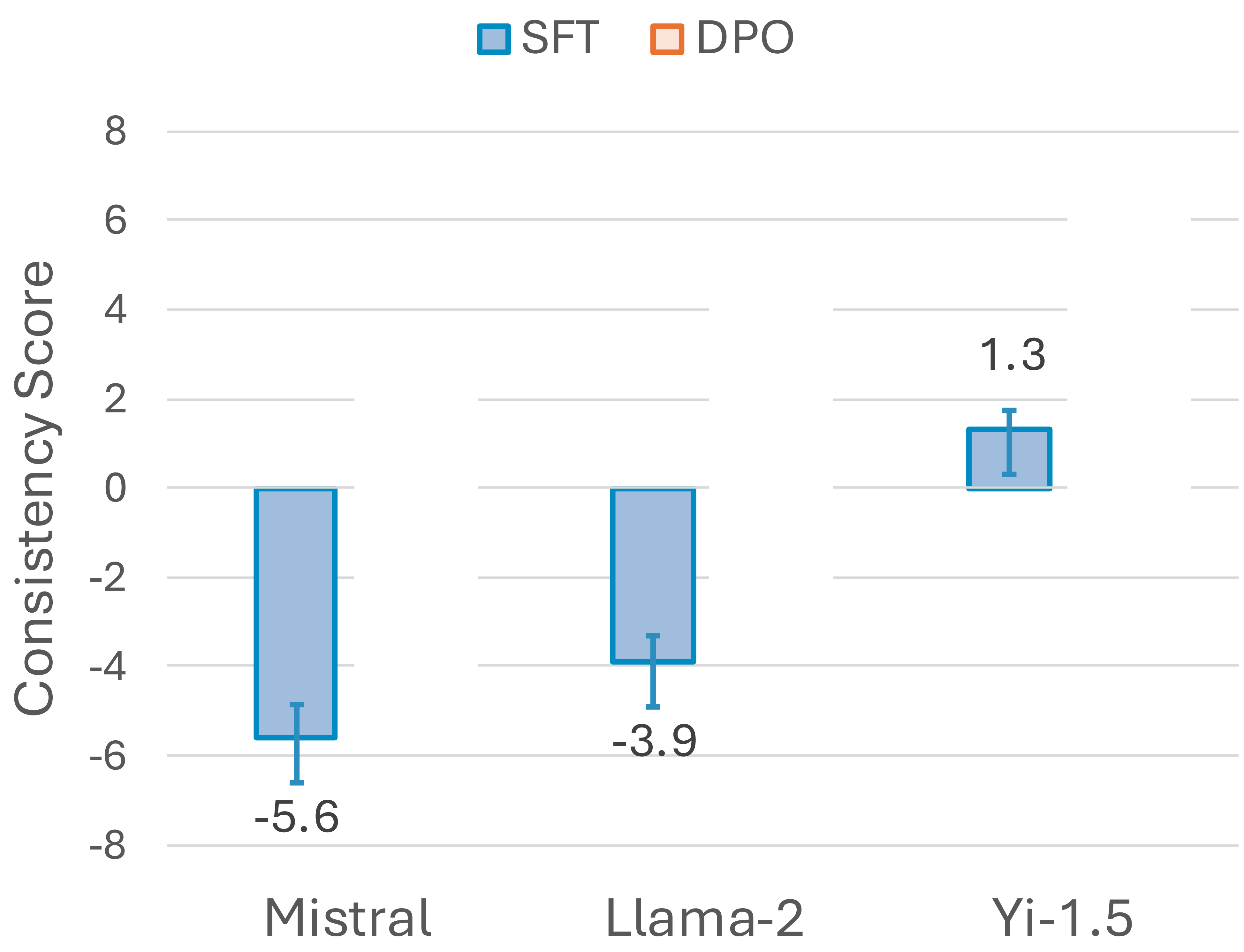


1. **Sampling** explanations from a language model
2. **Rank** explanations according to PEX consistency
3. **Optimize** explanation consistency using direct preference optimization (DPO):
   - *Preferred completion*: explanations with highest PEX consistency
   - *Dispreferred completion*: explanations with lowest PEX consistency
   - No human annotations needed

Optimizing explanation consistency with DPO: using PEX as signal

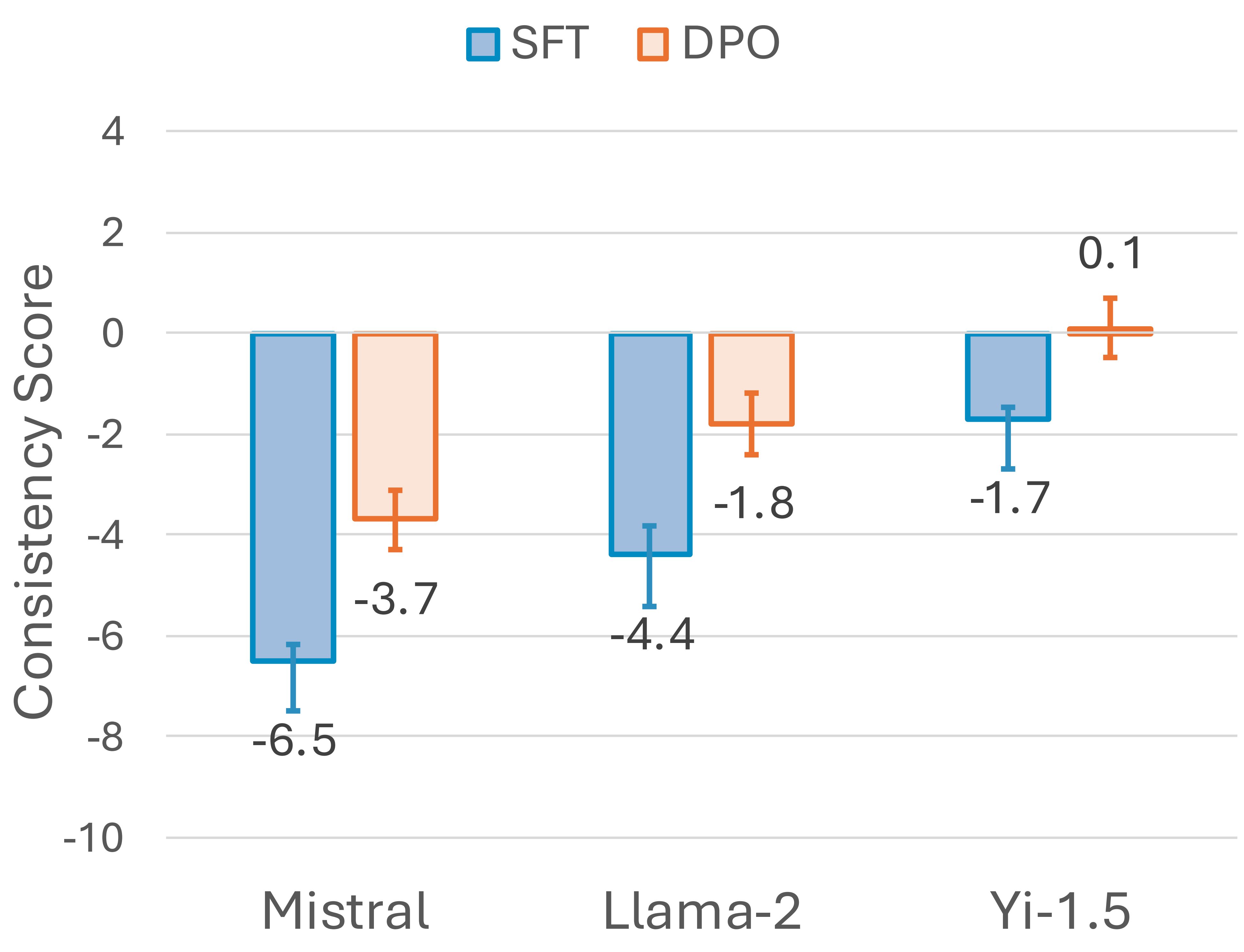


TripAdvisor hotel review dataset

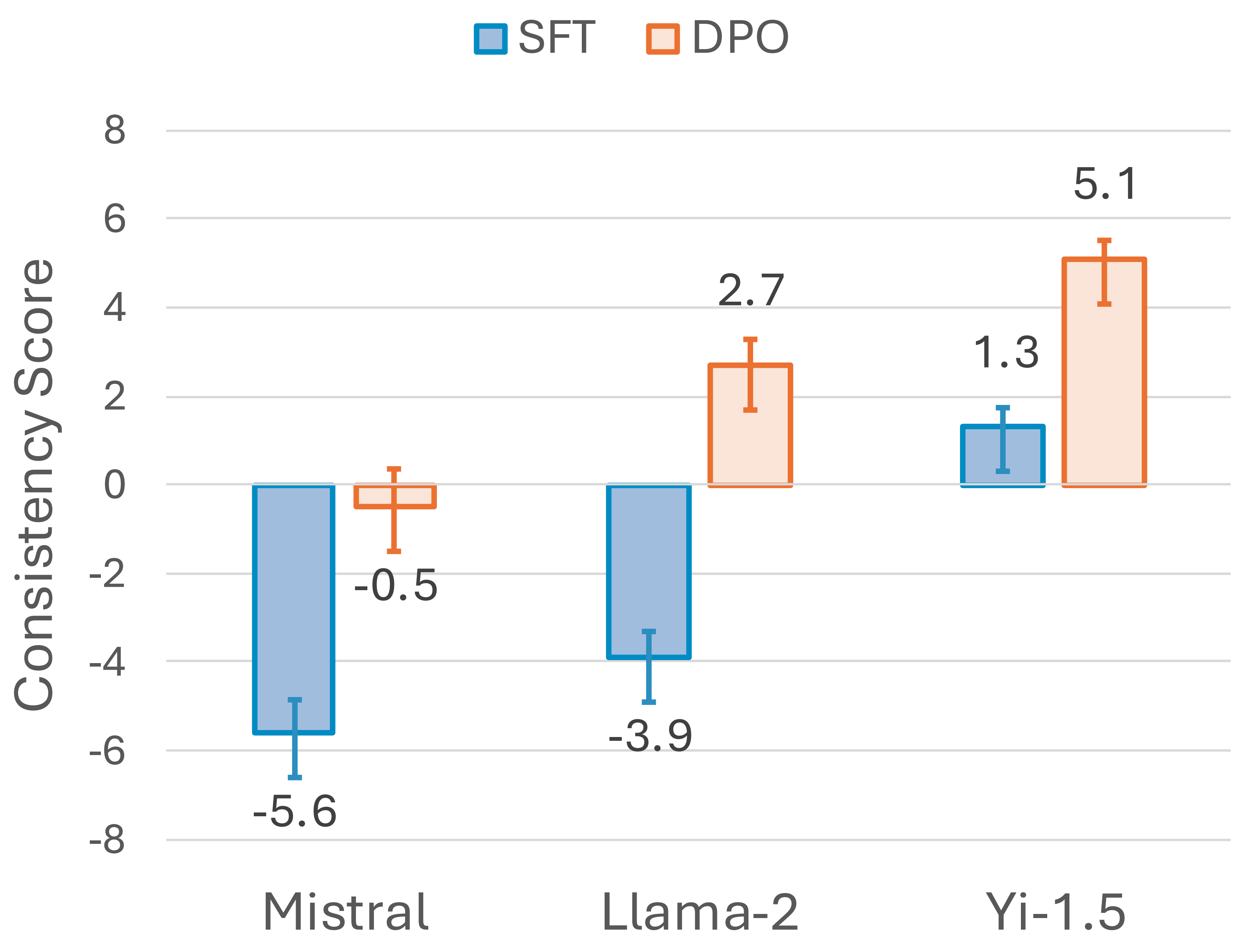


Amazon product review dataset

Optimizing explanation consistency with DPO: using PEX as signal



TripAdvisor hotel review dataset



Amazon product review dataset

- *Takeaway*: explanation consistency can be improved

Are consistency-optimized explanations also more **faithful**?

Accurately reflect the model's reasoning process

Need a faithfulness evaluation method

Faithfulness evaluation method: **simulatability**-based

- If model A’s explanations are more faithful

  ⇒ easier for model B to mimic model A’s prediction

  by using A’s explanation [1]

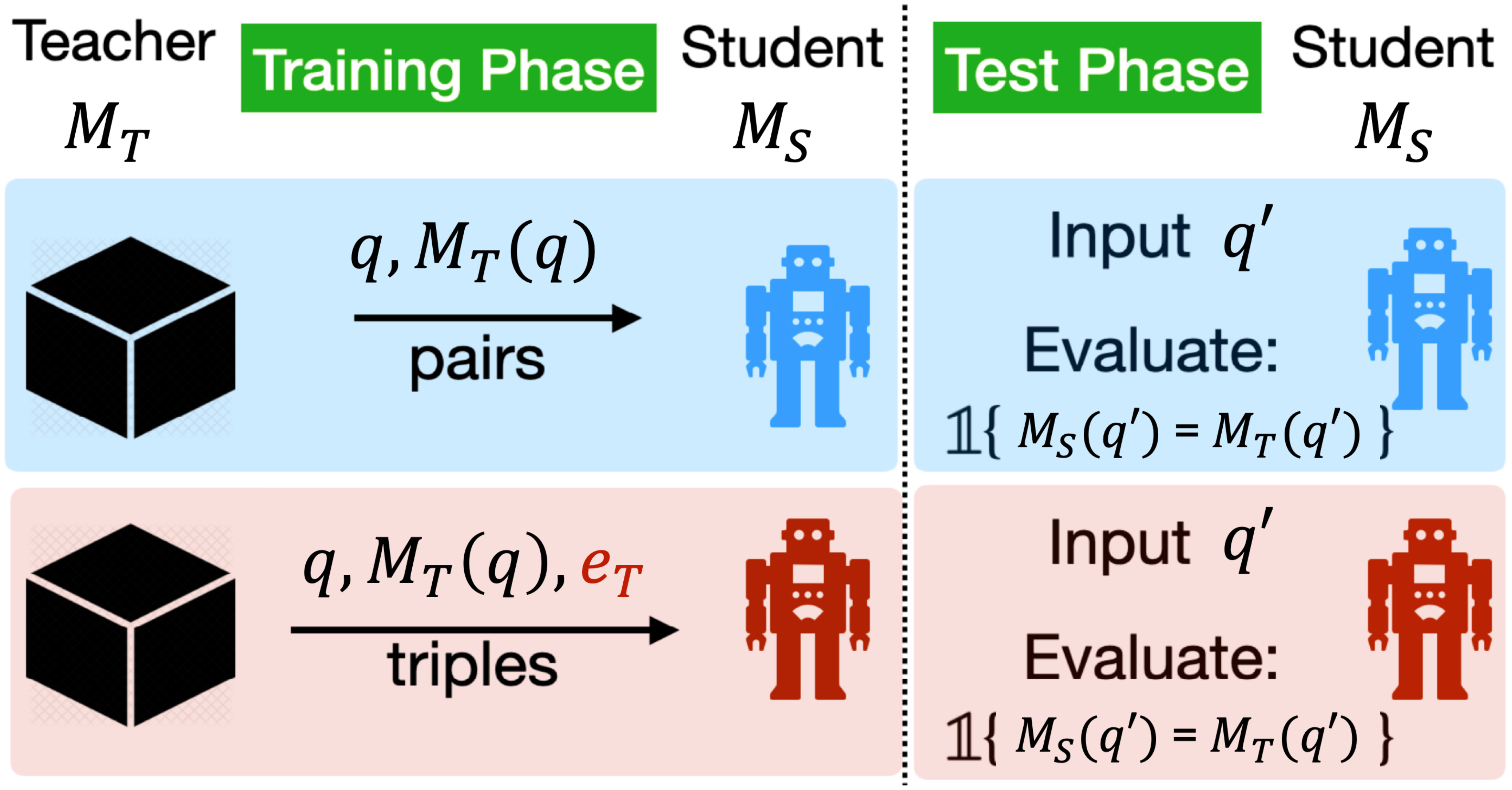
[1] Lyu Q, Apidianaki M, Callison-Burch C. Towards faithful model explanation in NLP: A survey. Computational Linguistics. 2024

Faithfulness evaluation method: **simulatability**-based

- **Student model**:
  - Training: use provided prediction + explanation from **teacher model**
  - Testing: *no* prediction/explanation provided

- **Eval metric**: student model test set F1 score (simulation performance)

- System-level evaluation

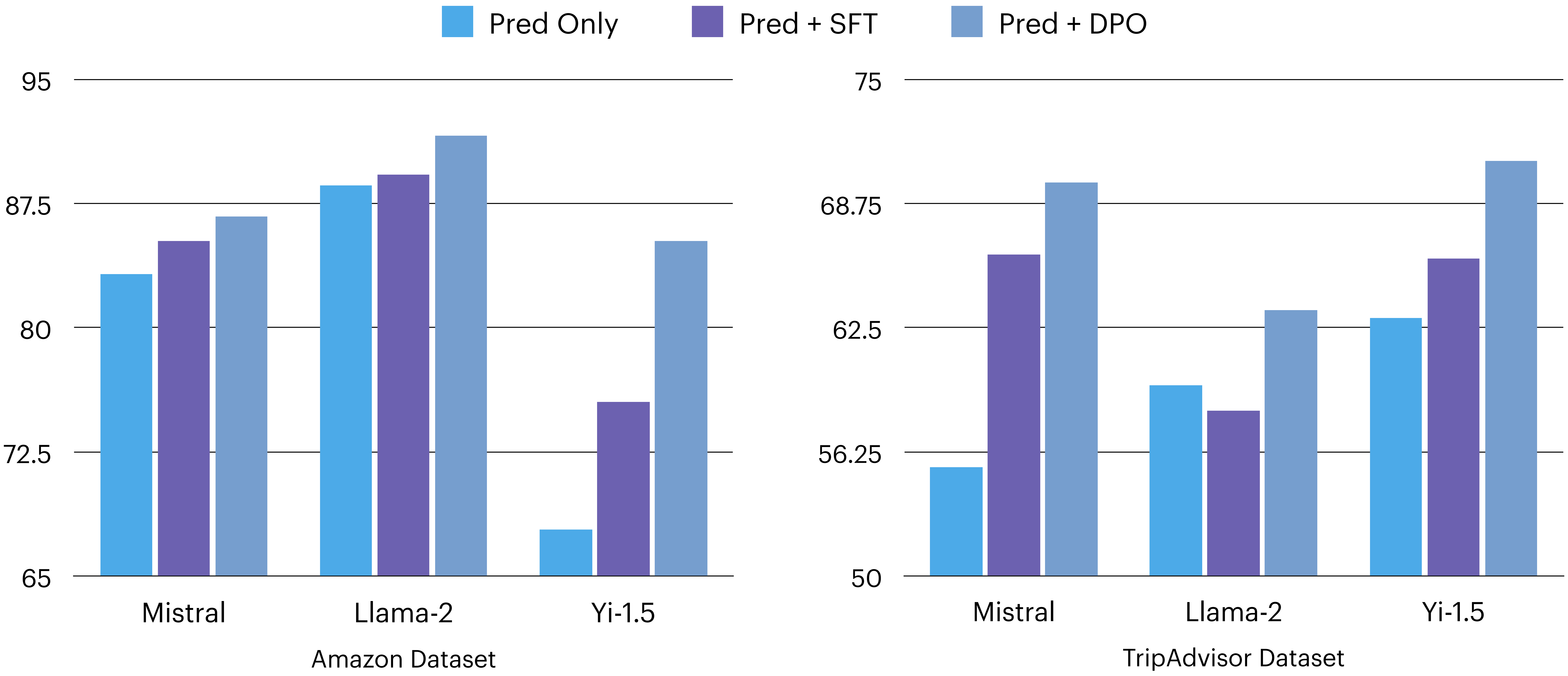


Explanation evaluation framework (figure reproduced from [1])

[1] Pruthi D, Bansal R, Dhingra B, Soares LB, Collins M, Lipton ZC, Neubig G, Cohen WW. Evaluating explanations: How much do explanations from the teacher aid students?. Transactions of the Association for Computational Linguistics. 2022

Optimizing PEX consistency **improves** explanation faithfulness: **1.5%-9.7%**

- Student model simulation performance (F1):



Takeaways

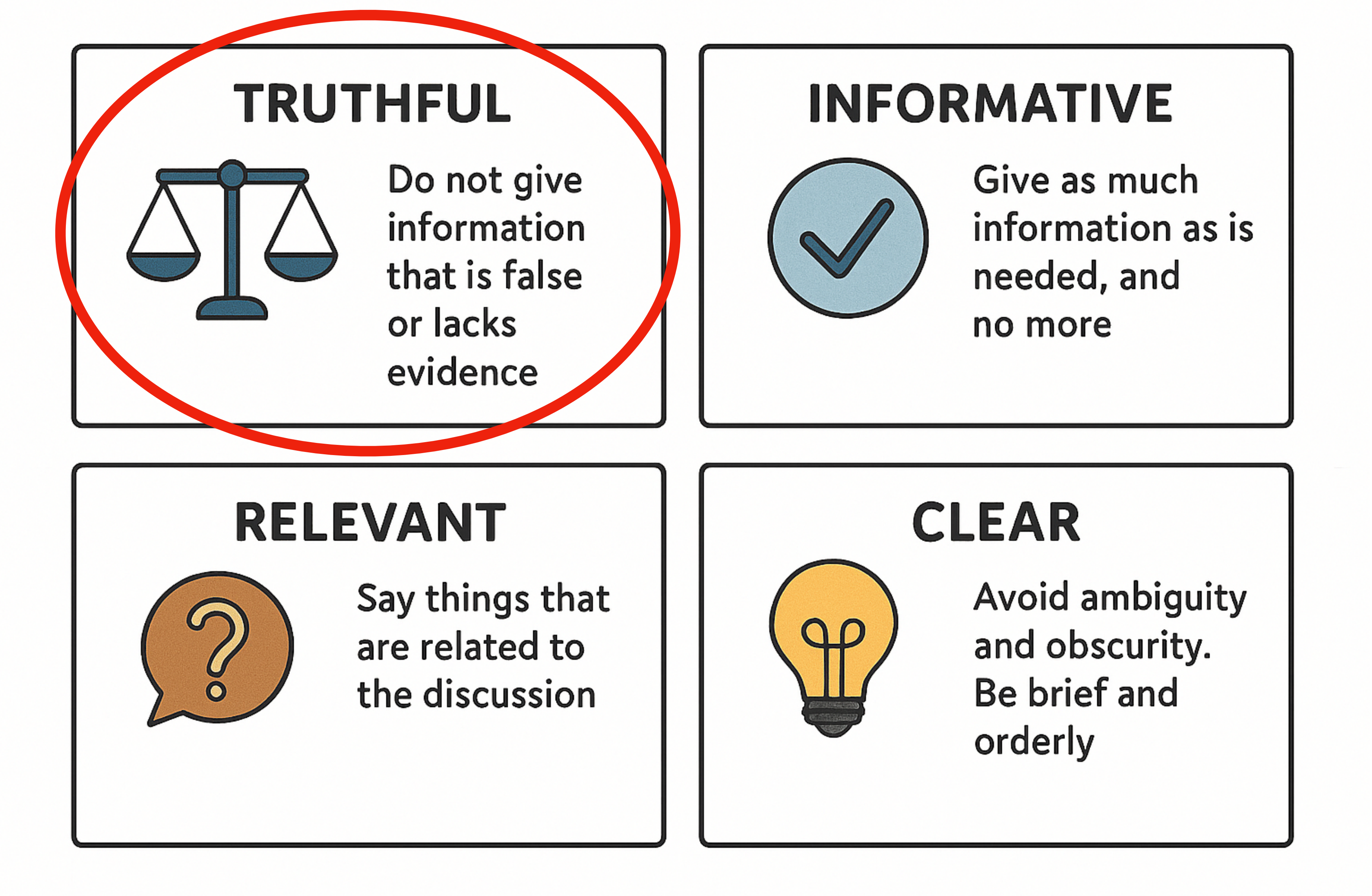
- Introduce Prediction-EXplanation (PEX) consistency:

  - 3 language models generate 62-86% inconsistent explanations

  ⇒ Undermine faithfulness

- Training approach: generate more consistent explanations

  ⇒ more faithful explanations: up to 10%



Focus on **improving**

Truthful Maxim

- Generate more faithful explanations (EMNLP 25)
- Communicate uncertainty more effectively (EMNLP 24 & 23)

Informative Maxim

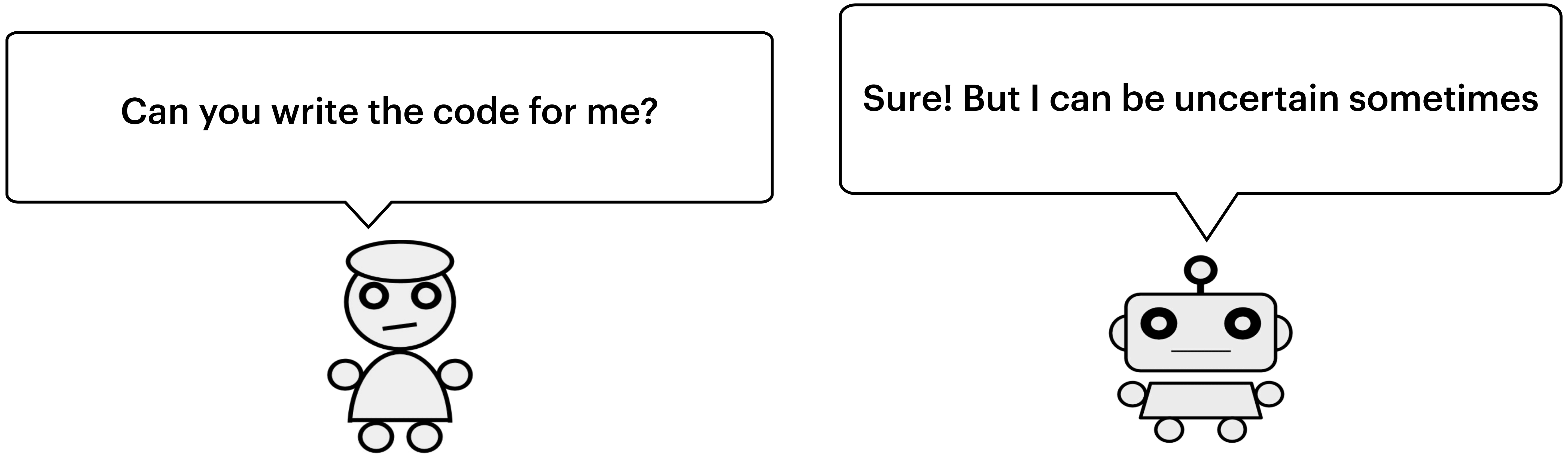
- Evaluate and improve cognitive capabilities for instruction generation (ACL 23)
- Culture pragmatics (ongoing)

Successfully Guiding Humans with Imperfect Instructions by Highlighting Potential Errors and Suggesting Corrections

Lingjun Zhao Khanh Nguyen Hal Daumé III

EMNLP 2024 Main Conference

Why is human-AI collaboration **important**?



Use experience/values to make informed decisions

Analyze data and provide solution quickly

- AI can make mistakes: e.g. language models <u>hallucinate</u>:

  Generate output factually incorrect, or not grounded with input

- Human as final decision maker: refer to AI's outputs and use their own judgement

  ⇒ Achieve better outcome

How to better support human-AI collaboration?

- Our approach (hypothesis): **communicate uncertainty** information more effectively
  - Goal: better human decision-making

- Why:
  - Clarify AI's limitations

  ⇒ Help human know when to trust AI / use own judgement

How to provide uncertainty information to assist humans?

- **Task**: Human navigate to a target location
  - Guided by a language model
  - Long horizon decision-making

- Evaluate AI communication efficacy:
  - **Human evaluation**: navigate using web interface
  - Measurable human's performance gain

- **Approach**: highlight potential hallucination spans

When to trust AI / use own judgement

Green box: ground truth destination

Walk past the couch and stop in front of the TV.

How to provide uncertainty information to assist humans?

Green box: ground truth destination

Walk past the couch and stop in front of the TV.

↓

Hallucination detection model

Is the span hallucination?

not grounded with input

Yes: highlight

No: no highlight

Problem: don't have human annotation

How to detect span-level hallucinations **without** human annotation?

Detecting **span-level hallucinations** without human annotation

- Tried a few unsupervised approaches: not working well
- Weakly supervised training approach:
  - Training: create **synthetic data** to train a hallucination detection model
  - Testing: actual language model-generated instructions

Creating **synthetic dataset** for training span-level hallucination detection model

- Each visual path: has human-written instruction

  - Create synthetic span-level hallucinations (different types)

When you see a couch, turn right, stop next to the bed

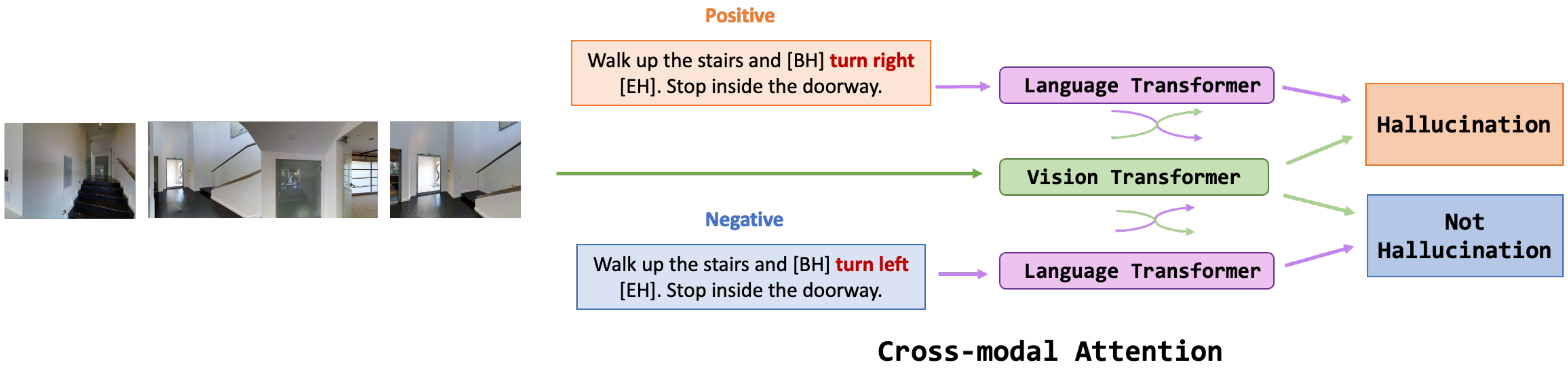
Human-written instruction

When you see a bed, turn right, stop next to the couch

Synthetic hallucination: swap objects

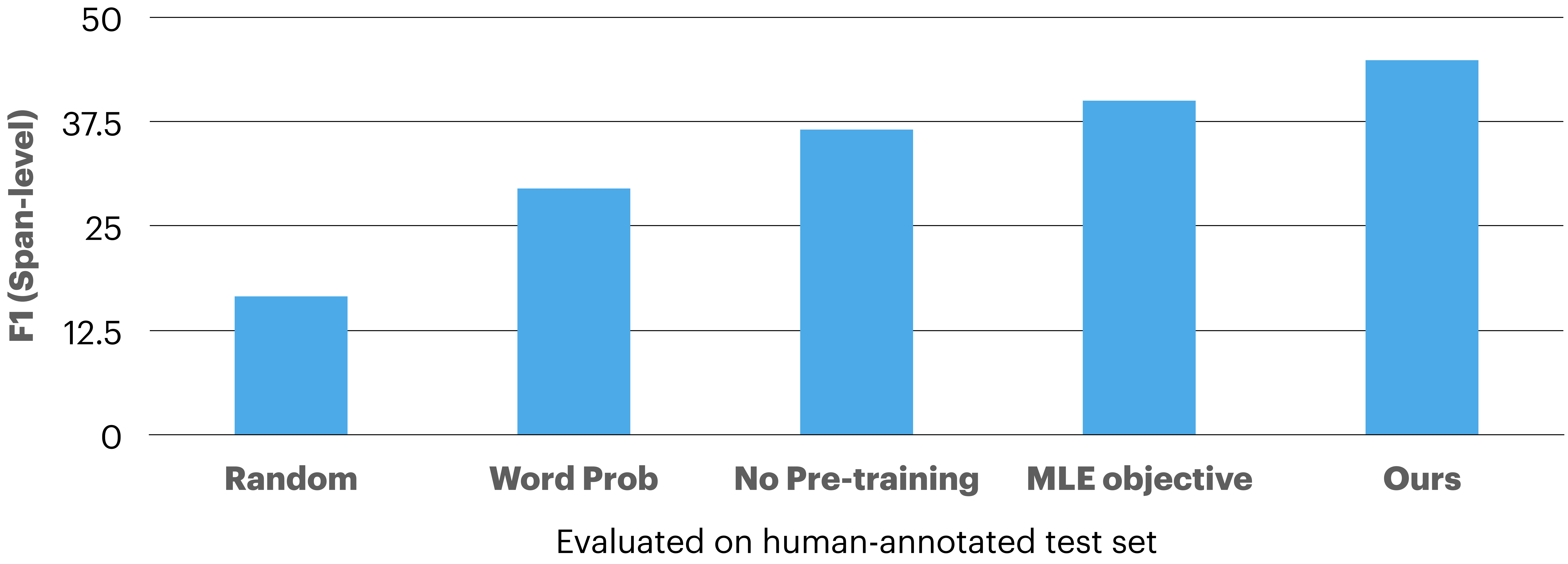
Green box: ground truth destination

Span-level hallucination detection model



- Initialization from a pre-trained visual-language model
  - Span representation: use special tokens ← pre-GPT technique
- Contrastive learning: distinguish hallucinated instruction from correct instruction
- Output: span-level **hallucination score** (normalized visual-text similarity score)

Model detects span-level hallucinations reasonably well

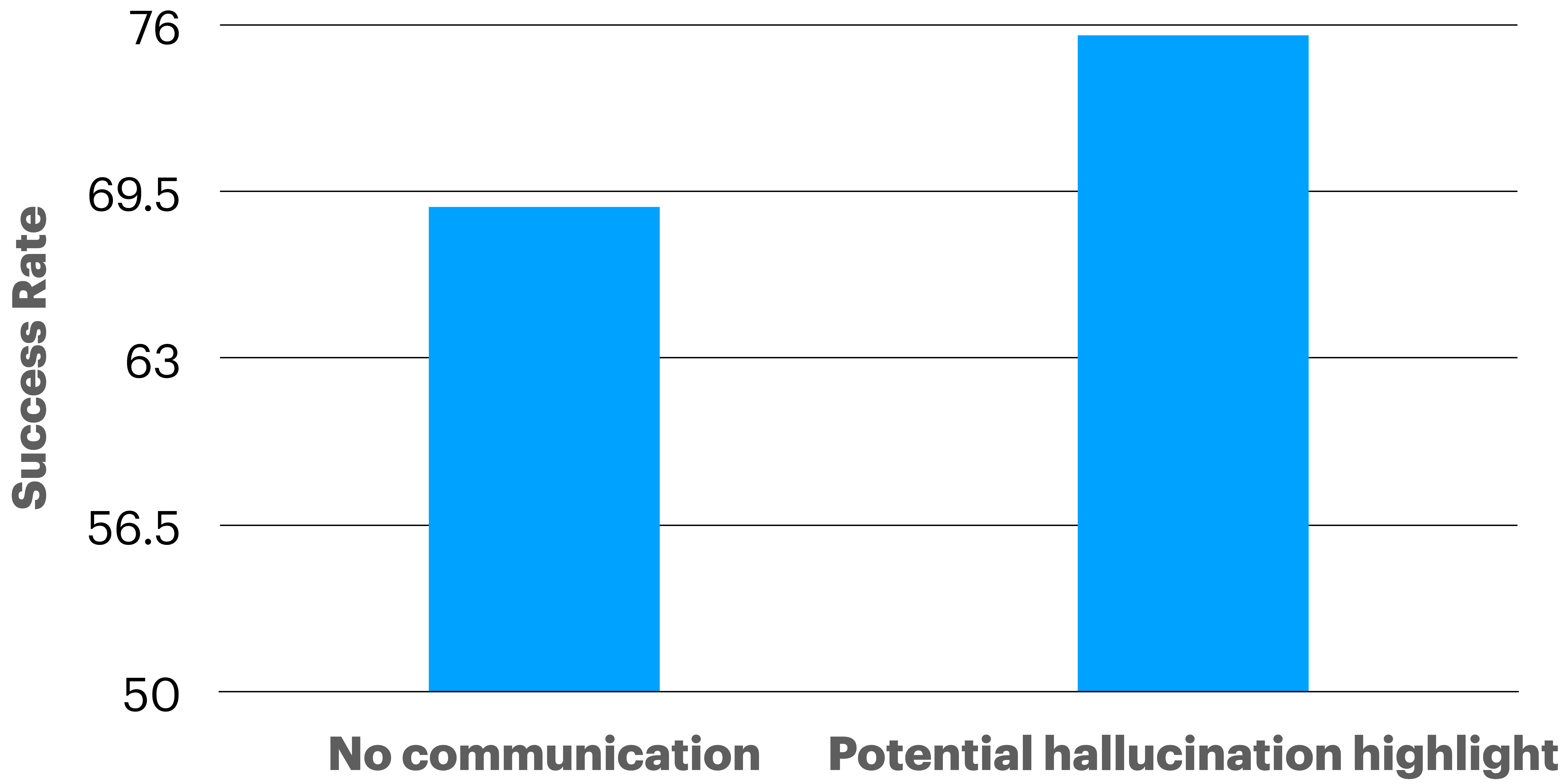


Does providing potential hallucination highlights improve **human** task performance?

Walk past the couch and stop in front of the TV.

Human evaluation: navigate using web interface

Potential hallucination highlights **improve 6.7%** human performance



Problem: Some users report don't know how to fix AI's mistakes

How to communicate to humans
how to **fix AI’s mistakes**?

Hallucination dEtection And Remedy (HEAR): **Rich** uncertainty communication

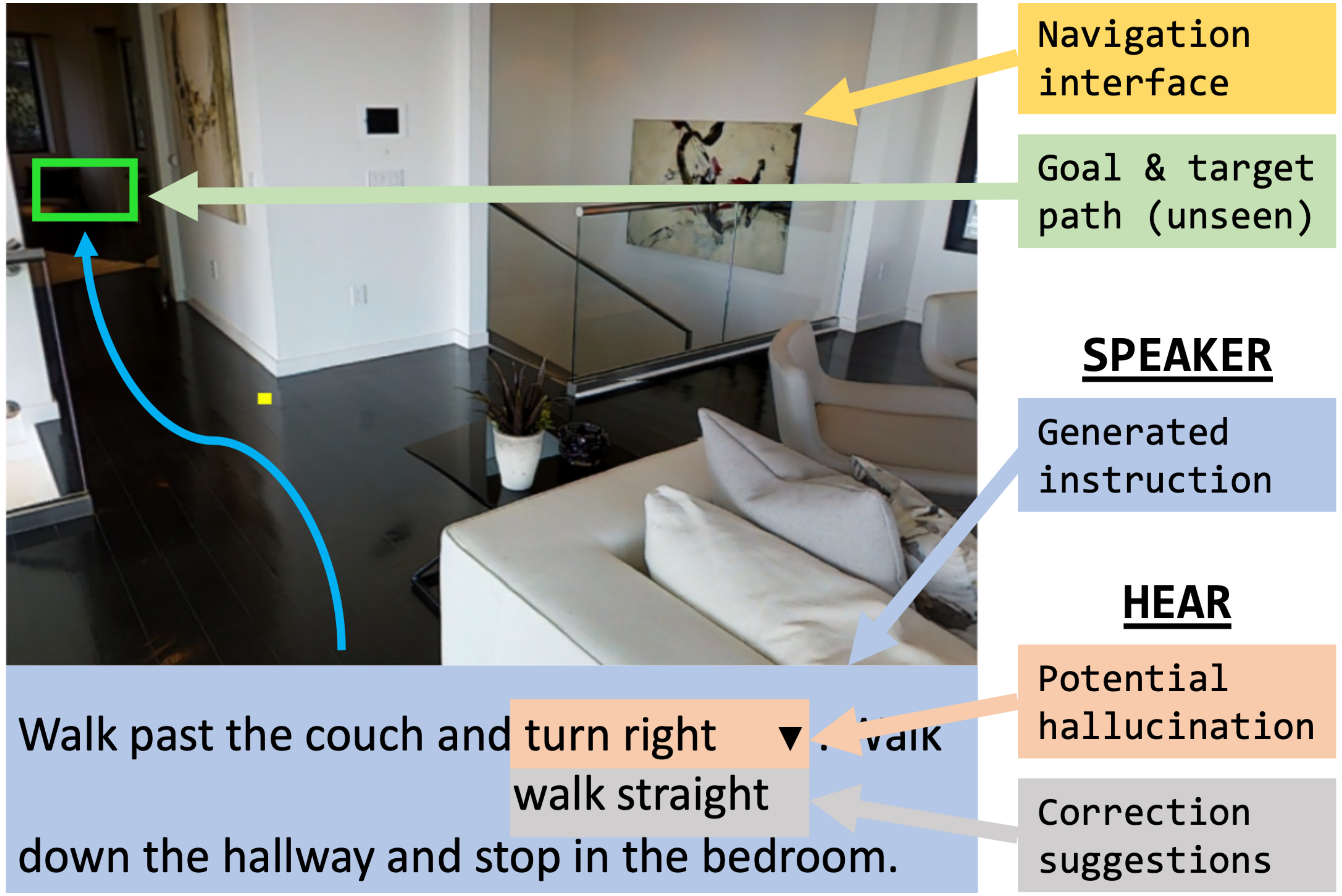
- Present to humans:

  - Potential hallucination spans

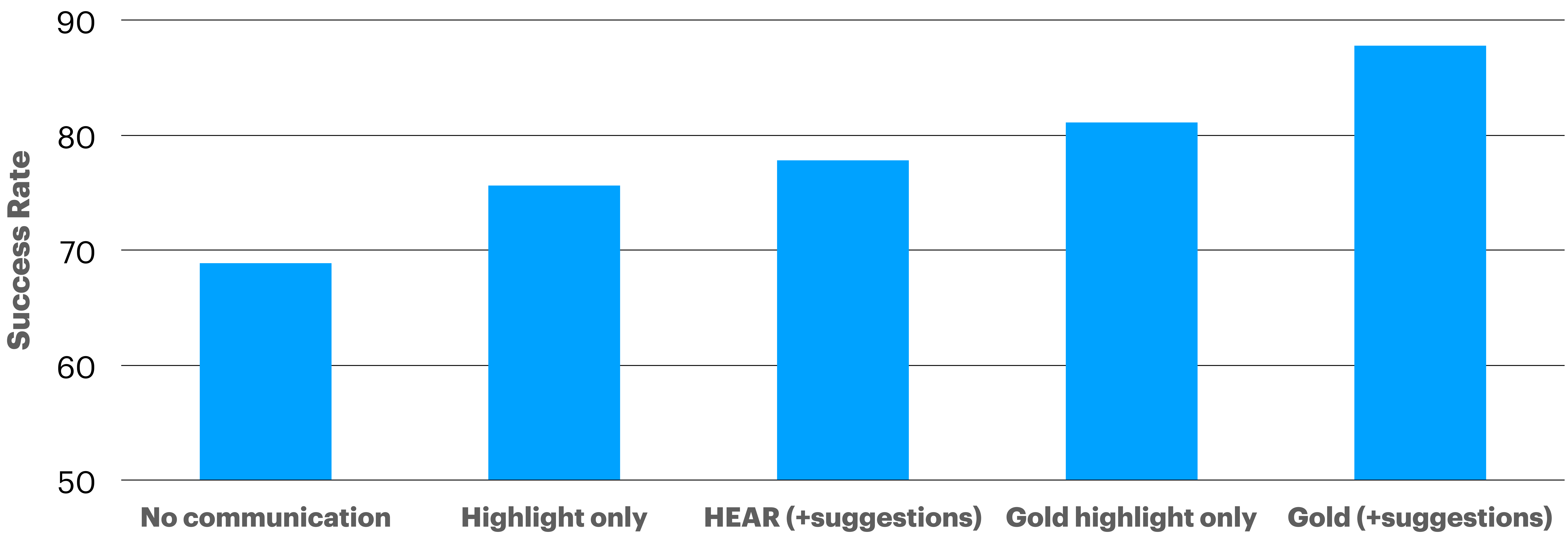
When to trust AI / use own judgement

  - Correction suggestions

How to fix AI's mistake



Highlights and suggestions **improve** human performance **8.9%-18.9%**



- *Takeaway*: better human-AI uncertainty communication ⇒ better human-AI collaboration

Takeaways

- Communicating **rich uncertainty** in LLM: better human-AI collaboration up to 19%

  - Modeling and training approach to generate uncertainty info

- Improving uncertainty communication:

  - A new direction for enhancing **human-AI collaboration**

Focus on **improving**

Truthful Maxim
- Generate more faithful explanations (EMNLP 25)
- Communicate uncertainty more effectively (EMNLP 24 & 23)

Informative Maxim
- Evaluate and improve cognitive capabilities for instruction generation (ACL 23)
- Culture pragmatics (ongoing)

Define, Evaluate, and Improve Task-Oriented Cognitive Capabilities for Instruction Generation Models

Lingjun Zhao*　　Khanh Nguyen*　　Hal Daumé III

Findings of ACL 2023

ICML Theory-of-Mind workshop Outstanding Paper

How to generate instructions for humans to easily follow?

- Why **important?**

  - Better human comprehension of AI's information

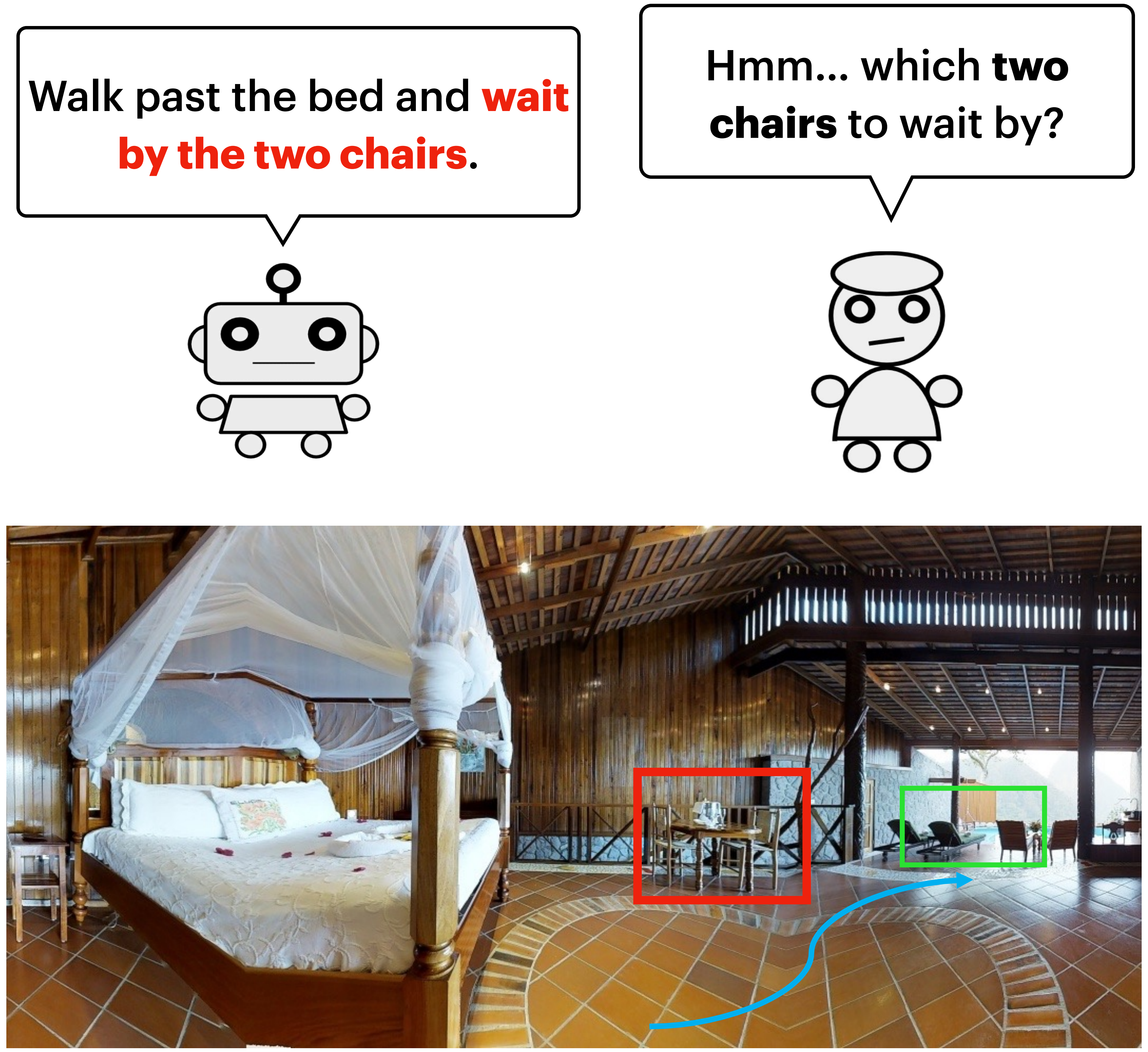
- Navigation task:

  - Measurable human interpretation of AI's communication

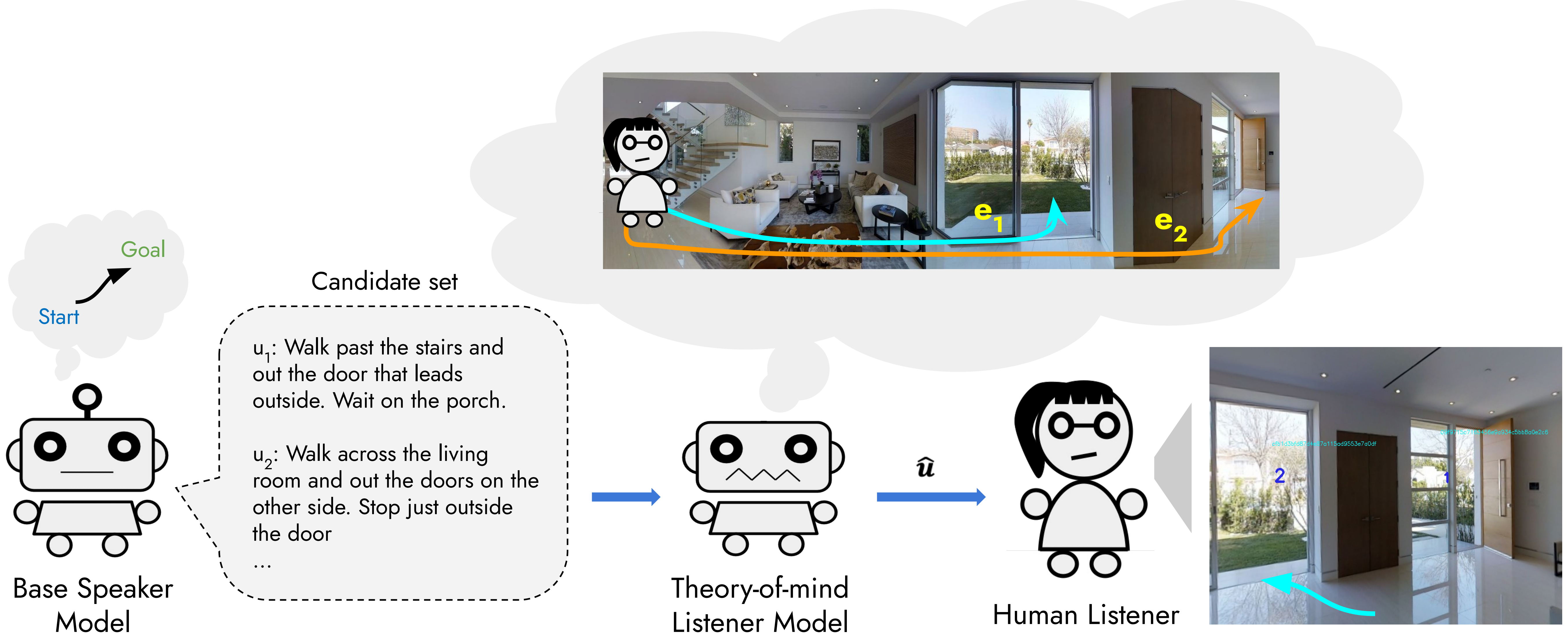
- Challenge: Model *fails* to communicate well with humans to achieve the goal

- **Task-oriented** speaker agent:

  - Generate instructions effectively help human accomplish a task



Bounded pragmatic speaker = Base speaker + Theory-of-mind Listener



- **Base Speaker:**

  Generates candidate instructions for a path

- **Theory-of-mind Listener**:

  Simulates how a human would follow each instruction

  (In practice: reinforcement learning agent for simulation)

- **Human Listener**:

  Follow the selected instruction to reach destination

Frank, Michael C., and Noah D. Goodman. "Predicting pragmatic reasoning in language games." Science, 2012.

Can we improve the speaker communication efficacy with better **theory-of-mind** listener?

Walk past the couch and stop in front of the TV.

Human evaluation: navigate using web interface

Using ensemble listeners for theory-of-mind
Improves up to 11.1% speaker communication efficacy

ToM listener L_{ToM}	Base speaker S_{base} Fine-tuned GPT-2	EncDec-LSTM	EncDec-Transformer
None	37.7 (▲ 0.0)	45.3 (▲ 0.0)	49.4 (▲ 0.0)
Single VLN-BERT (Majumdar et al., 2020)	38.9 (▲ 1.2)	39.8 (▼ 5.5)	46.2 (▼ 3.2)
Ensemble of 10 EnvDrop-CLIP (Shen et al., 2022)	37.8 (▲ 0.1)	53.1† (▲ 7.8)	57.3† (▲ 7.9)
Ensemble of 10 VLN⟳BERT (Hong et al., 2021)	43.4 (▲ 5.7)	56.4‡ (▲ 11.1)	54.2 (▲ 4.8)
Humans (skyline)	72.9‡ (▲ 35.2)	76.2‡ (▲ 30.9)	75.2‡ (▲ 25.8)

Human navigation performance (NDTW) using different speaker models, some augmented with theory-of-mind capabilities

- Shrink the gap with human-level speaker by 36%

Takeaways

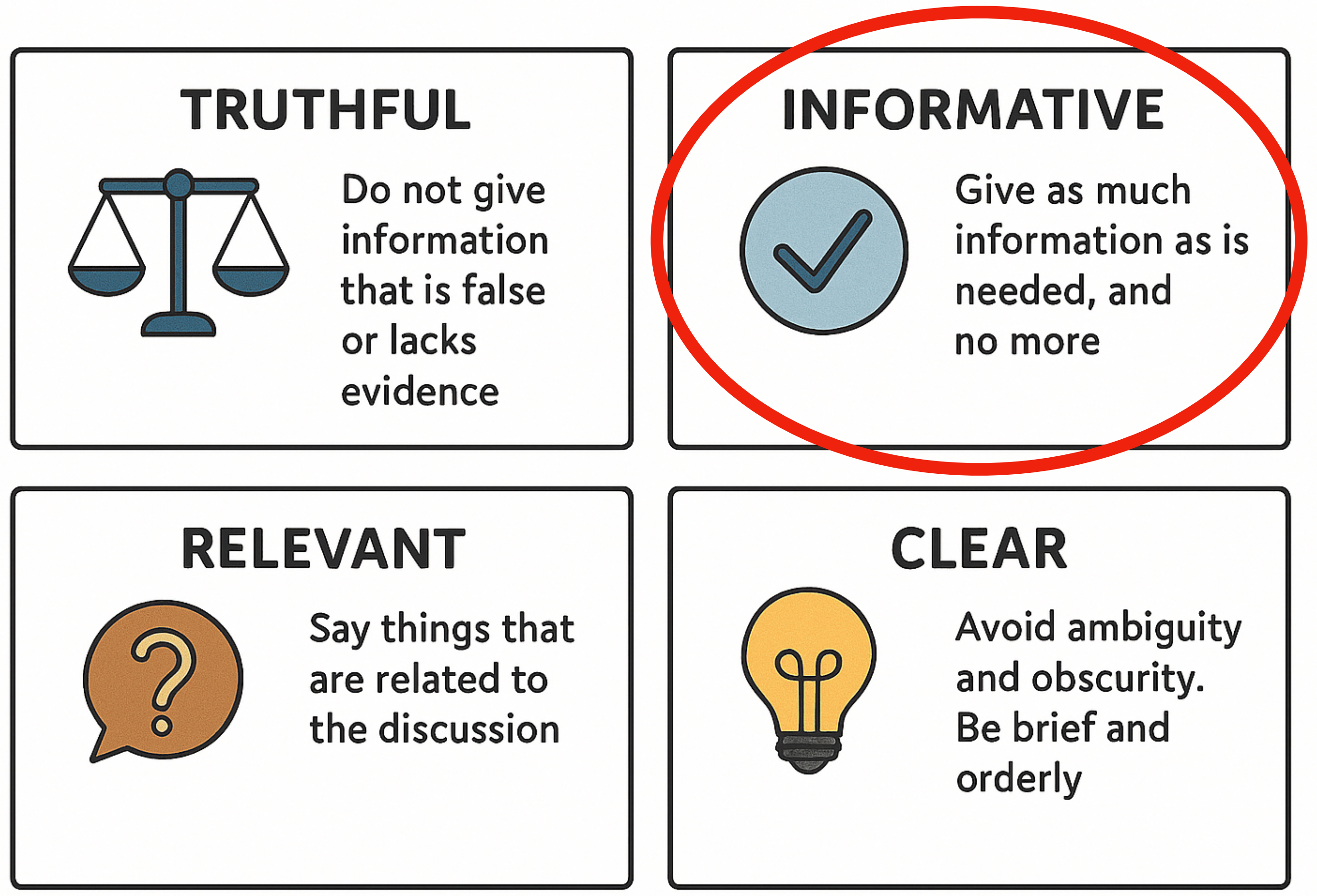
- Better **theory-of-mind** model improves task-oriented speaker agents

  - More informative AI communication ⇒ better human interpretation & performance

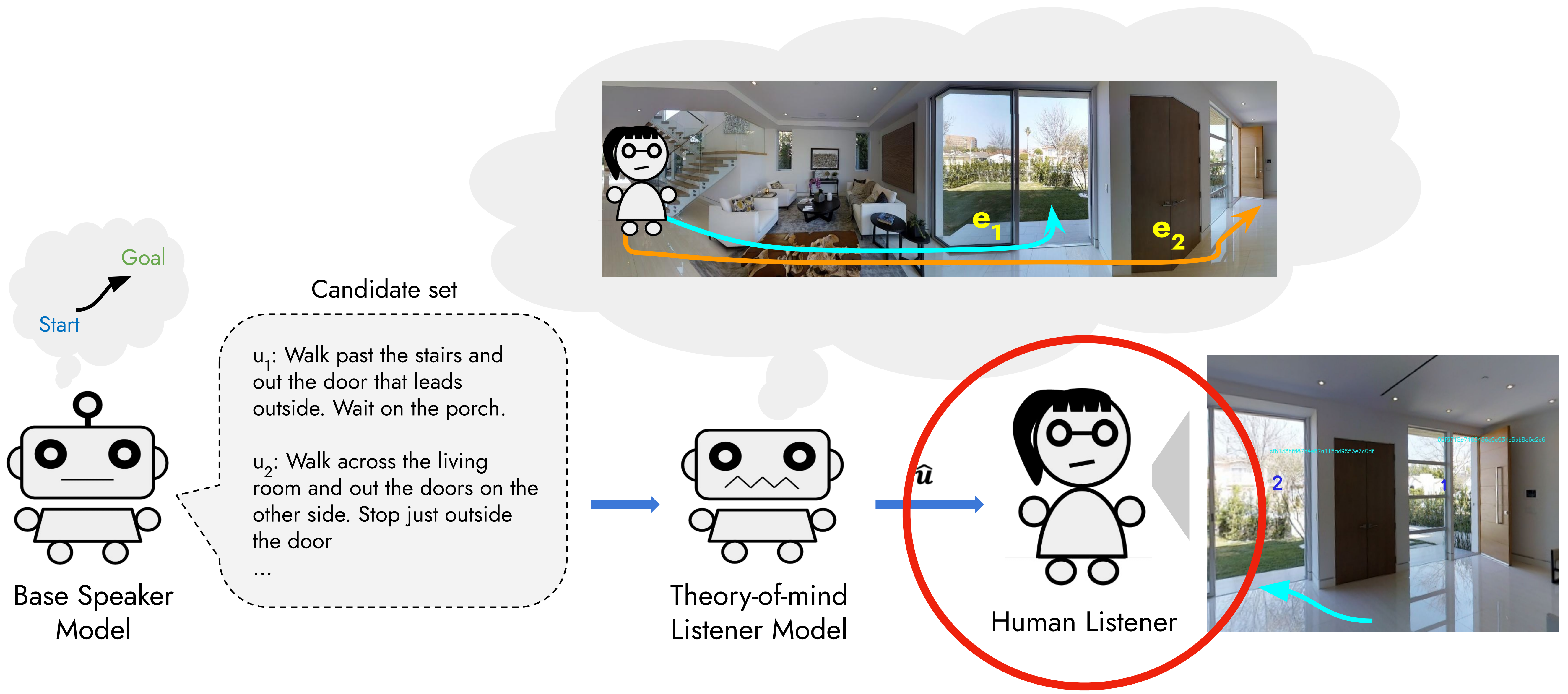
- Quantify the cognitive gaps between speaker agent and human speaker (in paper):

  - Search capability (candidate generation): good

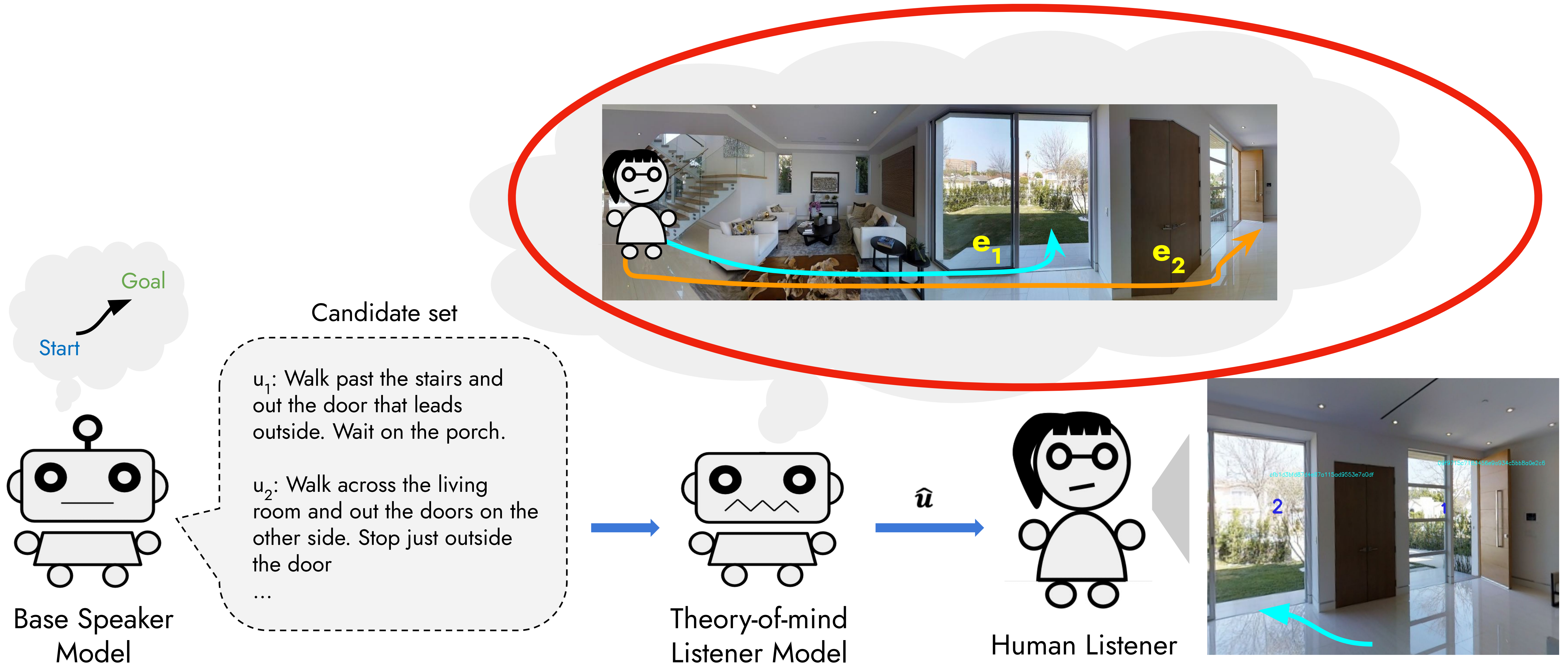
  - Theory-of-mind capability: still lacking



What if human listeners have **different** prior knowledge?



Real-world: What if we don't have a dataset to **evaluate** the theory-of-mind listener?



Focus on **improving**

Truthful Maxim

- Generate more faithful explanations (EMNLP 25)
- Communicate uncertainty more effectively (EMNLP 24 & 23)

Informative Maxim

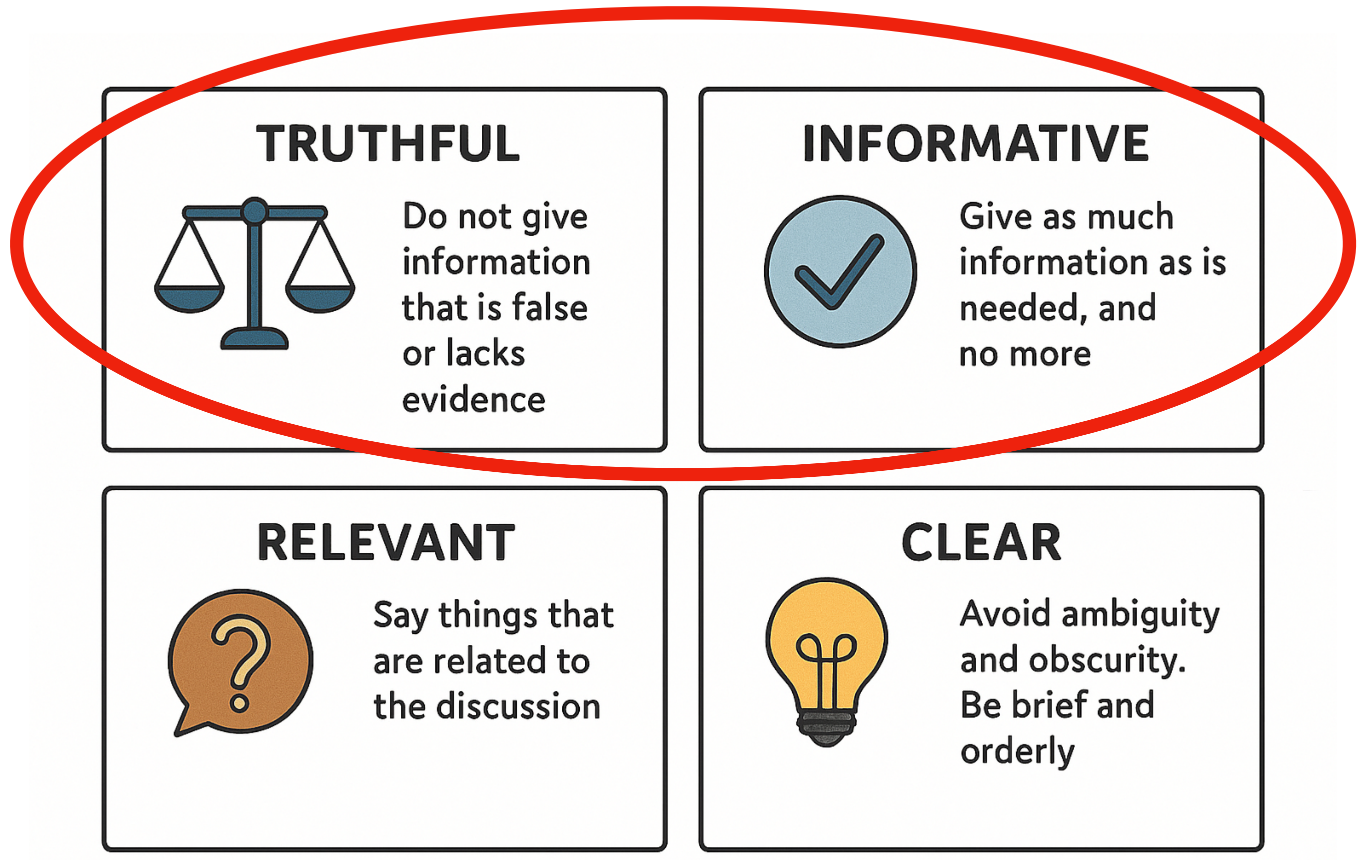
- Evaluate and improve cognitive capabilities for instruction generation (ACL 23)
- Culture pragmatics (ongoing)

Adapting Text Generation for Cultural Contexts

Lingjun Zhao Dayeon Ki Marine Carpuat Hal Daumé III

UNIVERSITY OF MARYLAND 1856

Summary



- We improve human-AI communication

  - by resembling human-human communication

Generate more faithful explanations

Communicate uncertainty more effectively

Support pragmatic communication

- Our methodology: circumvent annotation needs

Thank you! ☺️

Hal Daumé III

Khanh Nguyen

Dayeon (Zoey) Ki

Marine Carpuat

Acknowledgement: Some images in this talk were generated by Microsoft Copilot