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

Lingjun Zhao

Khanh Nguyen

Hal Daumé III

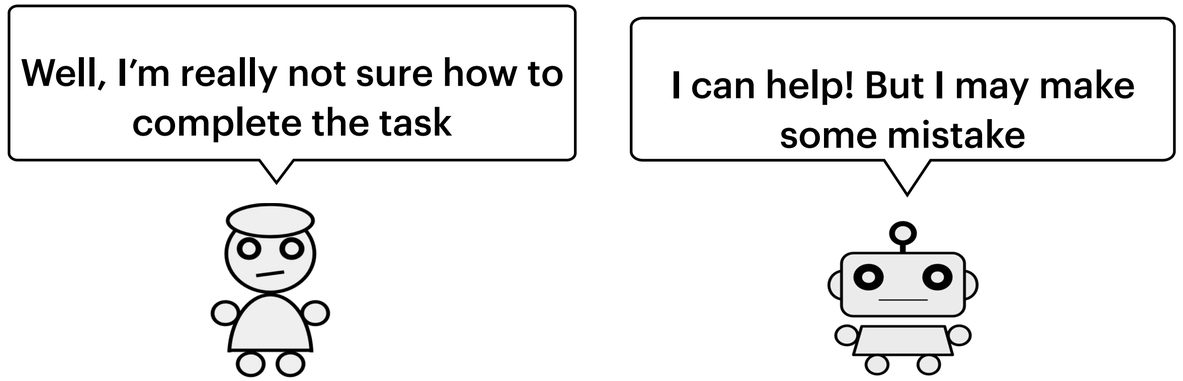
EMNLP 2024

Nov 12, 2024

Motivation: AI as Human's Collaborator for Decision Making

- **Problem**: Language models will **hallucinate** [1]
  - Can we enable them to assist human decision-making effectively?

- ~~AI as independent problem solver: higher benchmark scores~~
- **AI as human's collaborator**: improve human's decision making
- **How**: provide rich uncertainty information



[1] Calibrated language models must hallucinate. Kalai et al., 2024

Problem: How to better assist human navigation with fallible language models?

- **Task**: Human navigation in situated environment guided by a fallible language model

- **Question**: How can this model enable users to navigate successfully despite generating imperfect instructions?
  - Simply highlight potential hallucination is not enough:

    No actionable alternatives

Walk past the couch and stop in front of the TV

Grounded LLM: generates inaccurate description of stopping location

- **Solution**: communicate **uncertainties** to humans
  1. AI's potential mistakes ➡ **When to trust AI / use own judgement**
  2. Correction suggestions ➡ **How to fix AI's mistake**

Contributions

- A system that **detects and correct hallucinations** in visually grounded instructions
  - Modeling and training approach
  - Human evaluation: **29% reduction in human navigation error**

- Significance:
  - First study to show the **benefits of uncertainty communication** on a long-horizon decision-making task
  - Motivate improving uncertainty communication as a new direction for enhancing human-AI collaboration

Hallucination dEtection And Remedy (HEAR)

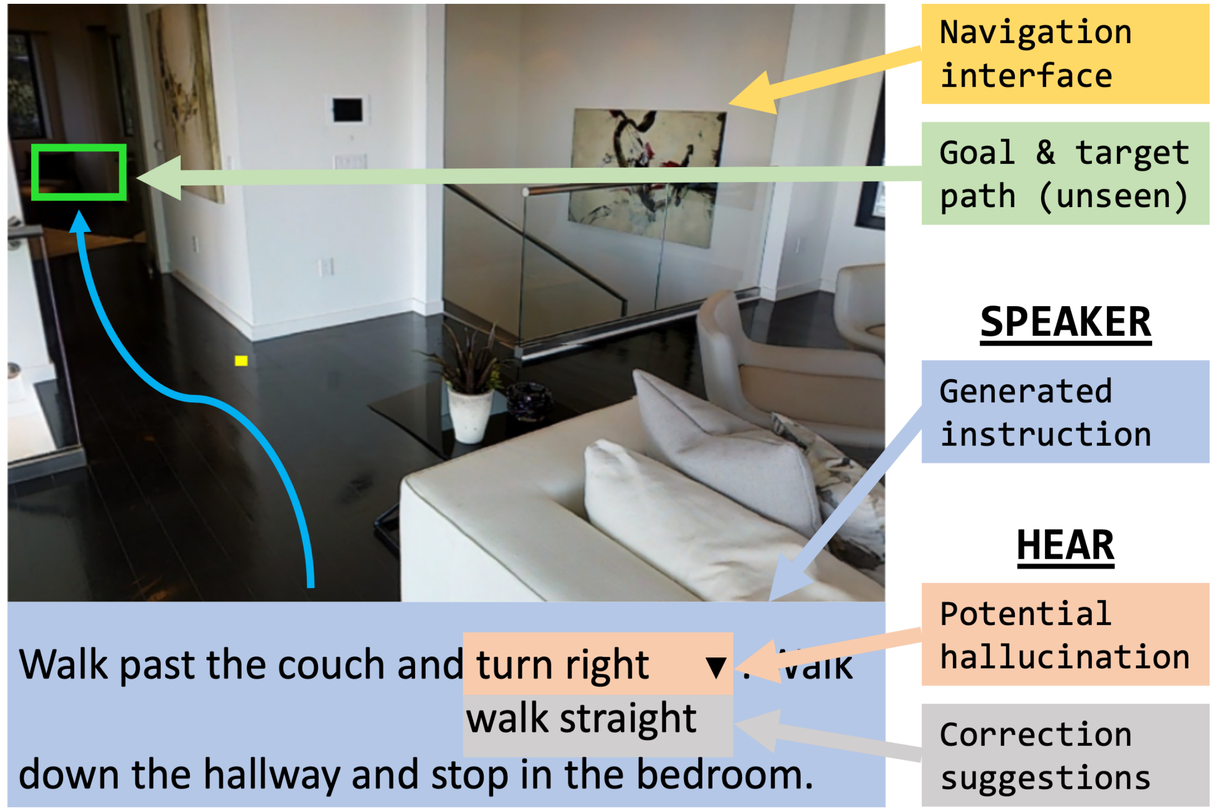
- Setting:
  - Users navigate in situated environments
  - Find destination using model-generated language instruction

- Present to humans:
  - **Potential hallucination spans**

    Hallucination detection model
  - **Correction suggestions**

    Suggestion ranking model

- No annotated data for training:
  - Create synthetic perturbations



HEAR: an example

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

Is it hallucination?

Hallucination detection model

Yes: highlight

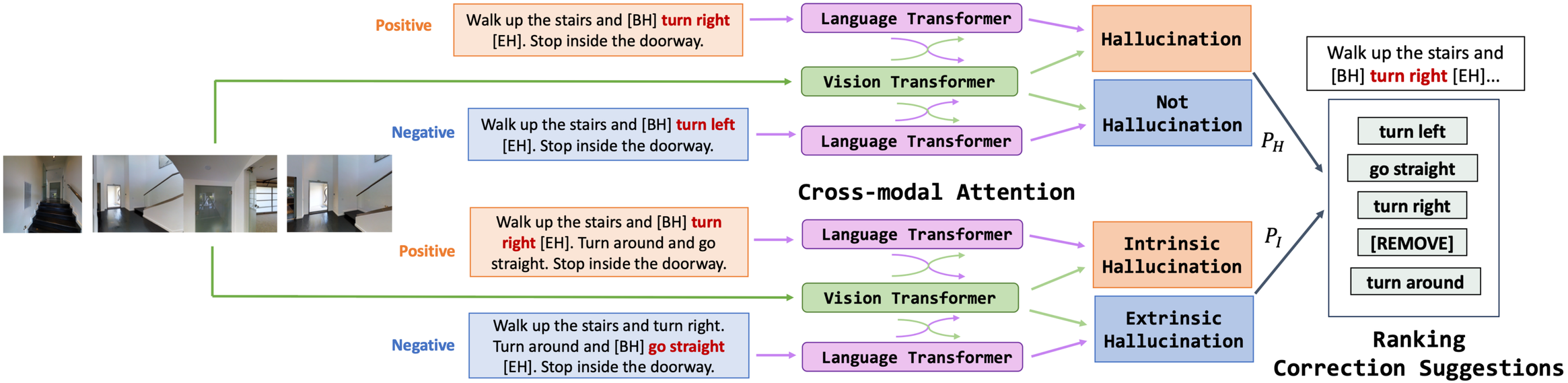
No: no highlight

Generate correction suggestions

Suggestion ranking model

Green box: ground truth destination

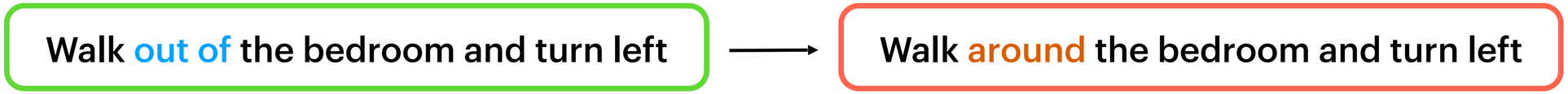
HEAR model design



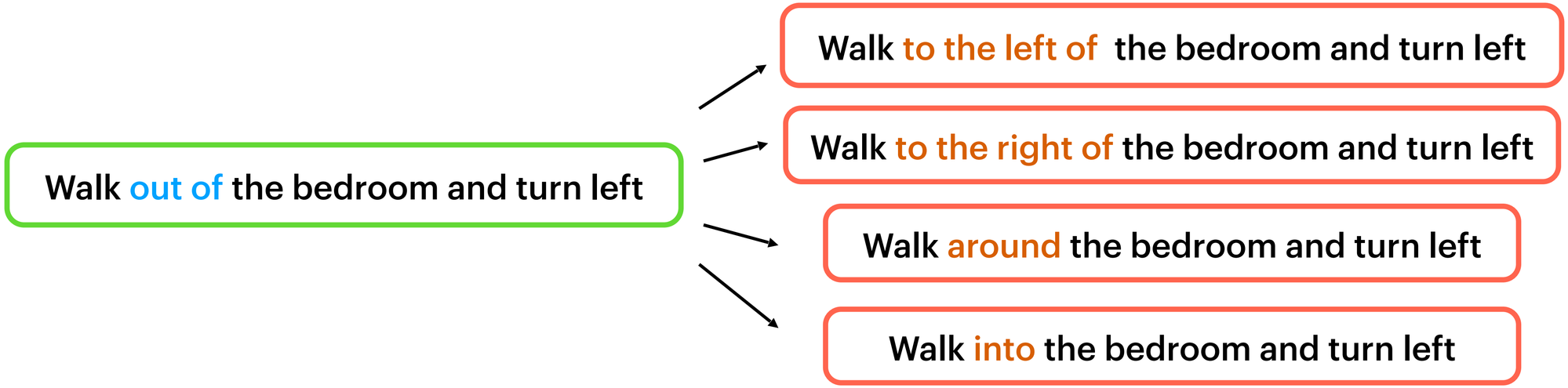
- Train two-stage model with **contrastive learning**:
  - Hallucination detection: on each span from the instruction
  - Hallucination type classification: ranking suggestions for each hallucination

Synthesize hallucinations and suggestion candidates

- For directions: GPT can generate meaningful perturbations

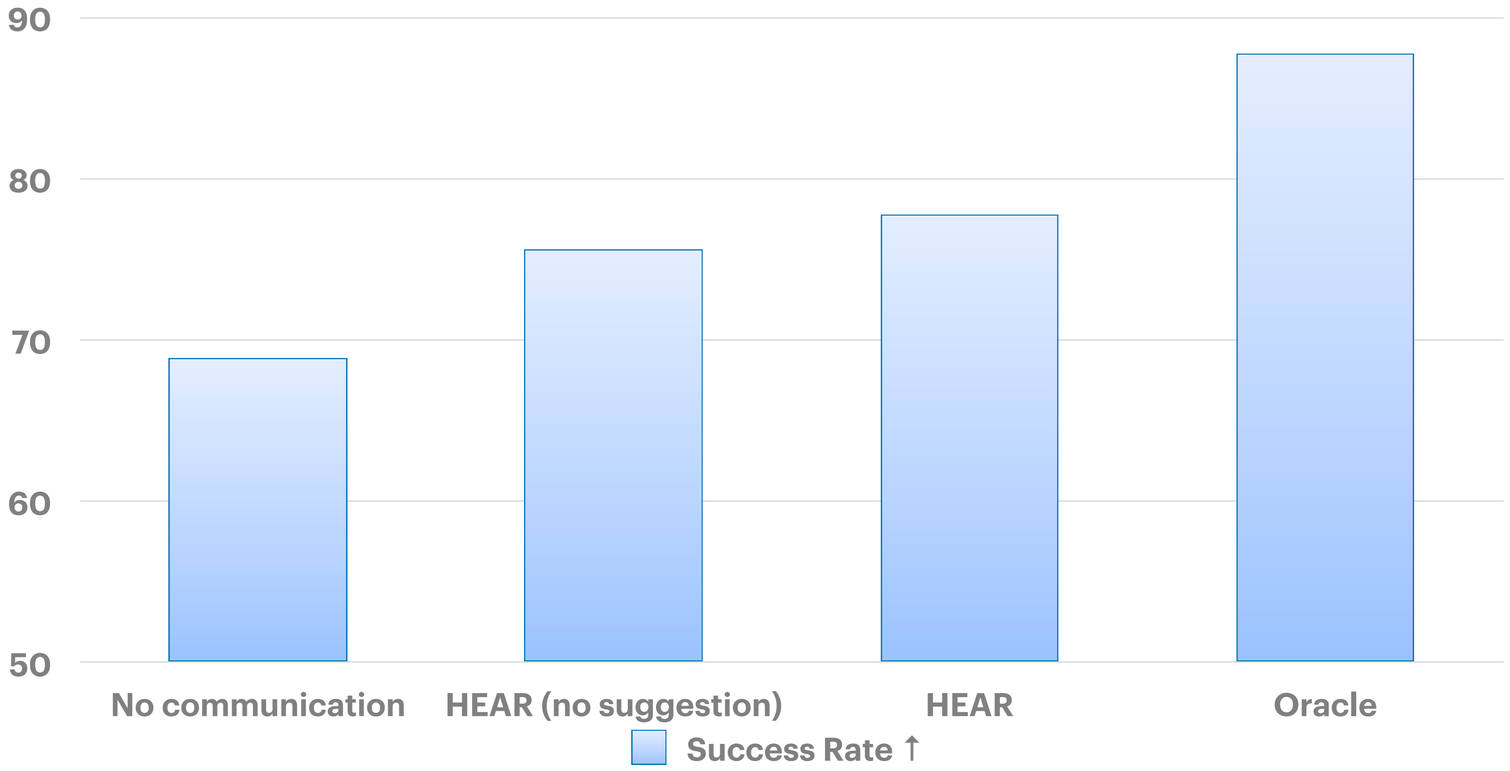


- And can generate meaningful suggestion candidates



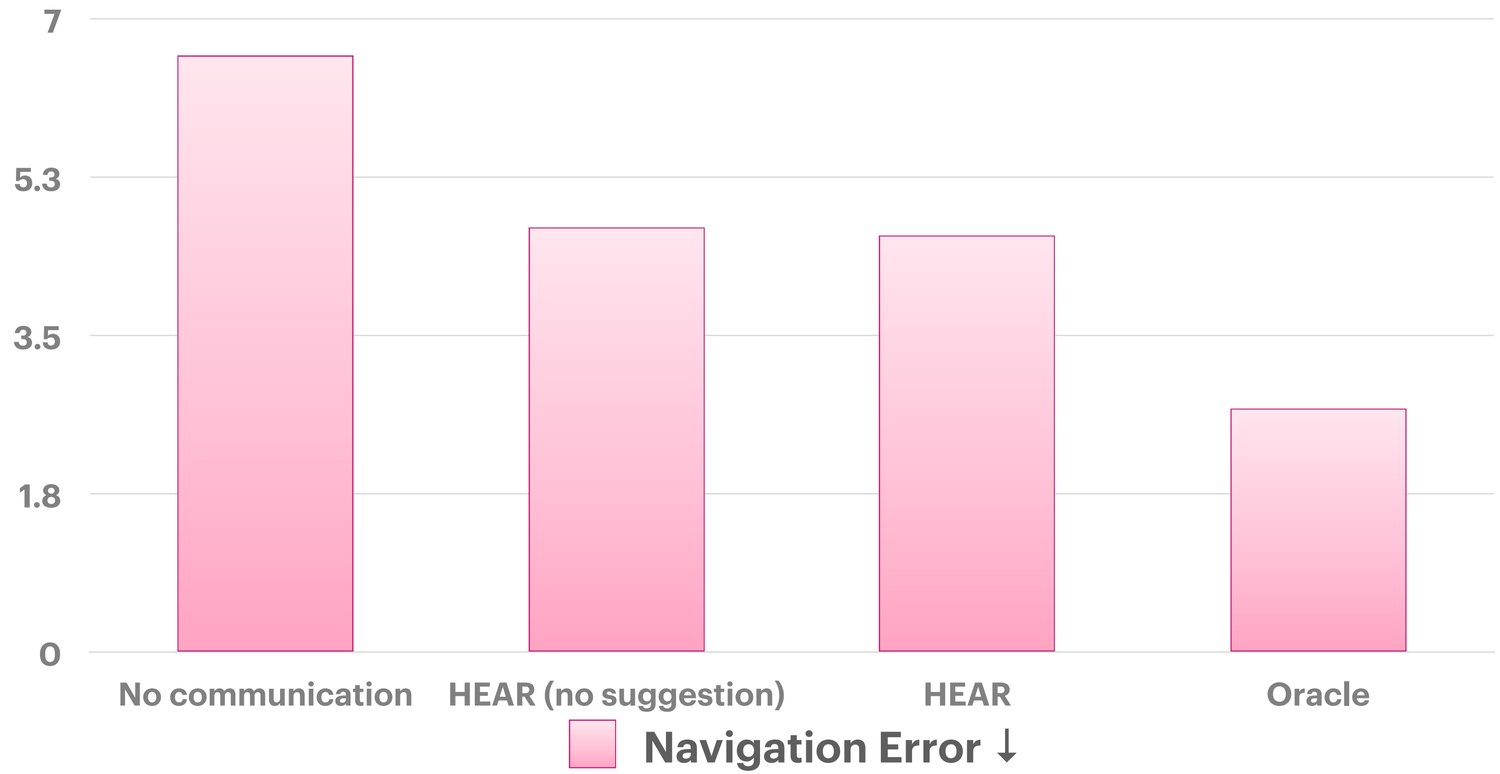
- Objects and rooms: rule-based procedures

Highlights and suggestions improve human navigation performance



Improve success rate: +13%

Highlights and suggestions improve human navigation performance



Reduce navigation error: -29%

Thank you!

- Our related paper:
  - **Hallucination detection for grounded instruction generation**. L Zhao, K Nguyen, H Daumé III. EMNLP Findings 2023.
- Try the **HEAR demo** from our project website (QR code below)!
- Dataset & codes released