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

**Lingjun Zhao** [*1] **Khanh Nguyen** [*2] **Hal Daumé III** [1 3]

## Abstract

Recent work studies the cognitive capabilities of language models through psychological tests designed for humans. While these studies are helpful for understanding the general capabilities of these models, there is no guarantee that a model possessing sufficient capabilities to pass those tests would actually *use* those capabilities in performing real-life tasks. In this work, we formulate *task-oriented* cognitive capabilities, which are human-like cognitive capabilities that language models leverage to perform tasks. These capabilities are (i) the ability to quickly generate good candidate utterances (the search capability) (ii) the ability to predict how a listener interprets those utterances and choose the most appropriate one (the pragmatic capability). We design an evaluation scheme for comparing these capabilities of a language model with those of a human. Applying this scheme to examine various models in a navigation instruction generation problem, we find that their pragmatic capability is severely lacking. This insight leads us to augment them with better models of the listener and obtain a significant boost of 11% in success rate in guiding real humans. Our work advocates for having a principled procedure for aligning language models with humans that involves (i) formulating task-oriented capabilities, (ii) devising a method to quantify their deficiency, and (iii) iteratively improving them.

## 1. Introduction

To communicate successfully with humans, language models must possess cognitive capabilities similar to those that facilitate human communication. Examining the cognitive capabilities of language models is notoriously challenging because the operations of these models are largely unintelligible to humans. Psychologists faced similar challenges when investigating human cognition, and have devised various behavioral tests to diagnose human cognitive capabilities (Premack & Woodruff, 1978; Wimmer & Perner, 1983; Baron-Cohen et al., 1985; Gopnik & Astington, 1988). Recent work (Sap et al., 2022; Kosinski, 2023; Ullman, 2023) applies these tests to evaluate large language models by inputting the tests to these models as prompts and verifying whether they behave like a normal human would.

While this approach is helpful for understanding the general limitations of language models, it has two potential drawbacks. First, it is applicable to only large language models that can comprehend human-written prompts, entangling linguistic capability with reasoning capability. Second, it shows that a language model can or cannot demonstrate certain mental skills, but does not imply that the model would employ those skills to perform a downstream task. For example, passing false-belief tests does not guarantee that a model will reason about the interpretation of the readers when generating summaries. In general, scoring high on psychological tests may not be sufficient to ensure language models would behave like humans in real-life scenarios.

In this work, we take a different approach to evaluating the cognitive capabilities of language models. We define and evaluate *task-oriented* cognitive capabilities, which are human-like capabilities that a model actually employs to perform the task it is designed for. Enhancing these capabilities thus warrants improved performance on the task. To identify these capabilities, we build on two lines of work from socio-cognitive science: Bayesian models of cooperative communication (Wang et al., 2020; Goodman & Frank, 2016; Shafto et al., 2014) and studies on how humans implement Bayesian reasoning (Sanborn & Chater, 2016; Sanborn et al., 2010; Vul et al., 2014; Mamassian et al., 2002). We propose a mathematical cognitive model called *bounded pragmatic speaker*, which can reasonably characterize the

[*] Equal contribution [1] Department of Computer Science, University of Maryland, College Park [2] Princeton University [3] Microsoft Research. Correspondence to: Lingjun Zhao <lzhao123@umd.edu>.

*Proceedings of the First Workshop on Theory of Mind in Communicating Agents*, Honolulu, Hawaii, USA. PMLR 202, 2023.

reasoning processes of both humans and language models. Casting humans and language models in the same way enables us to juxtapose their cognitive capabilities. We mathematically formulate two capabilities that a bounded pragmatic agent must possess in order to generate optimally pragmatic utterances. These conditions correspond to well-known cognitive capabilities of humans: (i) the ability to efficiently generate relevant utterances (the *search* capability) (Bloom & Fischler, 1980; Gold et al., 2000; Trosborg, 2010) and (ii) the ability to accurately simulate the listener's interpretations of their utterances (the *pragmatic* capability) (Premack & Woodruff, 1978; Gopnik & Astington, 1988; Tomasello, 2019; Call & Tomasello, 2011; Frank & Goodman, 2012). We design a simple procedure to quantitatively evaluate these capabilities of a language model. To evaluate each capability, we compute the task performance gap between the model and an *oracle* model, which is identical except that the evaluated capability of this model is at human level. Figure 1 illustrates our procedure, which theoretically can be applied to any language model.

We evaluate various language models on a navigation instruction generation problem (Anderson et al., 2018b), where a model generates English instructions to guide real humans in photo-realistic 3D environments. [1] Our evaluation reveals an interesting finding: all evaluated agents possess relatively efficient search capability but inadequate pragmatic capability. We improve the pragmatic capability of the evaluated models by enabling them to reason probabilistically about human listeners (Andreas & Klein, 2016; Fried et al., 2018a), employing state-of-the-art instruction-following agents (Magalhaes et al., 2019; Shen et al., 2022; Hong et al., 2021) as models of human listeners. We obtain significant improvement in success rate over the original agents, shrinking the gap with human performance on held-out data by 36%. Towards eliminating the remaining gap, we illustrate with empirical evidence a major challenge in developing better listener models. Specifically, when the instruction-following agents are employed as listener models for the instruction-generating agent, they are required to evaluate *AI-generated* instructions, which may be significantly different from human-generated instructions. Hence, a standard supervised-learning training scheme that only exposes these models to human-generated instructions would be inadequate for learning reliable models. We thus call for construction of novel datasets, algorithms, and evaluation methods for developing the pragmatic capability of language models.

[1] Our human-evaluation dataset and interface are publicly released at https://lingjunzhao.github.io/coop_instruction.html.

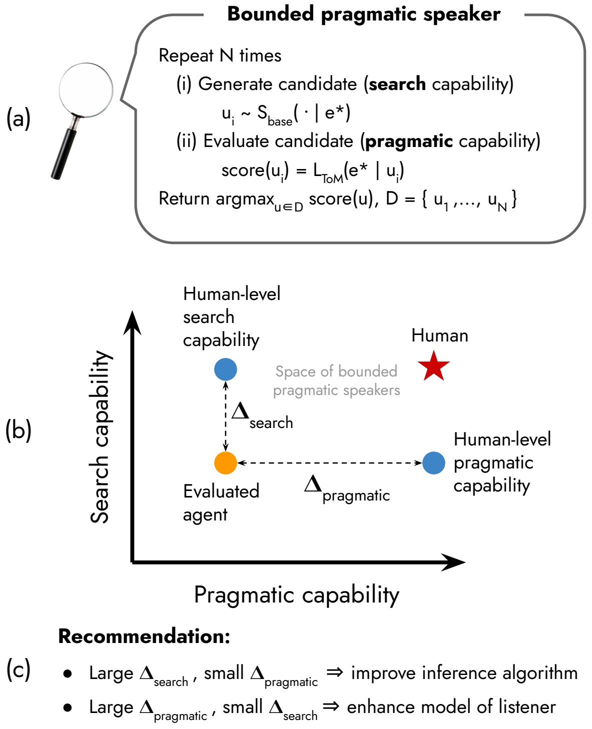

(a)
Bounded pragmatic speaker
Repeat N times
(i) Generate candidate (**search** capability)
$u_i \sim S_{base}(\cdot \mid e^*)$
(ii) Evaluate candidate (**pragmatic** capability)
$score(u_i) = L_{ToM}(e^* \mid u_i)$
Return $\text{argmax}_{u \in D}$ score(u), D = { $u_1, \ldots, u_N$ }
(b)
Search capability
Pragmatic capability
Human-level search capability
Human
Space of bounded pragmatic speakers
$\Delta_{search}$
Human-level pragmatic capability
Evaluated agent
$\Delta_{pragmatic}$
(c)
Recommendation:
• Large $\Delta_{search}$, small $\Delta_{pragmatic}$ ⇒ improve inference algorithm
• Large $\Delta_{pragmatic}$, small $\Delta_{search}$ ⇒ enhance model of listener

Figure 1: We propose a framework called *bounded pragmatic speaker* which can characterize pragmatic reasoning in both humans and language models (a). A bounded pragmatic speaker is composed of a base speaker $S_{\text{base}}$, representing prior knowledge that helps generate instructions efficiently, and a theory-of-mind (ToM) listener $L_{\text{ToM}}$, a hypothetical model of how the real listener interprets instructions. Viewing language models and humans through this unifying lens enables comparing their cognitive capabilities (b). To evaluate a capability of a model, we compare it with an oracle model which is identical except that the evaluated capability is at human level. The outcome of our evaluation can better inform the future direction for improving the model (c).

## 2. Related Work

**Navigation Instruction Generation.** Instruction generation has been commonly studied in navigation settings (Anderson et al., 1991; Byron et al., 2009; Koller et al., 2010; Striegnitz et al., 2011; Goeddel & Olson, 2012; Fried et al., 2018a;b). The Matterport3D simulator and the accompanying datasets (R2R (Anderson et al., 2018b), R4R (Jain et al., 2019), and RxR (Ku et al., 2020)) offer more challenging settings by combining photo-realistic scenes with long, verbally rich instructions. Recent work on evaluating instruction generation agents (Zhao et al., 2021) reveals the ineffectiveness of standard learning and modeling approaches to this problem. Wang et al. (2021) improve the accuracy and interpretability of instructions in the RxR setting. Kamath et al. (2023) leverage this model to synthesize additional data for training instruction-following agents. Our work aim to offer useful principles to further improve these models.

**Mathematical Models of Human Communication.** Different from communication within agents (Lazaridou et al., 2020; Roman Roman et al., 2020), human communication is a cooperative act (Grice, 1975; Scott-Phillips, 2014; Tomasello, 2019). Pragmatic communication in humans may involve different cognitive capabilities like basic understanding of language and social rules (Trosborg, 2010) and reasoning about the physical world (Bender & Koller, 2020) and human behavior (Enrici et al., 2019; Rubio-Fernandez, 2021). Our work describes similar capabilities but provides a formal mathematical description. Development of mathematical models of human communication have been greatly useful for understanding human behaviors (Ho et al., 2016; Sumers et al., 2022) and building communication agents (Andreas & Klein, 2016; Fried et al., 2018a;b; FAIR, 2022; Lin et al., 2022; Zhu et al., 2021; Bao et al., 2022). Wang et al. (2020) unify these models under an optimal-transport framework. The model we propose in this work is a generalized version capturing the essence of these models.

**Evaluating Cognitive Capabilities of Neural Networks.** Many benchmarks for evaluating the cognitive capabilities of AI-based agents have been created, focusing on theory-of-mind capabilities (Le et al., 2019; Roman Roman et al., 2020), grounding (Lachmy et al., 2022; Udagawa & Aizawa, 2019; Haber et al., 2019), or commonsense reasoning (Talmor et al., 2019; Levesque et al., 2012; Zellers et al., 2019; Sap et al., 2019). Large language models have demonstrated exceptional performance on following human instructions and solving complex reasoning tasks (Bubeck et al., 2023; Anil et al., 2023), raising the question of whether their cognitive capabilities are similar or as advanced as those of humans. Mahowald et al. (2023) advocate for separating formal competence (knowledge about linguistic rules and patterns) from their functional competence (knowledge about the world usage in the world) when assessing these models. Our bounded pragmatic speaker framework mathematically formalizes this description, allowing for quantitative evaluation of these competencies. Recent work (Sap et al., 2022; Kosinski, 2023; Ullman, 2023; Hu et al., 2023) examines cognitive capabilities of large language models through tests inspired by human psychological tests. The goal of these studies is to determine the limits of large language models, potentially calibrating the expectation on them. On the other hand, our focus is to devise a method that can be applied to language models of any size and benchmark cognitive capabilities that are relevant for accomplishing a specific task.

## 3. Problem Setting

We are concerned with instruction generation: learning a speaker agent $r$ that generates language instructions to guide a human listener $h$ to reach states in an environment.

**Human Listener.** We imagine a human listener $h$ acting in a partially observed environment with states $s$. The human does not have access to states but only observations $o^h$ and takes actions $a^h$. An *instruction* $\boldsymbol{u} \in \mathcal{U}$ is a language utterance consisting of words. A *trajectory* $\boldsymbol{e} = (s_1, o_1, a_1, \cdots, s_T, o_T, a_T)$ is an execution of an instruction. The human can follow instructions to generate trajectories in the environment. For example, in an indoor navigation setting, upon hearing "*go the kitchen and stop next to the oven*", a human walks to the specified location. We define $L_h(\boldsymbol{e} \mid \boldsymbol{u})$ as the probability that the human generates $\boldsymbol{e}$ upon hearing $\boldsymbol{u}$.

**Speaker Agent.** In each task, the speaker agent first imagines an *intended trajectory* $\boldsymbol{e}^\star = (s_1, o^r_1, a^r_1, \cdots, s_T, o^r_T, a^r_T)$, which specifies a path to get to an intended goal state $s_T$ from the human's current state $s_1$. Because the human's actions and perception may differ from those of the speaker, they may not be able to comprehend $\boldsymbol{e}^\star$ even if it is presented to them. Thus, the speaker needs to translate the trajectory into an instruction $\hat{\boldsymbol{u}}$ that the human can understand and follow. To do so, it implements a *language model* $S_r(\boldsymbol{u} \mid \boldsymbol{e})$, and an *inference algorithm* $\text{Gen}(S_r, \boldsymbol{e})$ to craft instructions based on the model (e.g., greedy or beam-search decoding). The speaker's objective is to generate instructions that maximize the expected chance of the listener reconstructing the intended trajectories

$$\arg\max_{S_r} \mathbb{E}_{\boldsymbol{e}^\star} \left[ L_h(\boldsymbol{e}^\star \mid \text{Gen}(S_r, \boldsymbol{e}^\star)) \right] \tag{1}$$

**Evaluation.** The speaker is evaluated using a dataset $\mathcal{D}_{\text{eval}}$ of held-out trajectories. For each trajectory $\boldsymbol{e}^\star_k \in \mathcal{D}_{\text{eval}}$, we generate an instruction $\hat{\boldsymbol{u}}_k = \textsc{Gen}(S_r, \boldsymbol{e}^\star_k)$. The instruction is then presented to a (real) human listener to follow, producing a trajectory $\boldsymbol{e}^{\text{h}}_k \sim L_h(\cdot \mid \hat{\boldsymbol{u}}_k)$. The performance of the speaker, denoted by $\rho(r)$, is the average similarity, $\Psi$, between the human-generated and the intended trajectories:

$$\rho(r) \triangleq \frac{1}{|\mathcal{D}_{\text{eval}}|} \sum_{\boldsymbol{e}^\star_k \in \mathcal{D}_{\text{eval}}} \Psi(\boldsymbol{e}^{\text{h}}_k, \boldsymbol{e}^\star_k) \tag{2}$$

We will specify the choices for the metric $\Psi$ in the experimental setup section (Section 6).

## 4. Task-Oriented Cognitive Capabilities

Human have evolved highly effective cognitive capabilities for communication. How can we endow a speaker agent with similar capabilities and quantify the degree of resemblance between its capabilities and those of a human?

We propose a mathematical framework that reasonably characterizes the human cognitive process for instruction generation (Section 4.1). We show that this model can also describe the operation of language models, which allows us to compare them with humans on specific cognitive capabilities. We identify two capabilities that are requisite for any agent implementing our framework to generate optimal instructions (Section 4.2), and introduce an evaluation scheme for collating these capabilities (Section 4.3).

### 4.1. A Mathematical Cognitive Model of Instruction Generation

To formulate how humans generate instructions, we build on mathematical models of cooperative communication (Wang et al., 2020; Goodman & Frank, 2016; Shafto et al., 2014). We consider a general version where a speaker agent constructs a *pragmatic speaker* model $S_{\text{prag}}(\boldsymbol{u} \mid \boldsymbol{e})$ based on two constituents: a *base speaker* model $S_{\text{base}}(\boldsymbol{u} \mid \boldsymbol{e})$ and a *theory-of-mind (ToM) listener* model $L_{\text{ToM}}(\boldsymbol{e} \mid \boldsymbol{u})$. The base speaker represents general knowledge of the agent about the world and the language it speaks. The ToM listener reflects situated knowledge about the listener, simulating how they would behave in the environment given an instruction. A pragmatic speaker aims to maximize the chance of the listener interpreting its instruction correctly, but it is still influenced by its general knowledge (e.g., social biases, language style). Formally, it is defined as:

$$S_{\text{prag}}(\boldsymbol{u} \mid \boldsymbol{e}) \propto L_{\text{ToM}}(\boldsymbol{e} \mid \boldsymbol{u}) S_{\text{base}}(\boldsymbol{u} \mid \boldsymbol{e}) \tag{3}$$

To convey an intended trajectory $\boldsymbol{e}^{\star}$, this speaker utters an instruction of maximum probability under its model:

$$\begin{aligned} \hat{\boldsymbol{u}}_{\text{prag}} &\triangleq \arg\max_{\boldsymbol{u} \in \mathcal{U}} S_{\text{prag}}(\boldsymbol{u} \mid \boldsymbol{e}^{\star}) \\ &= \arg\max_{\boldsymbol{u} \in \mathcal{U}} L_{\text{ToM}}(\boldsymbol{e}^{\star} \mid \boldsymbol{u}) S_{\text{base}}(\boldsymbol{u} \mid \boldsymbol{e}^{\star}) \end{aligned} \tag{4}$$

**Humans as bounded pragmatic speakers.** The pragmatic speaker model accounts for human behaviors highly accurately on problems where $\mathcal{U}$ is a small discrete space (Frank & Goodman, 2012). However, in problems like instruction generation where $\mathcal{U}$ is an unbounded set of linguistic expressions, it is unlikely that humans, which are known to be agents with bounded rationality (Simon, 1957), are able to compute the optimal utterance in Eq 4 exactly. A hypothesis, supported by empirical evidence, is that humans perform approximate inference via Monte-Carlo sampling (Sanborn & Chater, 2016; Sanborn et al., 2010; Vul et al., 2014; Mamassian et al., 2002). Applying this hypothesis to our setting, we derive a more practical model of how human generate instructions, in which they perform the search for the best utterance on only a subspace $\mathcal{U}_{\text{sub}}$ of $\mathcal{U}$ defined by a set of candidates sampled from $S_{\text{base}}$

$$\hat{\boldsymbol{u}}_{\text{bounded-prag}} \triangleq \arg\max_{\boldsymbol{u} \,\in\, \mathcal{U}_{\text{sub}} \,\subset\, \mathcal{U}} L_{\text{ToM}}(\boldsymbol{e}^{\star} \mid \boldsymbol{u}) \tag{5}$$

where $\mathcal{U}_{\text{sub}} = \{\boldsymbol{u}_i \sim S_{\text{base}}(\cdot \mid \boldsymbol{e}^{\star}) \mid 1 \leq i \leq N\}$. We call an agent that generates instructions according to Eq 5 a *bounded pragmatic speaker* (Figure 2). For such a speaker, instruction generation involves two tasks: candidate generation (performed by $S_{\text{base}}$) and candidate evaluation (performed by $L_{\text{ToM}}$). The former task ensures that the generation of an instruction is efficient, while the latter guarantees the generated instruction conveys the intended meaning to the human listener.

### 4.2. Formulating Task-Oriented Cognitive Capabilities

What cognitive capabilities enable humans to generate effective instructions? Viewing humans as bounded pragmatic speakers allows us to mathematically characterize those capabilities. Specifically, we require a bounded pragmatic speaker to be able to output the optimal utterance, i.e. satisfying

$$\hat{\boldsymbol{u}}_{\text{bounded-prag}} = \boldsymbol{u}^{\star} \triangleq \arg\max_{\boldsymbol{u}} L_h(\boldsymbol{e}^{\star} \mid \boldsymbol{u}) \tag{6}$$

where $L_h$ is the human listener.

For this equation to hold, the constituent models $S_{\text{base}}$ and $L_{\text{ToM}}$ of the bounded pragmatic speaker must meet certain conditions. The condition for $S_{\text{base}}$ is that the candidate set it generates must contain the optimal instruction, i.e. $\boldsymbol{u}^{\star} \in \mathcal{U}_{\text{sub}}$. This condition requires $S_{\text{base}}$ to be capable of quickly generating candidates and placing high probability on $\boldsymbol{u}^{\star}$ so the instruction can be found by sampling a few candidates. We refer to this capability as the *search capability*.

Meanwhile, the condition for $L_{\text{ToM}}$ is that it has to rank $\boldsymbol{u}^{\star}$ first among the candidates in $\mathcal{U}_{\text{sub}}$. Meeting this condition demands having the capability of constructing a mental emulation of the human listener and simulating the actions of the listener after receiving an instruction. We refer to this capability as the *pragmatic capability*.

The search and pragmatic capabilities are orthogonal and complementary. An agent with flawless pragmatic capability can evaluate the goodness of instructions given to it, but may not be able to efficiently generate good instructions by itself. In contrast, an agent with effective search capability can quickly bring to attention highly relevant utterances but cannot select the best one to output if its ToM model is erroneous.

### 4.3. Evaluating Task-Oriented Cognitive Capabilities

We have defined two cognitive capabilities that are requisite for humans in instruction generation. In this section, we will prove that a language model can also be cast as a bounded

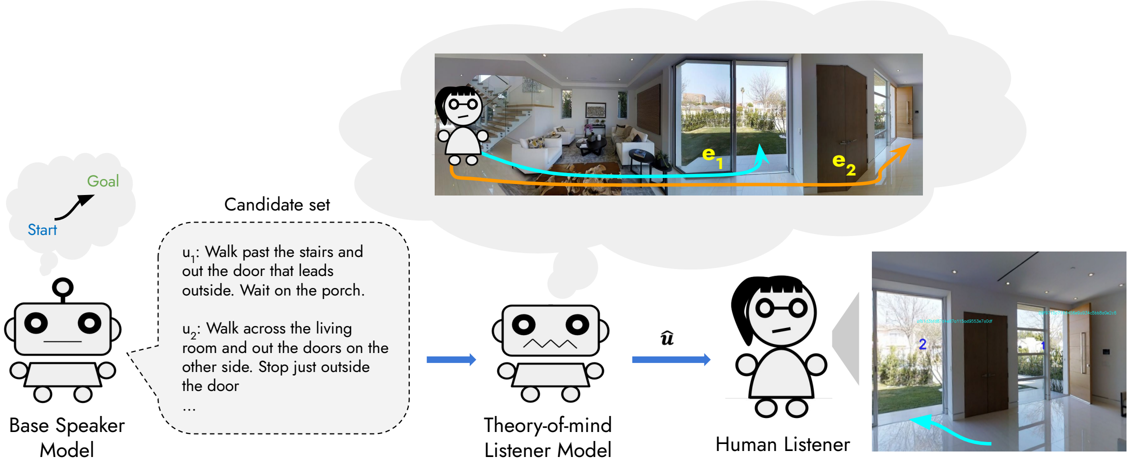

Goal
Start
Candidate set
u1: Walk past the stairs and out the door that leads outside. Wait on the porch.
u2: Walk across the living room and out the doors on the other side. Stop just outside the door
...
e1
e2
û
2
Base Speaker Model
Theory-of-mind Listener Model
Human Listener

Figure 2: The cognitive process of a bounded pragmatic speaker. In every task, the speaker first imagines a trajectory it wants to convey to the human listener. To reduce the search space, it then uses the *base speaker* to generate a small set of relevant candidate instructions. After that, it employs the *theory-of-mind listener* to simulate how the human listener would follow each instruction in the candidate set. The speaker finally elects the candidate instruction that causes the theory-of-mind listener to generate the trajectory most similar to the intended trajectory. The output instruction is finally sent to the human listener for a real execution in the environment.

pragmatic speaker. Hence, we can compare it with a human on the two cognitive capabilities.

**Language models as bounded pragmatic speakers.** We consider a speaker agent $r$ that learns a language model $S_r(\boldsymbol{u} \mid \boldsymbol{e})$ and runs an inference algorithm to compute an output $\hat{\boldsymbol{u}}_{\text{infer}} = \text{GEN}(S_r, \boldsymbol{e}^\star) \approx \arg\max_{\boldsymbol{u} \in \mathcal{U}} S_r(\boldsymbol{u} \mid \boldsymbol{e}^\star)$. Generative LSTM- or Transformer-based models that implement greedy or beam-search decoding are examples of this agent. We make the following assumption about the inference algorithm.[2]

**Assumption** (Better-than-sampling inference algorithm)**.** *We assume the inference algorithm is better at finding* $\arg\max_{\boldsymbol{u} \in \mathcal{U}} S_r(\boldsymbol{u} \mid \boldsymbol{e}^\star)$ *than drawing a small number of* $N$ *samples from* $S_r$*. Formally, let* $\gamma$ *be the probability of drawing* $\boldsymbol{e}^\star$ *and a set of* $N$ *instructions from* $S_r$ *such that* $S_r(\hat{\boldsymbol{u}}_{\mathit{infer}} \mid \boldsymbol{e}^\star) > \max_{\boldsymbol{u} \in \mathcal{U}_{\mathit{sub}}} S_r(\boldsymbol{u} \mid \boldsymbol{e}^\star)$*, where* $\hat{\boldsymbol{u}}_{\mathit{infer}} = \text{GEN}(S_r, \boldsymbol{e}^\star)$*. We assume that* $\gamma$ *is large for a small integer* $N > 0$*.*

If this assumption holds, then with high probability, the agent $r$ behaves identically to a bounded pragmatic speaker that computes its output as:

$$\hat{\boldsymbol{u}} \triangleq \arg\max_{\boldsymbol{u} \in \mathcal{U}^r_{\text{sub}}} S_r(\boldsymbol{u} \mid \boldsymbol{e}^\star) \tag{7}$$

$$\mathcal{U}^r_{\text{sub}} \triangleq \{\hat{\boldsymbol{u}}_{\text{infer}}\} \cup \{\boldsymbol{u}_i \sim S_r \mid 1 \leq i \leq N\} \tag{8}$$

This agent uses $S_r$ as both the base speaker $S_{\text{base}}$ and ToM listener $L_{\text{ToM}}$. Due to our assumption, on most inputs, the agent outputs $\hat{\boldsymbol{u}}_{\text{infer}}$, similar to the original agent. We employ this bounded pragmatic speaker as the proxy for the original agent in comparisons with humans, and also refer to it as $r$.

**Evaluation scheme.** To evaluate a cognitive capability (search or pragmatic) of a speaker $r$, we compute the performance gap between it and an oracle agent that is at human level on the evaluated capability, but is equally good as it is at the other capability. Specifically, we define $r^\star_{\text{search}}$ to be an oracle speaker that employs $S_r$ as the ToM model but is given a "gold" candidate set $\mathcal{U}^\star_{\text{cand}}$ that always contains a human-generated reference instruction $\boldsymbol{u}^\star$. It selects its output as follows

$$\boldsymbol{u}^\star_{\text{search}} \triangleq \arg\max_{\boldsymbol{u} \in \mathcal{U}^\star_{\text{cand}}} S_r(\boldsymbol{u} \mid \boldsymbol{e}^\star) \tag{9}$$

This agent has similar pragmatic capability as $r$ but human-level search capability. Next, we construct $r^\star_{\text{pragmatic}}$, an oracle that generates candidates using $S_r$ but employs a human $L_h$ to rank the candidates

$$\boldsymbol{u}^\star_{\text{pragmatic}} \triangleq \arg\max_{\boldsymbol{u} \in \mathcal{U}^r_{\text{sub}}} L_h(\boldsymbol{e}^\star \mid \boldsymbol{u}) \tag{10}$$

with $\mathcal{U}^r_{\text{sub}}$ from Eq 8. The search capability of $r^\star_{\text{pragmatic}}$ is as good as $r$ but its pragmatic capability is that of a human.

[2] We empirically verify that this assumption holds for the agents we evaluate with $N = 10$ and $\gamma^N$ ranging from 0.7 to 0.9. We estimate $\gamma^N$ by computing the fraction of evaluation examples where the agent's model ranks $\hat{\boldsymbol{u}}_{\text{infer}}$ above $N$ samples drawn from it.

We calculate the *prospective performance gain* (PPG) with respect to each capability as follows

$$\text{PPG}_{\text{search}}(r) \triangleq \rho(r^\star_{\text{search}}) - \rho(r) \tag{11}$$

$$\text{PPG}_{\text{pragmatic}}(r) \triangleq \rho(r^\star_{\text{pragmatic}}) - \rho(r) \tag{12}$$

where $\rho$ is the performance on held-out data (Eq 2 of §3). Each metric computes the potential improvement if the corresponding capability is upgraded to match with that of a human. Comparing the two metrics reveals which of the two capabilities of $r$ is currently more deficient and informs future development direction for the agent. For example, if $\text{PPG}_{\text{search}}(r)$ is large and $\text{PPG}_{\text{pragmatic}}(r)$ is small, it means that $r$ is scoring the candidate instructions highly accurately but it is bad at finding high-score instructions. In this case, developers may want to focus on devising a more effective inference algorithm to improve the search capability of $r$. On the other hand, if $r$ estimates poorly calibrated scores, signified by $\text{PPG}_{\text{pragmatic}}(r)$ being large, enhancing its inference algorithm is fruitless, but endowing it with a module that simulates the listener's behavior more accurately would boost its performance.

## 5. Improving Pragmatic Capability with Ensemble Instruction-Following Agents

In cases where our evaluation scheme indicates that the pragmatic capability of a language model is deficient, we improve it by installing a better ToM listener model. A common approach to learning this listener model is to use the same dataset used for learning the speaker model (Andreas & Klein, 2016; Fried et al., 2018a;b). We argue that this approach has a potential drawback. A ToM model learned in this way is only exposed to human-generated input instructions. At deployment time, it would likely experience a *covariate shift* because as a ToM model, the model is then asked to score instructions generated by a speaker model, not by humans. These instructions may be incorrect, ungrammatical, or may simply have a different style than human-generated instructions. This covariate shift would hamper the model's judgement. Our preliminary experiments (Appendix A.6) confirm that using a listener trained on only human-generated inputs as the ToM model hurts rather than improves the performance of various speakers.

We show that this problem can be alleviated by employing ToM models that have calibrated uncertainty on unseen instructions. We obtain calibrated models through ensembling (Lakshminarayanan et al., 2017): we train listener models $\hat{L}^{(k)}(\boldsymbol{e} \mid \boldsymbol{u}), k = 1 \dots K$, each on a random 90% subset of the training data with different random initial seeds.

We also leverage access to a simulation of the environment to construct better ToM models. Note that the probability that a ToM model $L_{\text{ToM}}$ assigns to an instruction can be seen as an expectation of a binary metric: $L_{\text{ToM}}(\boldsymbol{e}^\star \mid \boldsymbol{u}) = \mathbb{E}_{\boldsymbol{e} \sim L_{\text{ToM}}(\cdot \mid \boldsymbol{u})} [\mathbb{1}\{\boldsymbol{e} = \boldsymbol{e}^\star\}]$, which does not award credit if $\boldsymbol{e}$ overlaps only partially with $\boldsymbol{e}^\star$. We propose two augmentations: (i) replace the binary metric with a soft metric $\Psi(\boldsymbol{e}, \boldsymbol{e}^\star)$ that can measure partial similarity between trajectories and (ii) approximate the expectation by executing listeners $\hat{L}^{(k)}$ in the simulated environment to sample trajectories. Our final model selects its instruction as:

$$\hat{\boldsymbol{u}}_{\text{augment-ToM}} \triangleq \arg\max_{\boldsymbol{u} \in \mathcal{U}^r_{\text{sub}}} L_{\text{ToM}}(\boldsymbol{e}^\star \mid \boldsymbol{u}) \tag{13}$$

$$L_{\text{ToM}}(\boldsymbol{e}^\star \mid \boldsymbol{u}) \propto \frac{1}{KM} \sum_{k=1}^{K} \sum_{j=1}^{M} \Psi(\boldsymbol{e}_j(\hat{L}^{(k)}, \boldsymbol{u}), \boldsymbol{e}^\star)$$

$$\mathcal{U}^r_{\text{sub}} \triangleq \{\hat{\boldsymbol{u}}_{\text{infer}}\} \cup \{\boldsymbol{u}_i \sim S_r \mid 1 \leq i \leq N\}$$

where $\boldsymbol{e}(L, \boldsymbol{u})$ denotes a trajectory sampled from a listener model $L$ conditioned on an instruction $\boldsymbol{u}$, and $M$ is the number of trajectories we sample from each listener. Essentially, the score $L_{\text{ToM}}(\boldsymbol{e}^\star \mid \boldsymbol{u})$ of each candidate instruction is the average performance metric of $K$ listeners, each of which attempts to follow the instruction $M$ times.

## 6. Experimental Setup

**Environment and Dataset.** We employ Matterport3D (Anderson et al., 2018b), a photo-realistic simulator of the visual perception of a person walking in an indoor environment. At any location, an agent is provided with RGB images capturing the 360-degree panoramic view when looking from that location. We train and evaluate our models using the Room-to-Room (R2R) language-based navigation dataset. Each data point was collected by asking an English-speaking crowd-worker to write a verbal description of a path in an environment. The dataset is split into a training set (61 environments, 4,675 paths), a seen validation set (environments seen during training, 340 paths), and an unseen validation set (11 environments unseen during training, 783 paths). We train the models using the training set and perform model selection on the unseen validation set. Performance metrics are computed on the seen validation set.

**Speaker Models.** We evaluate three speaker architectures: (1) a decoder-only GPT-2 pre-trained on text (Radford et al., 2019); (2) an LSTM encoder-decoder (Shen et al., 2022); (3) a Transformer encoder-decoder (Vaswani et al., 2017). Parameters of the latter two models are randomly initialized. Details are in Appendix A.2.

**Human Evaluation.** We evaluate each speaker model on 75 paths in the seen validation data split. In the end, we have annotated 1,200 instructions generated by 16 different systems (humans, 3 speaker models, and their ablated and augmented versions). To evaluate a speaker model, we present its generated instructions to a human annotator and ask them to follow the instructions to navigate in Matterport3D environments. We adapt the PanGEA tool[3] to setup a web navigation interface and create a task on Amazon Mechanical Turk (MTurk). We recruit 213 human evaluators in total. More details about the setting are given in Appendix A.5.

**Performance Metrics.** The quality of a speaker is determined by the similarity between the intended trajectories and the actual trajectories that the human evaluators generate by following the speaker's instructions. We compute these similarity metrics: **Success rate (SR)** averages binary indicators of whether the final location of a human-generated trajectory is within three meters of the final location of the intended trajectory; **SPL** (Anderson et al., 2018a) weights the success indicator with the ratio between the intended traveling distance and the actual one; and **NDTW and SDTW** are metrics based on dynamic time-warping alignment (Magalhaes et al., 2019), capturing the similarity between two point sequences. NDTW computes only a sequence similarity score while SDTW weights the score with the success indicator.

## 7. Experiments

We investigate the following questions:

(a) *How well do the speakers perform on our problem?* We find that, despite implementing advanced architectures, these speakers perform poorly compared to human speakers.

(b) *What causes their performance deficiency?* Using our evaluation scheme, we identify that the speakers possess decent search capability but inadequate pragmatic capability.

(c) *Can we improve the speakers by equipping them with better ToM listeners?* We employ ensembles of state-of-the-art instruction-following agents as ToM listeners for the speakers, and obtain significant improvements.

(d) *What are the challenges in bridging the performance gap with human speakers?* We show that instruction-following agents trained with only human-generated instructions are not optimal for serving as ToM listener models.

**How well do the speakers perform on our problem?** As seen in Figure 3, there is a wide margin between the agent speakers and the human instructors. The best model speaker (EncDec-Transformer) lags behind the humans by 21.6 NDTW points. The encoder-decoder architecture with cross-attention of EncDec-Transformer outperforms the decoder-only self-attention architecture of GPT-2 (+11.7 NDTW), indicating that fusing the vision and language features too early in an architecture may be detrimental. On the other hand, EncDec-Transformer leads over EncDec-LSTM by 4.1 points NDTW, suggesting that the Transformer architecture is more effective than LSTM in this problem.

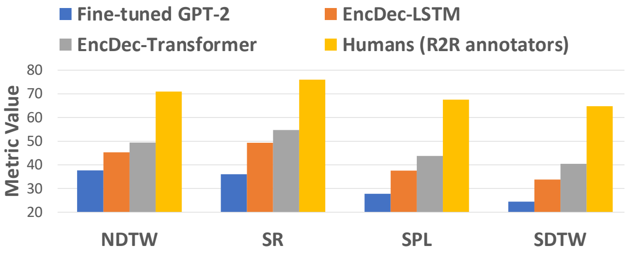


Figure 3: Performance of different speakers on held-out evaluation data, grouped by performance metrics (NDTW, SR, SPL, SDTW). Human speakers are annotators of the R2R dataset. There is a considerable gap between model and human speakers.

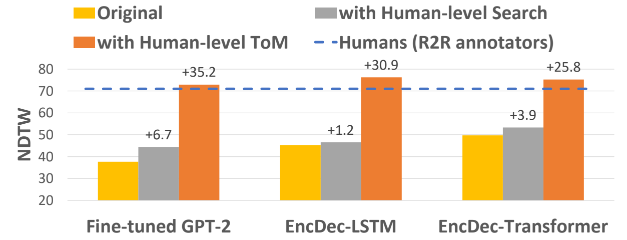


Figure 4: Performance (in NDTW) of the speakers and their human-augmented versions. Possessing human-level pragmatic capability improves performance of the speakers, showing that their original pragmatic capability is highly deficient compared to that of a human.

**What causes the speakers' deficiency?** Next, we investigate whether the lack of search or pragmatic capability is responsible for the deficiency of the speakers. The prospective performance gains presented in Figure 4 show that it is under-performed pragmatic capability that primarily causes the models to perform poorly. Specifically, while equipping the models with oracle search capability only improves their performance by 9.4% on average, granting them oracle pragmatic capability nearly doubles their performance metrics. In fact, the search capability of the models is already as good as that of the humans we employ, because the models with oracle pragmatic capability achieve even slightly higher NDTW scores than the human speakers.

**Can we improve the speakers by equipping them with better ToM models?** Following the procedure described in Section 5, we train state-of-the-art instruction-following agents to serve as ToM listener models for the speakers.

[3] https://github.com/google-research/pangea

| | | Base speaker $S_{\text{base}}$ | |
|---|---|---|---|
| ToM listener $L_{\text{ToM}}$ | 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) | $53.1^{\dagger}$ (▲ 7.8) | $57.3^{\dagger}$ (▲ 7.9) |
| Ensemble of 10 VLN↻BERT (Hong et al., 2021) | 43.4 (▲ 5.7) | $56.4^{\ddagger}$ (▲ 11.1) | 54.2 (▲ 4.8) |
| Humans (skyline) | $72.9^{\ddagger}$ (▲ 35.2) | $76.2^{\ddagger}$ (▲ 30.9) | $75.2^{\ddagger}$ (▲ 25.8) |

Table 1: Performance (in NDTW) of the speakers when equipped with different ToM models. Each base speaker generates 11 candidates (i.e. $N = 10$). Ensemble listeners significantly improve performance. $^{\ddagger}$ and $^{\dagger}$ indicate results that are significantly higher than those of "None" (row 1) with $p < 0.05$ and $p < 0.1$, respectively (two-related-sample t-test).

| | | Listener | |
|---|---|---|---|
| Instructions generated by | VLN-BERT | EnvDrop-CLIP | VLN↻BERT |
| Humans (R2R dataset) | 65.4 (▼ 0.0) | 47.2 (▼ 0.0) | 65.0 (▼ 0.0) |
| Fine-tuned GPT-2 | $43.1^{\ddagger}$ (▼ 22.3) | $31.6^{\ddagger}$ (▼ 15.6) | $39.9^{\ddagger}$ (▼ 25.1) |
| EncDec-LSTM | $50.0^{\ddagger}$ (▼ 15.4) | 43.7 (▼ 3.5) | $49.3^{\ddagger}$ (▼ 15.7) |
| EncDec-Transformer | $52.1^{\ddagger}$ (▼ 13.3) | 41.5 (▼ 5.5) | $51.9^{\ddagger}$ (▼ 13.1) |

Table 2: Agreement (in NDTW) of human and model listeners on instructions generated by different speakers. The level of agreement decreases substantially when shifting from human-generated to model-generated instructions. $^{\ddagger}$ indicate results that are significantly lower than the human skyline (row 1) with $p < 0.05$ (according to a two-related-sample t-test).

Performances of different combinations of speakers and listeners are given in Table 1. We see the largest improvement (+7.9 NDTW) over the best base speaker (EncDec-Transformer) by augmenting this speaker with an ensemble of 10 EnvDrop-CLIP listeners as the ToM model. We show qualitative examples where having a ToM listener enables the speaker to generate a more accurate instruction in Appendix A.7.

We observe that ensemble models consistently outperform single models. More results about the effectiveness of ensemble listeners compared to single listeners are given in Appendix A.6.

**What are the challenges in bridging the performance gap with human speakers?** Despite the promising improvements, there remains a large gap of 17.9 NDTW points between our best speaker and the human speakers. As suggested by Figure 4, this gap can be closed by developing accurate ToM models. We argue that optimal ToM models cannot be simply obtained by learning optimal instruction-following agents, because the latter is learned to execute *human-generated* instructions while the former is asked to rank *model-generated* instructions. To illustrate the difference, we measure the agreement between human and model listeners on instructions generated by different speakers. We define the agreement score between a human $L_h$ and a model $\hat{L}$ as

$$\begin{aligned}&\text{Agreement}(L_h, \hat{L})\\&= \text{Average}_{\boldsymbol{u} \in \mathcal{D}_{\text{eval}}} \left(\text{NDTW}(\boldsymbol{e}_h(\boldsymbol{u}), \hat{\boldsymbol{e}}(\boldsymbol{u}))\right) \qquad (14)\end{aligned}$$

where $\boldsymbol{e}_h(\boldsymbol{u})$ and $\hat{\boldsymbol{e}}(\boldsymbol{u})$ are the trajectories generated by $L_h$ and $\hat{L}$ given $\boldsymbol{u}$, respectively, and $\mathcal{D}_{\text{eval}}$ denotes the R2R seen validation set.

As seen in Table 2, the listener agents agree more with the humans on human-generated instructions than on model-generated ones. The results imply even an optimal instruction-following agent can fail to improve a base speaker in the presence of an input distribution mismatch. We thus advocate for developing ToM models that are robust or can adapt quickly against covariate shift, and for evaluating performance of these models on model-generated instructions.

## 8. Conclusion

This work introduces a framework for analyzing task-oriented cognitive capabilities of instruction-generation language models. We show that insights from the analysis are helpful in directing development on these models. Our results highlight the necessity of constructing better ToM models for improving these models. We argue that learning accurate ToM listener models is met with novel, distinct challenges. We hope that our findings will motivate the community to focus more on evaluating task-oriented cognitive capabilities and to create datasets, training methods, and evaluation procedures for enhancing the pragmatic capability of language models.

## References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. The hcrc map task corpus. *Language and speech*, 34(4):351–366, 1991.

Anderson, P., Chang, A., Chaplot, D. S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018a.

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and Van Den Hengel, A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674–3683, 2018b.

Andreas, J. and Klein, D. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1173–1182, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1125. URL https://aclanthology.org/D16-1125.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Bao, Y., Ghosh, S., and Chai, J. Learning to mediate disparities towards pragmatic communication. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2829–2842, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.202. URL https://aclanthology.org/2022.acl-long.202.

Baron-Cohen, S., Leslie, A. M., and Frith, U. Does the autistic child have a "theory of mind"? *Cognition*, 21(1): 37–46, 1985.

Bender, E. M. and Koller, A. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198, 2020.

Bloom, P. A. and Fischler, I. Completion norms for 329 sentence contexts. *Memory & cognition*, 8(6):631–642, 1980.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Byron, D., Koller, A., Striegnitz, K., Cassell, J., Dale, R., Moore, J., and Oberlander, J. Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pp. 165–173, Athens, Greece, March 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-0628.

Call, J. and Tomasello, M. Does the chimpanzee have a theory of mind? 30 years later. *Human Nature and Self Design*, pp. 83–96, 2011.

Enrici, I., Bara, B. G., and Adenzato, M. Theory of mind, pragmatics and the brain: Converging evidence for the role of intention processing as a core feature ofhuman communication. *Pragmatics & Cognition*, 26(1):5–38, 2019.

FAIR. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 2022.

Frank, M. C. and Goodman, N. D. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

Fried, D., Andreas, J., and Klein, D. Unified pragmatic models for generating and following instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1951–1963, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1177. URL https://aclanthology.org/N18-1177.

Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., and Darrell, T. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31, 2018b.

Goeddel, R. and Olson, E. Dart: A particle-based method for generating easy-to-follow directions. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1213–1219. IEEE, 2012.

Gold, J. M., Murray, R. F., Bennett, P. J., and Sekuler, A. B. Deriving behavioural receptive fields for visually completed contours. *Current Biology*, 10(11):663–666, 2000.

Goodman, N. D. and Frank, M. C. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

Gopnik, A. and Astington, J. W. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, pp. 26–37, 1988.

Grice, H. P. Logic and conversation. In *Speech acts*, pp. 41–58. Brill, 1975.

Haber, J., Baumgärtner, T., Takmaz, E., Gelderloos, L., Bruni, E., and Fernández, R. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1895–1910, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1184. URL https://aclanthology.org/P19-1184.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., and Austerweil, J. L. Showing versus doing: Teaching by demonstration. *Advances in neural information processing systems*, 29, 2016.

Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., and Gould, S. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1643–1653, 2021.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., and Gibson, E. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2212.06801.

Jain, V., Magalhaes, G., Ku, A., Vaswani, A., Ie, E., and Baldridge, J. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1862–1872, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1181. URL https://aclanthology.org/P19-1181.

Kamath, A., Anderson, P., Wang, S., Koh, J. Y., Ku, A., Waters, A., Yang, Y., Baldridge, J., and Parekh, Z. A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10813–10823, 2023.

Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., Moore, J. D., and Oberlander, J. Report on the second nlg challenge on generating instructions in virtual environments (give-2). In *Proceedings of the 6th international natural language generation conference*. The Association for Computer Linguistics, 2010.

Kosinski, M. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

Ku, A., Anderson, P., Patel, R., Ie, E., and Baldridge, J. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.356. URL https://aclanthology.org/2020.emnlp-main.356.

Lachmy, R., Pyatkin, V., Manevich, A., and Tsarfaty, R. Draw Me a Flower: Processing and Grounding Abstraction in Natural Language. *Transactions of the Association for Computational Linguistics*, 10:1341–1356, 11 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00522. URL https://doi.org/10.1162/tacl_a_00522.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lazaridou, A., Potapenko, A., and Tieleman, O. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7663–7674, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.685. URL https://aclanthology.org/2020.acl-main.685.

Le, M., Boureau, Y.-L., and Nickel, M. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5872–5877, 2019.

Levesque, H., Davis, E., and Morgenstern, L. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

Lin, J., Fried, D., Klein, D., and Dragan, A. Inferring rewards from language in context. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8546–8560, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.585. URL https://aclanthology.org/2022.acl-long.585.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Magalhaes, G. I., Jain, V., Ku, A., Ie, E., and Baldridge, J. General evaluation for instruction conditioned navigation using dynamic time warping. In *NeurIPS Visually Grounded Interaction and Language (ViGIL) Workshop*, 2019.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*, 2023.

Majumdar, A., Shrivastava, A., Lee, S., Anderson, P., Parikh, D., and Batra, D. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*, pp. 259–274. Springer, 2020.

Mamassian, P., Landy, M., and Maloney, L. T. Bayesian modelling of visual perception. *Probabilistic models of the brain*, 13:36, 2002.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526, 1978.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Roman Roman, H., Bisk, Y., Thomason, J., Celikyilmaz, A., and Gao, J. RMM: A recursive mental model for dialogue navigation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1732–1745, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.157. URL https://aclanthology.org/2020.findings-emnlp.157.

Rubio-Fernandez, P. Pragmatic markers: the missing link between language and theory of mind. *Synthese*, 199(1): 1125–1158, 2021.

Sanborn, A. N. and Chater, N. Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893, 2016.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, 117(4):1144, 2010.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454.

Sap, M., Le Bras, R., Fried, D., and Choi, Y. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.248.

Scott-Phillips, T. *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Bloomsbury Publishing, 2014.

Shafto, P., Goodman, N. D., and Griffiths, T. L. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71:55–89, 2014.

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. How much can clip benefit vision-and-language tasks? 2022.

Simon, H. A. Models of man; social and rational. *wiley*, 1957.

Striegnitz, K., Denis, A. A., Gargett, A., Garoufi, K., Koller, A., and Theune, M. Report on the second second challenge on generating instructions in virtual environments (give-2.5). In *13th European workshop on natural language generation*, 2011.

Sumers, T., Hawkins, R. D., Ho, M. K., Griffiths, T. L., and Hadfield-Menell, D. How to talk so ai will learn: Instructions, descriptions, and autonomy. In *Advances in Neural Information Processing Systems*, 2022.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Tan, H., Yu, L., and Bansal, M. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2610–2621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1268. URL https://aclanthology.org/N19-1268.

Tomasello, M. Becoming human. In *Becoming Human*. Harvard University Press, 2019.

Trosborg, A. *Pragmatics across languages and cultures*, volume 7. De Gruyter Mouton, 2010.

Udagawa, T. and Aizawa, A. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7120–7127, 2019.

Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vul, E., Goodman, N., Griffiths, T. L., and Tenenbaum, J. B. One and done? optimal decisions from very few samples. *Cognitive science*, 38(4):599–637, 2014.

Wang, P., Wang, J., Paranamana, P., and Shafto, P. A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, 33:17582–17593, 2020.

Wang, S., Montgomery, C., Orbay, J., Birodkar, V., Faust, A., Gur, I., Jaques, N., Waters, A., Baldridge, J., and Anderson, P. Less is more: Generating grounded navigation instructions from landmarks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15407–15417, 2021.

Wimmer, H. and Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Zhao, M., Anderson, P., Jain, V., Wang, S., Ku, A., Baldridge, J., and Ie, E. On the evaluation of vision-and-language navigation instructions. *European Chapter of the Association for Computational Linguistics*, pp. 1302–1316, 2021.

Zhu, H., Neubig, G., and Bisk, Y. Few-shot language coordination by modeling theory of mind. In *International Conference on Machine Learning*, pp. 12901–12911. PMLR, 2021.

| Hyperparam | GPT-2 | Transformer |
|---|---|---|
| Learning rate | $10^{-4}$ | $10^{-4}$ |
| Batch size | 4 | 32 |
| Optimizer | AdamW | AdamW |
| Num. of training iterations | $2 \times 10^5$ | $16 \times 10^4$ |
| Max. action steps | 15 | 35 |
| Max. instruction length | 100 | 80 |
| Image feature size | 2048 | 512 |
| Orientation feature size | 128 | 128 |
| Embedding dropout | 0.1 | 0.3 |
| Hidden size | 768 | 512 |
| Num. of hidden layers | 1 | 1 |
| Hidden-layer dropout rate | 0.0 | 0.6 |
| Num. of encoder layers | - | 2 |
| Num. of decoder layers | 12 | 2 |
| Transformer dropout rate | 0.1 | 0.3 |
| Beam size | 5 | 1 |

Table 3: Hyperparameters for training the GPT-2 EncDec-Transformer speakers.

# A. Appendices

## A.1. The Room-to-Room dataset

The R2R dataset (Anderson et al., 2018b) was originally created for training instruction-following agents. Each data point was collected by asking a crowd-worker to write a verbal description of a path in an environment. In the end, each path was annotated with three instructions. Each instruction contains 29 words on average. The dataset is split into a training set (61 environments, 4,675 paths), a seen validation set (340 paths) whose paths are sampled in the training environments, and an unseen validation set (11 environments unseen during training, 783 paths). We do not use the unseen test split because it does not provide ground-truth paths of the descriptions. We use the dataset consistent to their MIT License.

## A.2. Implementation of Speaker Models

We train the speakers with a standard maximum-likelihood objective using the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of $10^{-4}$.

The speaker models take a sequence of visual observations and actions from the trajectory $\boldsymbol{e}^\star$ as input and output a text instruction $\boldsymbol{u}$. The model is trained to estimate conditional probability $S_\theta(\boldsymbol{u}|\boldsymbol{e}^\star)$. We use grid search to select the model and training hyperparameters, and the best-found values are listed in Table 3.

**Input.** The input trajectory $\boldsymbol{e}^\star$ is a sequence of panoramic views and actions. Each panoramic view at time step $t$ is represented by 36 vectors $\{o_{t,i}\}_{i=1}^{36}$, each of which is a visual feature vector extracted from a pre-trained vision model concatenated with orientation features describing the agent's current gaze direction. The image features of the GPT-2 model are extracted from a ResNet-152 model (He et al., 2016), whereas those of the encoder-decoder models are from a CLIP model (Radford et al., 2021). Each ground truth action $a_t^\star$, which moves the agent to an adjacent location, is represented by image features from the gaze direction of the agent when looking towards that adjacent location, and orientation features capturing the direction of the adjacent location relative to the agent's current gaze direction.

**Output.** The output of a speaker model is a language instruction describing the input trajectory. At test time, the GPT-2 model employs beam search, and the encoder-decoder models generate instructions via greedy decoding (Shen et al., 2022).

**Training Objective.** We train the speakers with maximum-likelihood objective:

$$\max_\theta \sum_{(\boldsymbol{u}^\star,\boldsymbol{e}^\star)\in\mathcal{D}_{\text{train}}} \sum_{t=1}^{|\boldsymbol{u}^\star|} \log S_\theta(\boldsymbol{u}_t^\star \mid \boldsymbol{e}^\star, \boldsymbol{u}_{<t}^\star) \tag{15}$$

where $\theta$ is the speaker model parameters, $\boldsymbol{u}_t^\star$ is $t$-th word of the ground-truth instruction, and $\boldsymbol{u}_{<t}^\star$ is the first $t-1$ words of the instruction.

We select the best model based on the unseen-validation BLEU score (Papineni et al., 2002) of the model-generated instructions with the respect to the ground-truth instructions.

**Tools.** We use SacreBLEU 2.2.1 to compute BLEU scores. For preprocessing and implementing the speaker models, we use Pytorch 1.7.1, NLTK 3.6.7, SentencePiece 0.1.97, and Huggingface Transformers 4.5.1.

**Computation.** The GPT-2 model has 124.4 million parameters, and was trained for 24 hours on single NVIDIA GEFORCE RTX 2080 Ti. The EncDec-LSTM model has 7.5 million parameters, taking 24 hours to train on single NVIDIA RTX A6000. The EncDec-Transformer model has 56.6 million parameters, trained on single NVIDIA RTX A6000 for 48 hours.

## A.3. Fine-tuning GPT-2 Speaker Model

To represent the trajectory features as a sequence of feature vectors to feed into the GPT-2 model, we first average the view features $\bar{o}_t$ for each time step:

| | Performance Metrics | | | | | |
|---|---|---|---|---|---|---|
| Speaker | SR ↑ | SPL ↑ | NDTW ↑ | SDTW ↑ | Path Len ↓ | Interpretability ↑ |
| *Without ToM listener* | | | | | | |
| Finetuned GPT-2 | 36.0 | 27.8 | 37.7 | 24.5 | 20.9 | 2.9 |
| EncDec-LSTM | 49.3 | 37.6 | 45.3 | 33.8 | 17.4 | 3.3 |
| EncDec-Transformer | 54.7 | 43.8 | 49.4 | 40.4 | 15.8 | 3.4 |
| *With 10 VLN↻BERT as ToM listener* | | | | | | |
| Finetuned GPT-2 | 46.7 | 30.9 | 43.4 | 28.1 | 21.2 | 3.0 |
| EncDec-LSTM | 54.7 | 46.0 | 56.4 | 41.9 | 14.0 | 3.1 |
| EncDec-Transformer | 52.0 | 44.0 | 54.2 | 41.6 | 17.7 | 3.2 |
| Humans (R2R dataset) | 76.0 | 67.6 | 71.0 | 64.8 | 14.2 | 3.6 |

Table 4: Humans evaluation results on instructions generated by the speaker models. The similarity metrics are defined in §6. *Path Len* measures the average length of the generated trajectories. *Interpretability* indicates how easy or difficult to follow the instructions according to human evaluators (without knowing the ground-truth trajectory).

$$\bar{o}_t = \frac{1}{36}\sum_{i=1}^{36} o_{t,i} \tag{16}$$

We compute the input features $\boldsymbol{e}_t^\star$ by concatenating the panoramic view features and ground truth action features:

$$\boldsymbol{e}_t^\star = [\bar{o}_t; a_t^\star] \tag{17}$$

The sequence of feature vectors $\boldsymbol{e}^\star$ representing a trajectory is calculated as follows

$$\boldsymbol{e}^\star = [\tanh(\boldsymbol{e}_1^\star W); \cdots; \tanh(\boldsymbol{e}_T^\star W)] \tag{18}$$

where $W$ is parameters of a linear layer.

For the instruction $\boldsymbol{u}^\star$, we perform an embedding look-up of its words. Then, we first prompt the model with $\boldsymbol{e}^\star$ and then train it to generate $\boldsymbol{u}^\star$ as a suffix.

### A.4. Training Encoder-Decoder Speaker Models

Our EncDec-LSTM model follows the implementation of the speaker in Shen et al. (2022). We implement the EncDec-Transformer model by replacing the LSTM layers of the speaker model described in Tan et al. (2019) with Transformer layers (Vaswani et al., 2017).

### A.5. Human Evaluation Interface and Data Collection

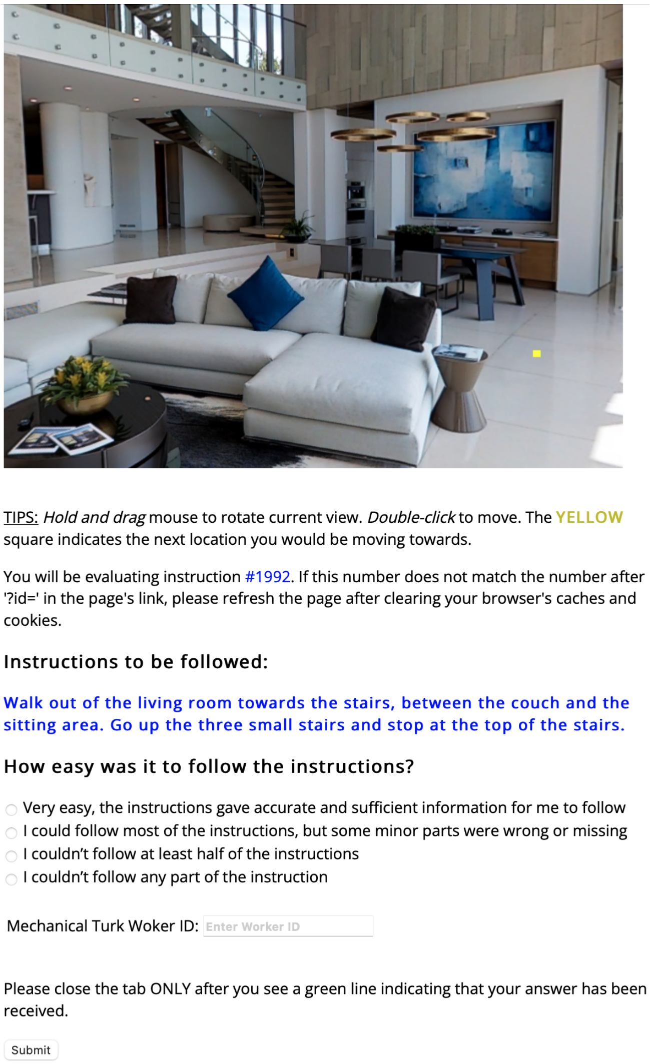
We pay the evaluator $5.20 per task which takes about 25 minutes, and the payment is decided by state minimum wage. For each task, we ask the evaluator to follow six instruction-following sessions. One of the six sessions, which appears in all tasks, is a quality-control test featuring an easy-to-follow human-written instruction. We only approve an evaluator if they navigate successfully to the goal destination in this test. Following Zhao et al. (2021), we instruct the judges to not explore the environments unnecessarily and not wander back and forth unless they are lost. We record the trajectories created by the human and use them to compute the performance metrics.

Figure 5 shows the interface for our human evaluation to collect annotations, which we adapted from the PanGEA tool[4] consistent with their Apache License v2.0. After a human evaluator finishes following an instruction, we recorded the path they generate and compute similarity metrics with respect to the ground-truth path. After the instruction-following sessions, we ask each evaluator to assess the interpretability of the instructions by asking them how easy (or difficult) it was for them to follow the instruction. We provide four rating levels ranging from "*1: I couldn't follow any part of the instruction*" to "*4: very easy, the instructions gave accurate and sufficient information for me to follow*". The answer of the evaluators is converted to a score between one and four.

Table 4 shows the human evaluation results of the three speaker models we evaluated.

For the human evaluation survey, participants will be restricted to those fluent in English. There are no other restrictions for this study. Participants must be at least 18 years old. Before completing the survey, participants will be shown information about the task requirement: *You are in a*

[4] https://github.com/google-research/pangea



TIPS: *Hold and drag* mouse to rotate current view. *Double-click* to move. The YELLOW square indicates the next location you would be moving towards.

You will be evaluating instruction #1992. If this number does not match the number after '?id=' in the page's link, please refresh the page after clearing your browser's caches and cookies.

**Instructions to be followed:**

Walk out of the living room towards the stairs, between the couch and the sitting area. Go up the three small stairs and stop at the top of the stairs.

**How easy was it to follow the instructions?**

○ Very easy, the instructions gave accurate and sufficient information for me to follow
○ I could follow most of the instructions, but some minor parts were wrong or missing
○ I couldn't follow at least half of the instructions
○ I couldn't follow any part of the instruction

Mechanical Turk Woker ID: Enter Worker ID

Please close the tab ONLY after you see a green line indicating that your answer has been received.

Figure 5: Human evaluation interface.

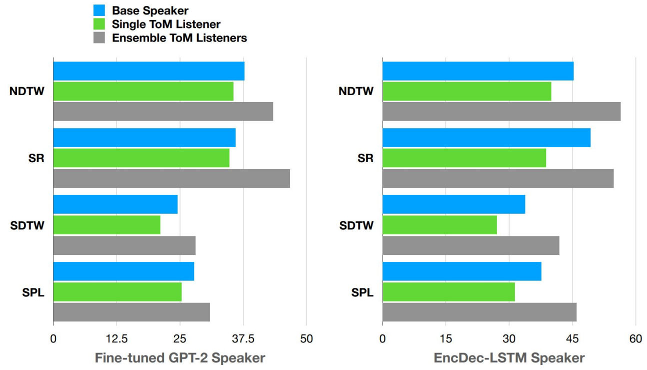


Figure 6: Comparison of using single and ensemble ToM listeners.

*building, and are provided with a short set of instructions to navigate to a target location. Please follow the instructions as closely as possible. Do NOT explore the building unnecessarily and do NOT wander back and forth unless you are lost. Please read ALL of the instructions before you start moving.*

We waive consent for this study for several reasons: 1) Minimal risk: The study collects minimal identifying information and there are no known risks for the subjects beyond everyday computer use. 2) Rights and welfare: All participants will be shown all information regarding task requirements before they complete our survey. They must consent to performing the task before they are shown the questions. 3) Practicality: Since the sessions are conducted online on a large scale, it would be infeasible to require all users to send a signed form. 4) Post participation information: We do not think there is any pertinent information that is not already shared with the participants before or during our experiments, so we do not feel it is necessary to provide any additional information after participation. PI information will be shared with the participants to enable them to obtain additional information about the study post completion.

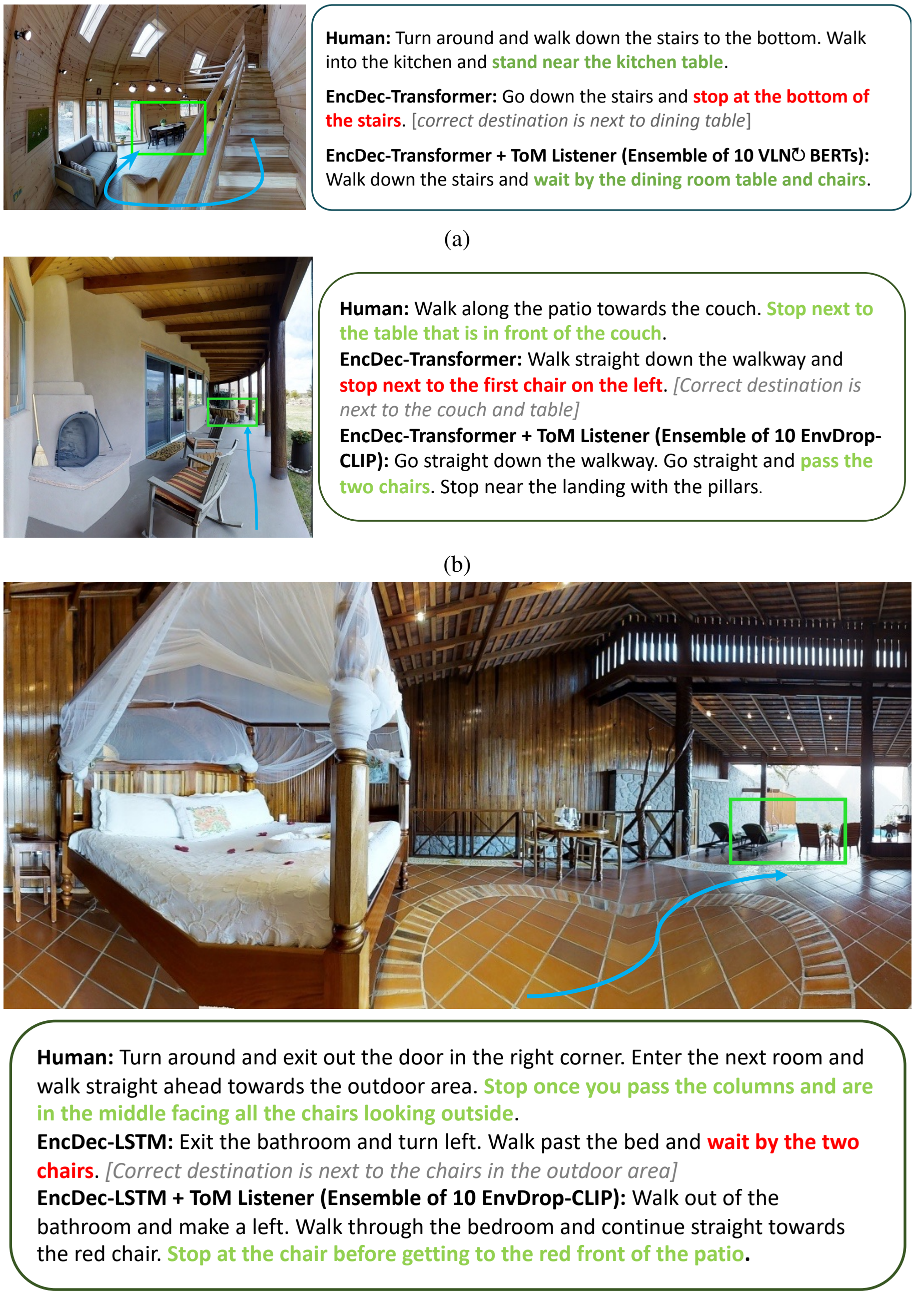
For data anonymization, we removed the only identifying information, Amazon Mechanical Turk ID, after collecting the human annotation data. This information would also be removed for future dataset release. The dataset will be released under MIT license terms, which are compatible with those of the tools used to create it, and will be intended for research usage.

### A.6. Single vs. Ensemble Listeners

As a preliminary experiment, we compare the effectiveness of a single and an ensemble of 10 VLN↻BERT agents when serving as the ToM model of a speaker. Results in Figure 6 show that the ensemble listener is significantly better than the single listener for two different speakers.

### A.7. Qualitative Examples

In Figure 7, we show qualitative examples where having a ToM listener enables the speaker to generate a more accurate instruction.



**Human:** Turn around and walk down the stairs to the bottom. Walk into the kitchen and **stand near the kitchen table**.

**EncDec-Transformer:** Go down the stairs and **stop at the bottom of the stairs**. [*correct destination is next to dining table*]

**EncDec-Transformer + ToM Listener (Ensemble of 10 VLN↻ BERTs):** Walk down the stairs and **wait by the dining room table and chairs**.

(a)

**Human:** Walk along the patio towards the couch. **Stop next to the table that is in front of the couch**.

**EncDec-Transformer:** Walk straight down the walkway and **stop next to the first chair on the left**. *[Correct destination is next to the couch and table]*

**EncDec-Transformer + ToM Listener (Ensemble of 10 EnvDrop-CLIP):** Go straight down the walkway. Go straight and **pass the two chairs**. Stop near the landing with the pillars.

(b)

**Human:** Turn around and exit out the door in the right corner. Enter the next room and walk straight ahead towards the outdoor area. **Stop once you pass the columns and are in the middle facing all the chairs looking outside**.

**EncDec-LSTM:** Exit the bathroom and turn left. Walk past the bed and **wait by the two chairs**. *[Correct destination is next to the chairs in the outdoor area]*

**EncDec-LSTM + ToM Listener (Ensemble of 10 EnvDrop-CLIP):** Walk out of the bathroom and make a left. Walk through the bedroom and continue straight towards the red chair. **Stop at the chair before getting to the red front of the patio.**

(c)

Figure 7: Qualitative examples where the pragmatic speaker (the last model) avoids missing information by simulating the interpretation of the human listener.