Speaking with Confidence: Investigating the effects of uncertainty in pragmatic language learning

Pawan Wirawarn[1], Rose E. Wang[1], Noah D. Goodman[1,2]
Departments of [1]Computer Science and [2]Psychology, Stanford University

Motivation

Main research question: How can we reliably train language models to generate contextually relevant utterances?

Prior work has investigated training pragmatic language models with communication-based objectives, where neural listeners stand in as communication partners. However…

Challenges include (a) obtaining a well-calibrated listener model, and (b) listener models are domain-specific, which often makes them overconfident about poorly generated utterances [1].

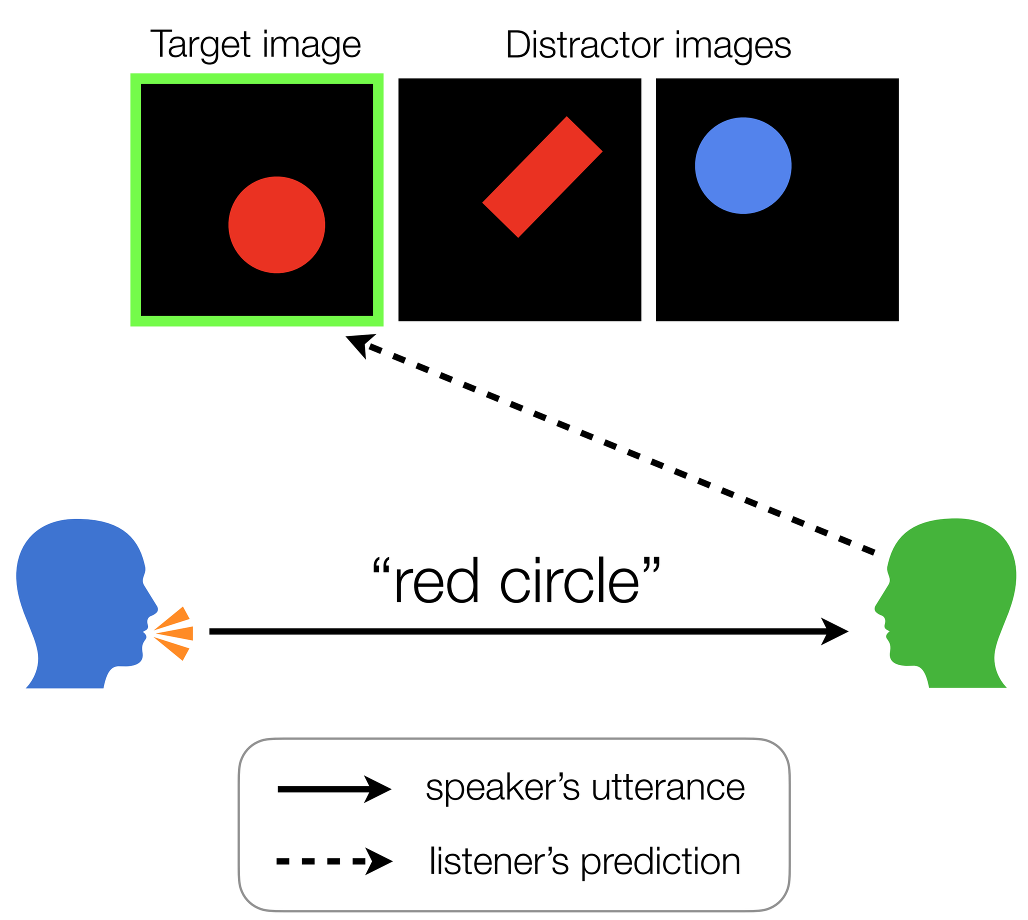
Our work explores whether pragmatic language learning is better with a well-calibrated domain-agnostic listener [2, 3].

Setup

We study the **problem of training a pragmatic speaker** for reference games with the ShapeWorld dataset [4].

A **reference game** $(\mathbf{I}, t)$ consists of n images $\mathbf{I} = (i_1, \ldots, i_n)$ and a target image i_t, with the index t known only to the speaker.

The **objective of the speaker** f_S is to produce an utterance u which allows the listener f_L to identify the target t given the images.



[1] *Calibrate your listeners! Robust communication-based training for pragmatic speakers.* Wang et al., 2021.
[2] *Revisiting the Calibration of Modern Neural Networks.* Minderer et al., 2021.
[3] *Using Pre-Training Can Improve Model Robustness and Uncertainty.* Hendrycks et al., 2019.
[4] *ShapeWorld - A new test methodology for multimodal language understanding.* Kuhnle and Copestake, 2017.
[5] *Learning Transferable Visual Models From Natural Language Supervision.* Radford et al., 2021.
[6] *Microsoft COCO: Common Objects in Context.* Lin et al., 2015.

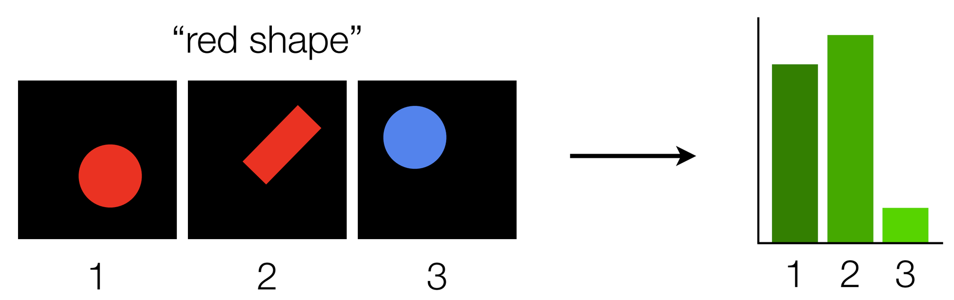
Method

Listeners (L)

We experiment with two types of listeners that differ in which dataset they were trained on. Both listeners are a distribution over possible targets in a reference game. Specifically:

$$f_L(t \mid \mathbf{I}, u) \ \propto \exp(g(i_t)^\top h(u))$$

where g and h are the listener's image and language encoders, respectively.



Domain-specific (DS) listener f_L^{DS} is trained on the ShapeWorld dataset.

Domain-agnostic (DA) listener f_L^{DA} is the CLIP model (Contrastive Language-Image Pre-training), which is pre-trained on 400 million (image, text) pairs collected from the internet [5].

Speakers (S)

Speakers are trained to produce an utterance for the listeners given a game and desired target. Specifically:

$$f_S(u \mid \mathbf{I}, t) \ = p_S(u \mid g(i_t), g(i_1), \ldots, g(i_{n-1}))$$

where g is the speaker's image encoder.

Our work considers three base speaker objectives:

- Domain-agnostic (DA) pragmatic training:

$$\mathcal{L}^{\mathrm{DA}}_{\mathrm{prag}}(\hat{u}|\mathbf{I}, t) = -\log f_L^{DA}(t|\mathbf{I}, \hat{u})$$

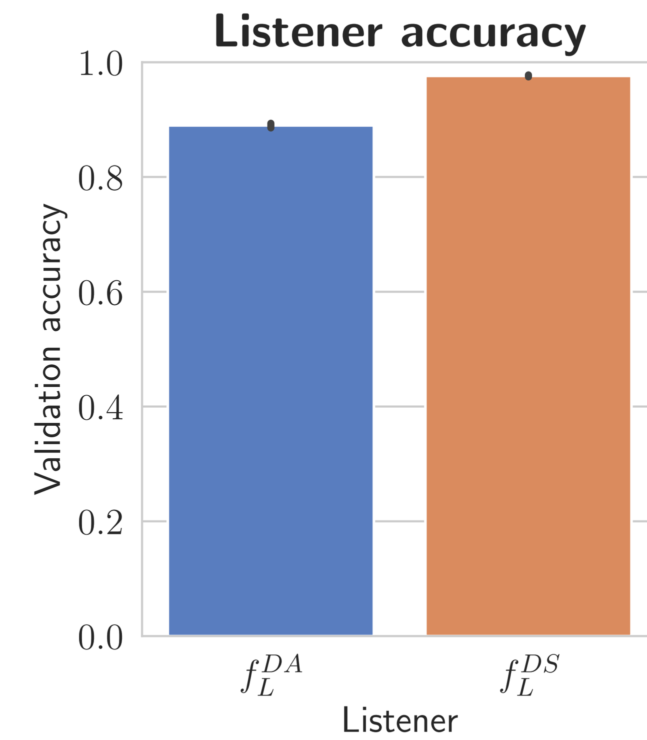
- Domain-specific (DS) pragmatic training:

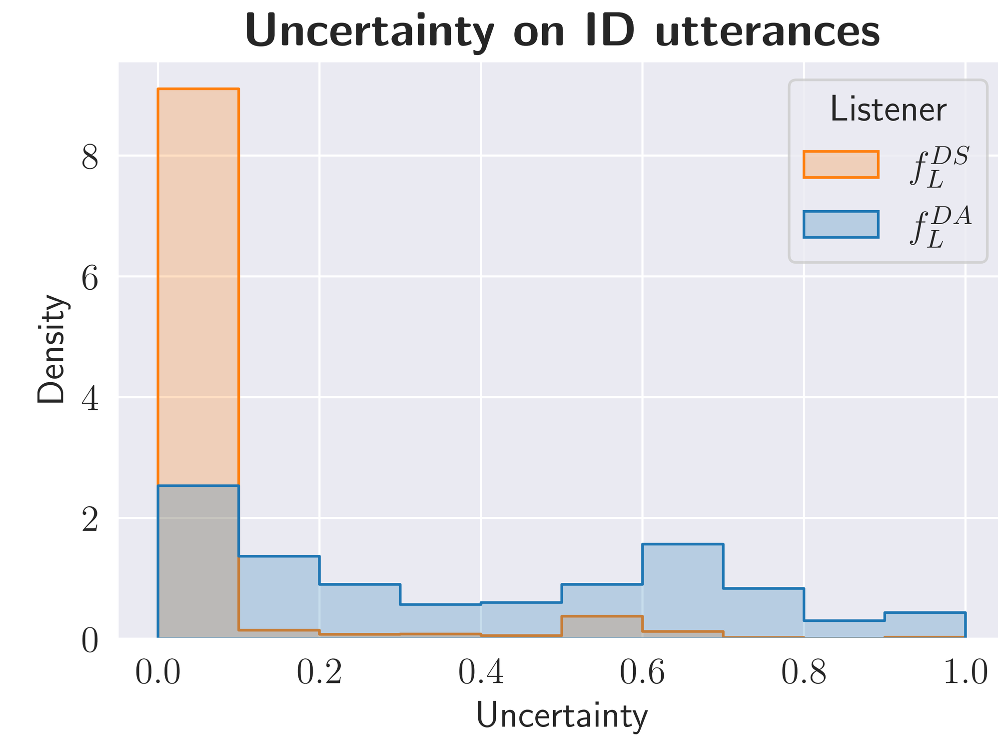
$$\mathcal{L}^{\mathrm{DS}}_{\mathrm{prag}}(\hat{u}|\mathbf{I}, t) = -\log f_L^{DS}(t|\mathbf{I}, \hat{u})$$

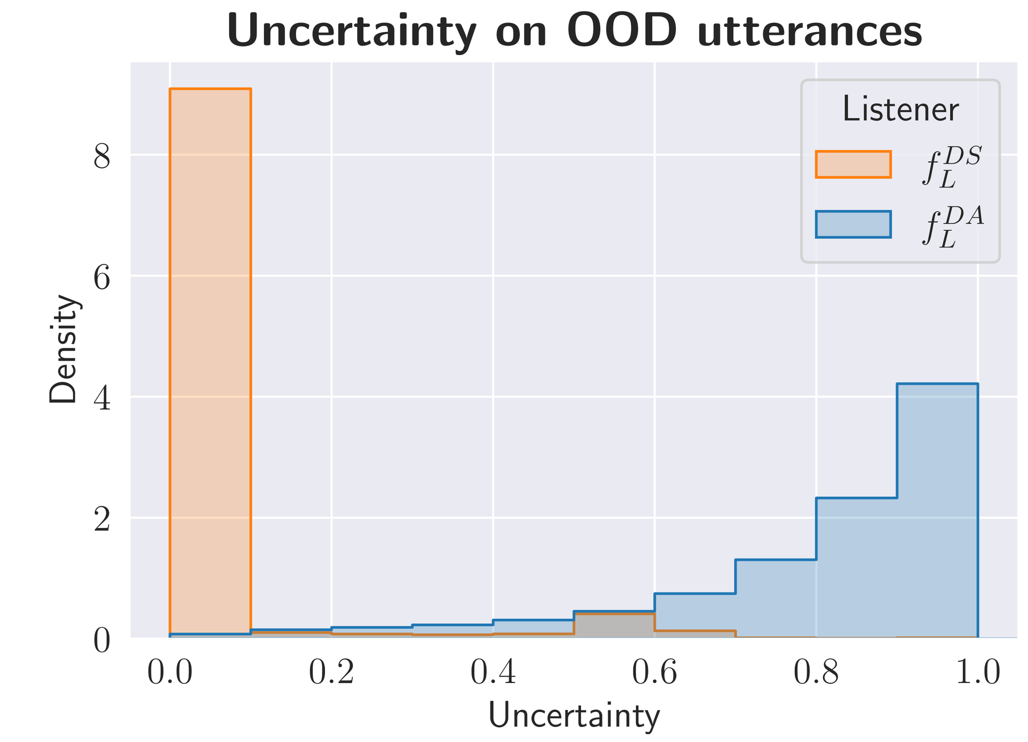
- Supervised (sup) training:

$$\mathcal{L}_{\mathrm{sup}}(\hat{u}, u) = -\sum_k \log p_S(\hat{u}_k = u_k | u_{<k}, \mathbf{I})$$

Results and Analysis

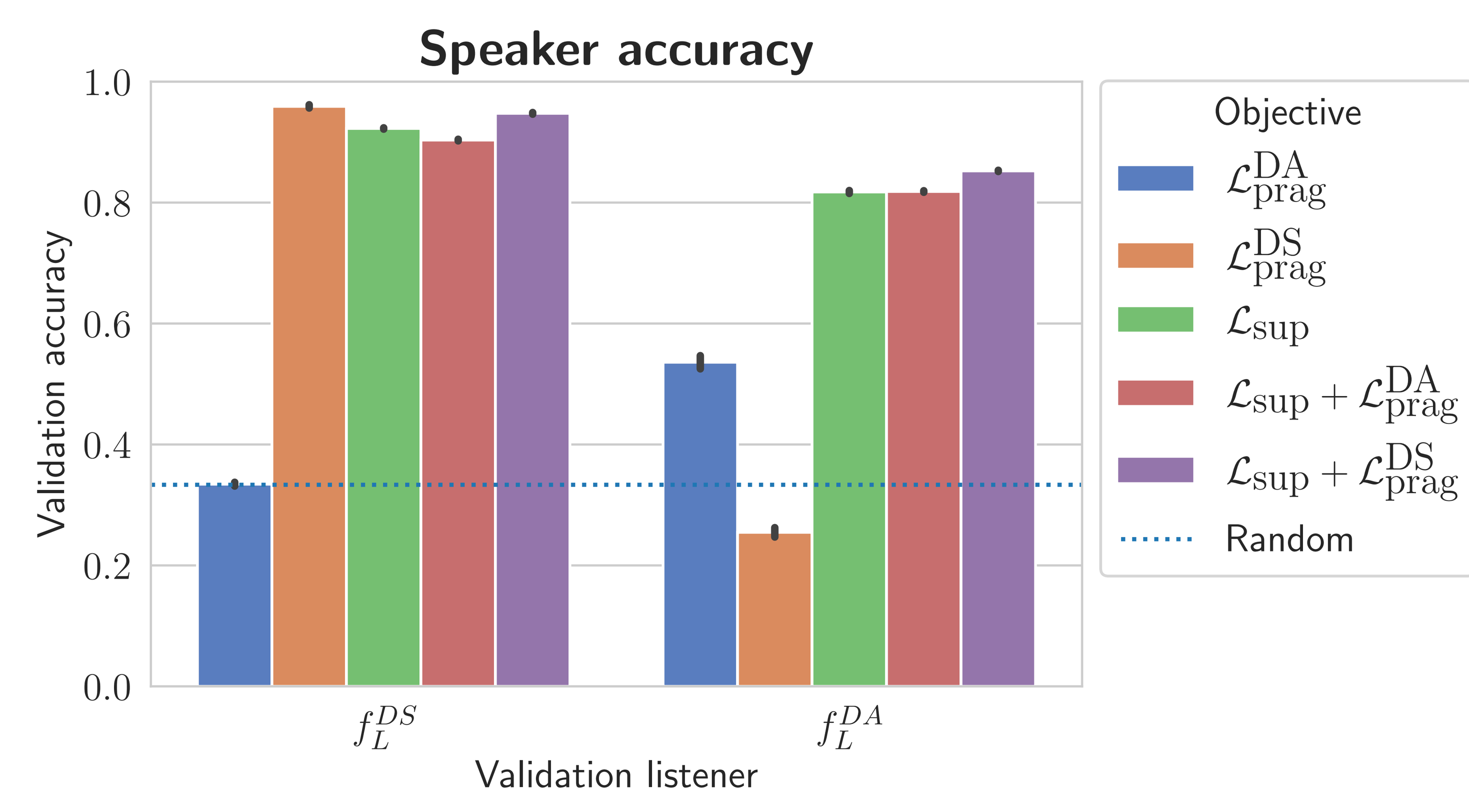


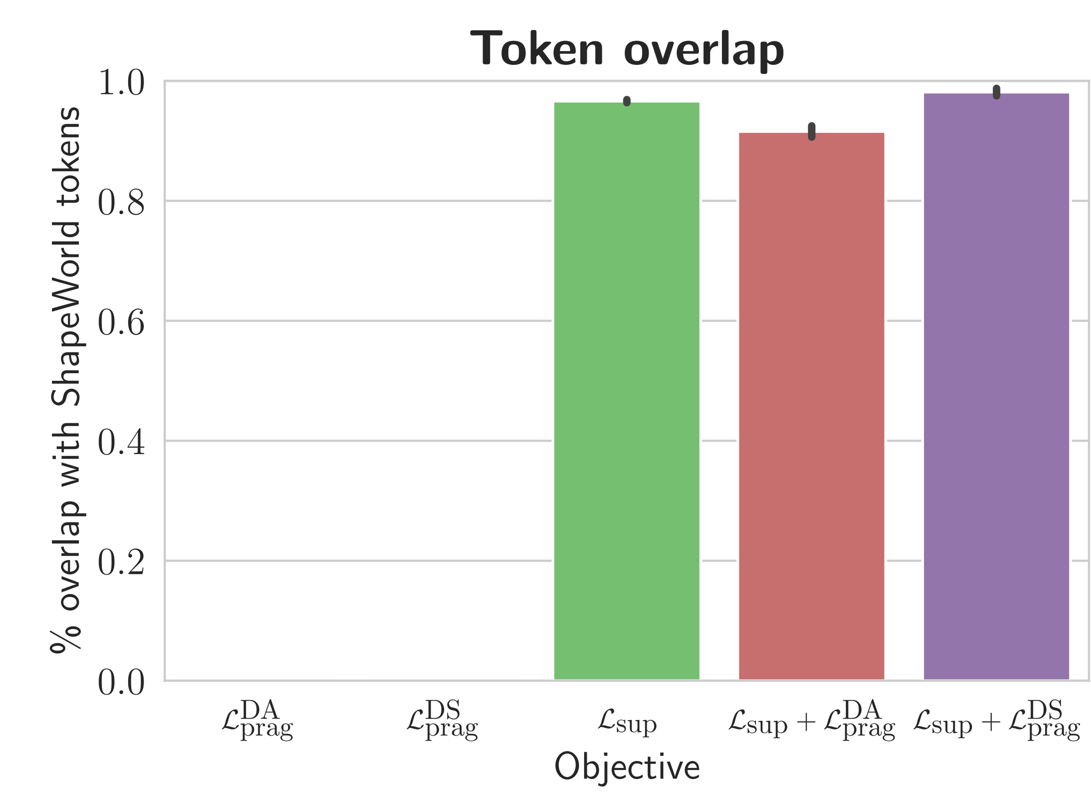




Listener takeaways:

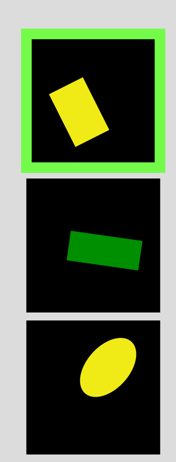
- Domain-agnostic (DA) listeners are better calibrated than domain-specific (DS) listeners: DA listeners can signal when an utterance is out-of-distribution (OOD).
- However, the DS listener is more confident about in-distribution (ID) utterances!





Speaker takeaways:

- Domain specificity and high-confidence in ID utterances is key to training pragmatic speakers: $\mathcal{L}_{\mathrm{sup}} + \mathcal{L}^{\mathrm{DS}}_{\mathrm{prag}}$ performs the best. Giving the speaker high rewards when it generates ID utterances is critical.
- Because the DS listener is more confident about ID utterances than the DA listener, the DS listener gives the speaker higher rewards for generating useful ShapeWorld utterances.



Examples of generated utterances

Objective	Utterance
ground truth	yellow rectangle
$\mathcal{L}^{\mathrm{DA}}_{\mathrm{prag}}$	siren dara dara dara
$\mathcal{L}^{\mathrm{DS}}_{\mathrm{prag}}$	lewis prize prize lewis
$\mathcal{L}_{\mathrm{sup}}$	yellow rectangle
$\mathcal{L}_{\mathrm{sup}} + \mathcal{L}^{\mathrm{DA}}_{\mathrm{prag}}$	religions
$\mathcal{L}_{\mathrm{sup}} + \mathcal{L}^{\mathrm{DS}}_{\mathrm{prag}}$	yellow rectangle

Discussion

- We show that the **domain specificity** of listeners and their **high confidence in in-domain utterances** is important for training pragmatic speakers.
- Our research can be extended to pragmatic language learning in other domains like COCO [6], where we can experiment with new variations of listener models and speaker objectives.