

# Calibrate your listeners!

## Robust communication-based training for pragmatic speakers

Rose E. Wang<sup>1</sup>, Julia White<sup>2</sup>, Jesse Mu<sup>1</sup>, Noah D. Goodman<sup>1,3</sup>

Departments of <sup>1</sup>Computer Science, <sup>2</sup>Electrical Engineering and <sup>3</sup>Psychology  
Stanford University

{rewang, jwhite, muj, ngoodman}@stanford.edu

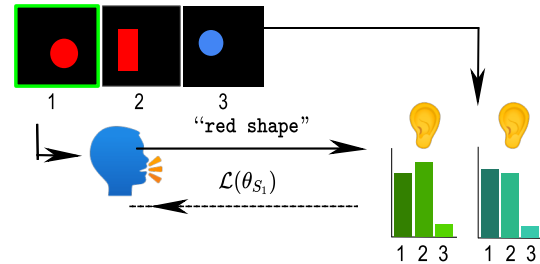
### Abstract

To be good conversational partners, natural language processing (NLP) systems should be trained to produce contextually useful utterances. Prior work has investigated training NLP systems with communication-based objectives, where a neural listener stands in as a communication partner. However, these systems commonly suffer from semantic drift where the learned language diverges radically from natural language. We propose a method that uses a population of neural listeners to regularize speaker training. We first show that language drift originates from the poor uncertainty calibration of a neural listener, which makes high-certainty predictions on novel sentences. We explore ensemble- and dropout-based populations of listeners and find that the former results in better uncertainty quantification. We evaluate both population-based objectives on reference games, and show that the ensemble method with better calibration enables the speaker to generate pragmatic utterances while scaling to a large vocabulary and generalizing to new games and listeners.<sup>1</sup>

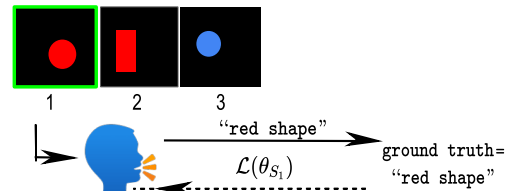
## 1 Introduction

To be good conversational partners, language models (LMs) should learn to produce fluent utterances that serve the needs of their listeners (Grice, 1975). However, they are often trained to capture the statistical, rather than communicative, properties of language via supervised learning (Bengio et al., 2003; Radford et al., 2019). Consider the reference game (Lewis, 1969) in Figure 1, where the goal of the speaker is to disambiguate a target image from distractor images. A naive LM may generate the literal description “the red shape”; this is semantically accurate, but fails to disambiguate the referent among its context. A more pragmatically useful description is “the red square”.

<sup>1</sup>The accompanying code can be found here: [https://github.com/rosewang2008/calibrate\\_your\\_listeners](https://github.com/rosewang2008/calibrate_your_listeners).



(a) Communication-based objectives, where a speaker is trained with a listener reward model.



(b) Traditional LM objectives, where a speaker is trained from a corpus of image descriptions.

Figure 1: The goal of the speaker (blue agent) is to generate context-aware utterances to refer to the image 1 and disambiguate from images 2 & 3. Our work uses (a) communication-based training objectives rather than (b) non communication-based objectives which rely on a corpus of ground-truth captions.

Prior work in training communicative NLP systems has used explicit models of pragmatics (Goodman and Frank, 2016) to fine-tune traditional LMs (Monroe et al., 2017; Andreas and Klein, 2016; Vedantam et al., 2017), or has trained LMs with external reward signals that indicate the contextual utility of an utterance. Human preferences are an ideal source of supervision for the latter approach (Stiennon et al., 2020; Ziegler et al., 2020), but are expensive to collect. One promising avenue is communication-based training (or *self play*), where a speaker learns to communicate with a *learned model* of a listener (Lazaridou et al., 2017; White et al., 2020). However, this approach commonly suffers from the problem of *semantic drift*. Our use of “semantic drift” follows the communication-

based training literature (Lazaridou et al., 2020; Lee et al., 2019): the speaker produces utterances that satisfy the listener but diverge from the semantics and conventions of natural language.

It can be difficult to diagnose the nature of semantic drift in complex, open domain models. In this paper, we isolate one form of semantic drift that occurs when scaling communication-based training to larger domains: by increasing the size of the vocabulary available to the speaker and listener, the speaker fails to generalize to new listeners. We identify miscalibrated listener uncertainty as the source of this problem: the listener is highly confident in the interpretation of utterances outside its training domain, so the speaker overfits and produces nonsensical utterances that fail to generalize. We propose to correct the calibration problem by using *populations* of listeners, drawing from both computational and cognitive science work that suggests deficiencies in just training with a single listener (Graesser et al., 2020; Raviv et al., 2019; Wagner et al., 2003) and improvements in emergent communication protocols by regularization effects of ensembled models (Li and Bowling, 2019; Tieleman et al., 2019). We find that ensemble-based populations of listeners are better calibrated, helping speakers avoid semantic drift and generalize to new games and listeners.

## 2 Approach

We study the problem of learning a pragmatic speaker for reference games with the ShapeWorld (Kuhnle and Copestake, 2017) dataset. A reference game  $(\mathbf{I}, t)$  consists of  $n$  images  $\mathbf{I} = (i_1, \dots, i_n)$  and a target image  $i_t$ , with the index  $t$  known only to the speaker. The speaker  $S$  must produce an utterance  $u$  which allows the listener  $L$  to identify the target  $t$  given the images.

Formally, given  $u$ , the listener  $L$  is a distribution over possible targets in a reference game:

$$L(t | u) \propto \exp(f_\theta(i_t)^\top g_\theta(u)) \quad (1)$$

where  $f_\theta$  and  $g_\phi$  are the listener’s image and language encoders, respectively.

The speaker is then trained to produce an utterance for the listener given a game and desired target. Specifically,  $S$  is parameterized by a gated recurrent neural network (GRU) (Cho et al., 2014):

$$S(u | \mathbf{I}) = p_{\text{GRU}}(u | f_\eta(i_t), f_\eta(i_1) \dots, f_\eta(i_{n-1})). \quad (2)$$

where  $f_\eta$  is the speaker’s image encoder used to initialize the GRU hidden state. The speaker is trained to maximize the listener’s probability of selecting the correct target given its utterance. Formally, the speaker’s loss over a game  $G$  is

$$\mathcal{L}(S; L, \mathbf{I}, t) = -\log L(t | \hat{u}), \hat{u} \sim S(u | \mathbf{I}). \quad (3)$$

This *informative communication* objective has been used several different ways in the past. Rational Speech Act models of pragmatic language use (Goodman and Frank, 2016) adopt it as a definition of speaker behavior, rather than an objective for training. Studies of emergent multi-agent communication (Lazaridou et al., 2017) use this objective to jointly train the speaker and listener, or alternatively train a pragmatic speaker against a fixed listener (White et al., 2020; Lazaridou et al., 2020).

We adopt the setting of White et al. (2020): we first train listener models on separate splits of ShapeWorld reference games, then train pragmatic speakers with the fixed listener models as the communication objective. The listeners and speakers do not share data splits among each other. **Training listener(s)** are used as the internal listener model for the speaker in Equation 3, and other **validation listeners** are held-out for speaker evaluation. All the speakers and listeners are trained on 15000 randomly generated reference games with a single target and 2 distractor images. More details on the model architecture, training, data, and the supplementary code can be found in Appendices A and B.

## 3 The problem of semantic drift

White et al. (2020) found successful pragmatic language production from a model trained to communicate informatively in the ShapeWorld domain. We first show that this success hinges on the restricted domain of utterances used. We consider two settings where the speaker learns to select utterances for two listeners: both are trained on the ground-truth ShapeWorld data, but one only has access to the ShapeWorld vocabulary (15 tokens), and one has access to the entire GPT-2 vocabulary (51k tokens; Radford et al. 2019). Given either listener, we train the speaker using Equation 2. We chose GPT-2 vocabulary, a byte-pair encoding (BPE) (Sennrich et al., 2016), to access a large tokenization space as we hope to extend communication-based training to using GPT-2 as the generation model. Since BPE is the typical

Table 1: Average speaker accuracy  $\pm$  standard error (middle 4 columns) and speaker-generated tokens overlapping with ShapeWorld tokens at test time (left-most column, “token overlap”) in % averaged over 10 seeds. Accuracy is reported with either the training or validation listener (train  $L$ , val  $L$ ), and on either training or validation set (train  $\mathcal{D}$ , val  $\mathcal{D}$ ). Gray cells indicate the speaker unable to generalize to new listeners and emitting novel tokens.

token space	train $L$ -train $\mathcal{D}$	train $L$ -val $\mathcal{D}$	val $L$ -train $\mathcal{D}$	val $L$ -val $\mathcal{D}$	token overlap
small	99 $\pm$ 1.40	90 $\pm$ 1.40	80 $\pm$ 1.40	80 $\pm$ 1.40	100 $\pm$ 0.0
large	98 $\pm$ 0.00	94 $\pm$ 0.01	37 $\pm$ 0.06	36 $\pm$ 0.06	1 $\pm$ 0.6

choice for most large pretrained language models, it is representative of large vocabulary spaces and introduces the challenges we might encounter in communication-based training for large speakers.

In Table 1, we report speaker communication accuracy with either the listener used for speaker training (train  $L$ ) or a separately trained held-out listener (val  $L$ ), and either reference games seen during training (train  $\mathcal{D}$ ) or held-out games (val  $\mathcal{D}$ ). To measure the extent the speaker uses domain-relevant vocabulary, we additionally report the percentage of speaker-generated tokens that overlap with ShapeWorld captions, such as “blue circle” or “red square”.

The results reveal two noticeable discrepancies between the small and large vocabulary settings. First, speakers trained over the large vocabulary space can successfully communicate with the training listener on new reference games – that is, they generalize to new contexts. However, they are unable to generalize to held-out listeners and achieve only random chance accuracy (33%). This pattern of overfitting to the training listener despite generalizing to new contexts with the training listener indicates that the speaker suffers a form of *semantic drift* (Lazaridou et al., 2020; Lee et al., 2019).

This semantic drift is shown most clearly in the second discrepancy: speakers trained with large vocabularies rarely use tokens related to shapes and colors (i.e. the domain-relevant tokens), whereas those trained with small vocabularies do so (by design). Upon qualitative examination of the speaker’s utterances (Table 2), we observe some examples of this drift: the speaker’s language deviates from natural language and the ShapeWorld domain that the neural listeners were trained on.

#### 4 Diagnosing semantic drift

We hypothesize that the observed semantic drift arises from poor uncertainty calibration in the neural listener, which the speaker then overfits to. If this is correct, then populations of listeners—

which have been shown to be better calibrated (Beluch et al., 2018; Li and Bowling, 2019; Tieleman et al., 2019)—may yield more useful listeners for communication-based training. To investigate this hypothesis we first explore the calibration of both single listeners and populations of listeners.

We compare the uncertainty measurements of three different internal listener implementations. One baseline is **single- $L_0$** , which approximates the internal listener  $L$  with a single neural listener, i.e. exactly the setup in Section 3 where existing work uses a single neural listener for communication-based training. The second is an **ensemble-based population** of  $n$  neural listeners; this is equivalent to substituting  $L(t | u) = \frac{1}{n} \sum_{j=1}^n L^j(t | u)$  in Equation 3 where each listener  $L^j$  is initialized randomly. They are trained on different data splits and different random seeds, and each achieves at least 92% accuracy on training and validation. While all listeners generalize to new in-distribution utterances, they generalize differently out-of-distribution. It is these disagreements among the ensemble that improve calibration compared to a single neural network like single- $L_0$ . The third is the **dropout-based population**, where we use a single listener to approximate a population of  $n$  listeners via MC-dropout (Gal and Ghahramani, 2016); this is equivalent to substituting  $L(t|u) = \frac{1}{n} \sum_{j=1}^n L^{\text{dropout}}(t | u, d_j)$  in Equation 3 where  $d_j$  is a randomly sampled dropout mask. In our experiments, we use a dropout rate of  $p = 0.1$  but found no differences with using larger rates while maintaining high listener accuracy; please refer to Appendix B for more details on the dropout implementation.

Intuitively, the ideal listener should be uncertain about utterances that are different from their training domain. To measure this, we evaluate listener uncertainty scores on utterances with varying degrees of overlap with their training domain. Specifically, for **high-overlap** utterances, we sample messages directly from the ground truth

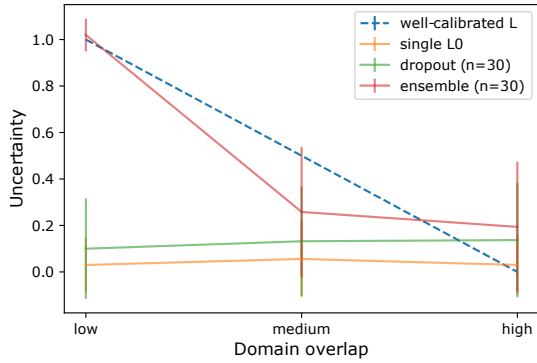


Figure 2: Listener uncertainty (entropy) by domain overlap. Scores come from either single  $L_0$ , dropout-based population, or an ensemble-based population. Dashed line is an idealized listener, for reference.

	Target	Distractors
<b>Method</b>		
gt	red shape	
single $L_0$	suppressed	suppressed Mix suppressed Pagan Mix Mix Pagan Mix
dropout	suppressed	suppressed suppressedops suppressed Mix Pagan imprison Mix
ensemble	rect rect	rect rect
<b>Method</b>		
gt	gray shape	
single $L_0$	ESEESE FDAWHAT Macedichickarios Macedichick	
dropout	ariosichickarios FDA FDAarios brainsichick Maced	
ensemble	gray gray gray	

Table 2: Examples of generated utterances. “gt” refers to the ground truth ShapeWorld utterance associated to the target image (left-most image). We show the utterance generated by the speaker trained with either the single  $L_0$ , dropout-based population, or ensemble-based population method. Based on token overlap and the reference game context, the ensemble-based population method results in more pragmatic and useful utterances.

ShapeWorld language. For **medium-overlap** utterances, we sample language from a well-calibrated

neural speaker (Table 2, ensemble). For **low-overlap** utterances, we sample from a poorly-calibrated speaker (Table 2, single  $L_0$ ). Refer to Appendix C for a complete discussion on utterance sampling. For these utterance types, the ideal listener should have maximum entropy over targets for low-overlap utterances, and zero entropy for high-overlap utterances. In Figure 2, we plot the listener’s uncertainty over utterances with varying domain overlap. We see, as hypothesized, that the single listener is poorly calibrated, yielding very low entropy for low-overlap utterances. Somewhat surprisingly, dropout-based populations do not help either. The ensemble-based populations, however, do show better calibration and are closer to the idealized listener: they are uncertain about out-of-domain utterances.

## 5 Countering semantic drift

Following our hypothesis, a well-calibrated listener should lead the speaker to generalize to new listeners *and* new contexts while reasoning pragmatically over a large vocabulary. We show that this is indeed the case in the following two experiments.

First, we show that with increasing population size  $n$ , the ensemble-based population objective closes the accuracy gap between training and validation listeners, while still generalizing to new reference games (Figure 3). This is not the case for the dropout listener with increasing the number of passes  $n$ . These results are averaged on 10 seeds, for  $n$  of  $\{1, 10, 20, 30\}$ . With ensemble-based population  $n = 30$ , we found that the speaker token overlap increased to  $48.4 \pm 12.3\%$ ; this is a much larger token overlap than the original speaker trained on single- $L_0$  ( $1.0 \pm 0.6$  from Table 1). We show utterance examples from speakers trained with the population objectives in Table 2.

To explore the semantic content of utterance distributions, we investigate the topic overlap between different speakers and the ground truth utterances. We find that speakers trained with well-calibrated listeners produce messages which are semantically similar to the ground truth utterances, according to distances in GloVe (Pennington et al., 2014) embedding space (Table 3). We embed the ground truth utterances by taking the sum of embeddings over the ground truth utterance. We denote this sum as  $z_{GT}$ . We compare these to the utterance embeddings of three speakers: one speaker which is constrained to Shapeworld vocabulary (used in



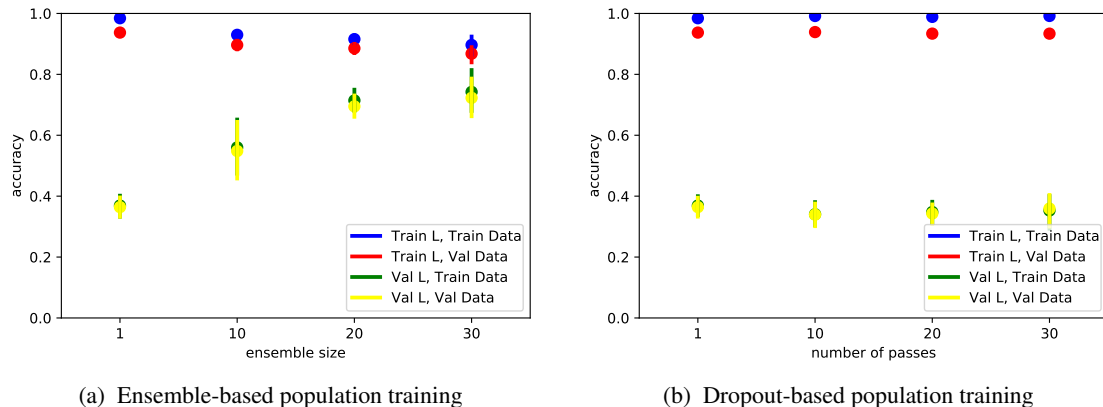


Figure 3: Accuracy and standard error for speaker trained with a (a) listener ensemble of size  $n = \{1, 10, 20, 30\}$  or (b) dropout listener with  $n = \{1, 10, 20, 30\}$  number of passes.  $n$  represents the population size. Results averaged over 10 seeds.

speaker type	$d(z_{GT}, \text{SUM}(z_S))$	$d(z_{GT}, \text{FIRST}(z_S))$
Limited	$9.2 \pm 3.74$	$6.43 \pm 2.54$
Calibrated	$10.37 \pm 2.41$	$8.52 \pm 0.50$
Miscalibrated	$13.03 \pm 3.62$	$8.72 \pm 0.44$

Table 3: Euclidean distance  $d$  in GloVe embedding space between the ground truth utterances  $z_{GT}$  and speaker utterances  $z_S$ . Lower distances indicate that the speaker’s utterance is close in topicality to the ground truth. The speakers are either trained only on the Shape-world vocabulary (Limited), trained with an ensemble-based listener population (Calibrated), or trained with single- $L0$  (Miscalibrated). The speaker utterance embeddings are either calculated as a sum of the uttered words ( $\text{SUM}(z_S)$ ) or as the embedding of the first uttered word ( $\text{FIRST}(z_S)$ ).

White et al. (2020) and in Section 3), one that is trained with well-calibrated listeners (ensemble-based population of size 30), and another that is trained with the poorly calibrated single- $L0$  listener. The speaker’s embeddings,  $z_S$ , are calculated either as a sum of uttered word embeddings or as the embedding of the first utterance word; we include the results from the latter because miscalibrated speakers often repeat words and taking only the sum might distort the calculated embedding distance. The distance reported in Table 3 is the Euclidean distance between  $z_{GT}$  and  $z_S$ . The results show that the calibrated speaker utters words that are similar to ground truth utterances, though not quite as similar as the limited speaker. The miscalibrated speaker utters words that are less similar to the ground truth utterances as indicated by the higher distance scores.

## 6 Conclusion

The true objective for language use is communication. Our work highlights the importance of well-calibrated listeners in communication-based training. We show that it’s important to understand the properties of speakers optimized to communicate with different listeners. We have found that naive communication-based training over unconstrained vocabularies is subject to a pernicious form of semantic drift that arises from poorly calibrated listener models. The overconfidence of listeners is exploited by the speaker during training and, as a result, speakers acquire niche linguistic properties, like conventions that fail to generalize to other listeners (Hawkins et al., 2017; Graesser et al., 2020). By contrast, an ensemble of listeners shows better calibration. Speakers trained to communicate with these listeners avoid semantic drift, generalizing to new games and new listeners. Future research on communication-based training for language models will thus benefit from listener ensembles or other methods for training listeners with properly calibrated uncertainty.

## 7 Acknowledgements

REW and JM are supported by the National Science Foundation Graduate Research Fellowship. This work was supported in part by the Stanford HAI Hoffman–Yee project ‘Towards grounded, adaptive communication agents.’ We thank CoCo-Lab and the anonymous reviewers for their helpful feedback on the paper, and Ke Alexander Wang for helpful input on Figure 1.

## References

- Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1182.
- William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. 2018. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.
- Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.
- Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.
- Laura Graesser, Kyunghyun Cho, and Douwe Kiela. 2020. Emergent linguistic phenomena in multi-agent communication games. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 3700–3710. Association for Computational Linguistics.
- H. P. Grice. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*. Academic Press, New York.
- Robert XD Hawkins, Mike Frank, and Noah D Goodman. 2017. Convention-formation in iterated reference games. In *CogSci*.
- E. Jang, S. Gu, and B. Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*.
- A. Kuhnle and A. Copestake. 2017. *ShapeWorld - a new test methodology for multimodal language understanding*. *arXiv preprint*.
- Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations (ICLR)*.
- Angeliki Lazaridou, Anna Potapenko, and Oliver Tieleman. 2020. 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 (ACL)*.
- Jason Lee, Kyunghyun Cho, and Douwe Kiela. 2019. Countering language drift via visual grounding. 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)*, pages 4385–4395.
- D. Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA.
- Fushan Li and Michael Bowling. 2019. Ease-of-teaching and language structure from emergent communication. *Advances in Neural Information Processing Systems*, 32:15851–15861.
- Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).
- Limor Raviv, Antje Meyer, and Shiri Lev-Ari. 2019. Larger communities create more systematic languages. *Proceedings of the Royal Society B*, 286(1907):20191262.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

- Olivier Tieleman, Angeliki Lazaridou, Shibl Mourad, Charles Blundell, and Doina Precup. 2019. Shaping representations through communication: community size effect in artificial learning systems. *arXiv*.
- Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Computer Vision and Pattern Recognition (CVPR)*, pages 251–260.
- Kyle Wagner, James A Reggia, Juan Uriagereka, and Gerald S Wilkinson. 2003. Progress in the simulation of emergent communication and language. *Adaptive Behavior*, 11(1):37–69.
- Julia White, Jesse Mu, and Noah D. Goodman. 2020. Learning to refer informatively by amortizing pragmatic reasoning. In *CogSci*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Huggingface’s transformers: State-of-the-art natural language processing](#).
- Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. [Fine-tuning language models from human preferences](#).

## A Software and data for reproducibility

The accompanying code can be found here: [https://github.com/rosewang2008/calibrate\\_your\\_listeners](https://github.com/rosewang2008/calibrate_your_listeners). It includes code for training the models, instructions, the data used for this work, and pretrained  $L$  models.

## B Implementation and training

**Model details** The listener and speaker have a separate vision encoder,  $f_\theta$  and  $f_\phi$ , which are convolutional neural networks with 4 blocks, each block containing a 64-filter 3x3 convolutional layer, batch normalization, ReLU activation, and 2x2 max pooling. Images are represented as 64x64x3 input, which produce 1024-d representations. For speakers, the image embeddings are projected down to the GRU hidden state size for hidden state initialization in  $p_{\text{GRU}}$ . For listeners, we take the dot product of the image embeddings and utterance embeddings to produce the target probability. The listener’s language encoder  $g(\cdot)$  is also a gated RNN. The listener and speaker’s GRU has two layers. It takes in an embedding dimension of 512 and has a hidden size of 100. The token space varies in our experiments: the small vocabulary setting has 15 tokens as in White et al. (2020) and the large setting has about 51k tokens from the GPT2 tokenizer (Wolf et al., 2020; Radford et al., 2019). The speaker’s maximum sequence length is 10 tokens.

For the dropout-based listener, we apply a dropout layer with dropout rate  $p = 0.1$  in the listener’s image and language encoders,  $f_\theta, g_\theta$ . Dropout is applied after every convolutional block except the last block in the image encoder. Dropout is applied on the outputs of each GRU layer except for the last layer. We experimented with higher dropout rates ( $p = 0.2, 0.3$ ) while maintaining high listener validation accuracy during listener evaluation, and found no differences in the reported experimental findings. Specifically, we found that using listeners with higher dropout rates did not change the uncertainty measurements shown in Figure 2. Speakers trained with listeners of higher dropout rates still obtained training and validation accuracies that were still within the margin of error reported in Figure 3.

The listeners in the ensemble-based population do not share any encoders and are trained on separate models.

**Training** All the models were trained on 100 epochs, with the Adam optimizer, a batch size of 32, using the Gumbel Softmax trick (Jang et al., 2017). The learning rate is 0.001 for speakers and 0.01 for listeners. All the listeners after training reach at least 92% validation accuracy, and are used for speaker training or validation.

## C Domain overlap in utterances

Intuitively, we want utterances with low domain overlap to be utterances that are unrelated ShapeWorld domain (e.g. those sampled from the speaker trained with the single- $L_0$  method as shown in Table 2), and utterances with high domain overlap to be captions correct for a given target image and perfectly related to the ShapeWorld domain and correct (e.g. those directly from the ShapeWorld dataset). Utterances with medium domain overlap should be an interpolation between low and high domain overlap. In our setting that is concerned with semantic language drift, we define overlap to be with respect to the percentage of utterance tokens that are ShapeWorld-related tokens.

We use the ShapeWorld image captions as the high domain overlap because these captions consist of only tokens related to the ShapeWorld setting, i.e. 100% overlap. We use utterances sampled from the speaker trained with ensemble-based population for the medium overlap utterances; these utterances generally use ShapeWorld-related tokens but in repeated fashion (see Table 2 for an example) and have 51.6% overlap. We use utterances sampled from the speaker trained with the single- $L_0$  method with 0% token overlap for the low domain overlap, as we found that this condition applied to a wide majority of utterances sampled from this speaker. We note that using token overlap as a proxy for domain overlap is an imperfect measurement as it doesn’t factor in whether an utterance is actually appropriate for a desired target image. For example, an utterance like “red green shape” uses ShapeWorld-related tokens, but may not be an ideal utterance for a red-shape target image. Nonetheless, since our work is concerned with semantic language drift that uses novel and out-of-domain tokens, we found this to be the best measurement of domain overlap compared to alternatives like n-gram overlap in tokens.