UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind

Takmaz, E.; Brandizzi, N.; Giulianelli, M.; Pezzelle, S.; Fernández, R.

DOI
10.18653/v1/2023.findings-acl.258

Publication date
2023

Document Version
Final published version

Published in
Findings of the Association for Computational Linguistics: ACL 2023

License
CC BY

Link to publication

Citation for published version (APA):
Takmaz, E., Brandizzi, N., Giulianelli, M., Pezzelle, S., & Fernández, R. (2023). Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023: July 9-14, 2023* (pp. 4198-4217). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.258

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Oct 2026

Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind

Ece Takmaz[◁*], **Nicolo' Brandizzi**[◇*],
Mario Giulianelli[◁], **Sandro Pezzelle**[◁], **Raquel Fernández**[◁]
[◁]University of Amsterdam [◇]Sapienza University of Rome
{ece.takmaz|m.giulianelli|s.pezzelle|raquel.fernandez}@uva.nl
brandizzi@diag.uniroma1.it

Abstract

Dialogue participants may have varying levels of knowledge about the topic under discussion. In such cases, it is essential for speakers to adapt their utterances by taking their audience into account. Yet, it is an open question how such adaptation can be modelled in computational agents. In this paper, we model a visually grounded referential game between a knowledgeable speaker and a listener with more limited visual and linguistic experience. Inspired by psycholinguistic theories, we endow our speaker with the ability to adapt its referring expressions via a simulation module that monitors the effectiveness of planned utterances from the listener's perspective. We propose an adaptation mechanism building on plug-and-play approaches to controlled language generation, where utterance generation is steered on the fly by the simulator without finetuning the speaker's underlying language model. Our results and analyses show that our approach is effective: the speaker's utterances become closer to the listener's domain of expertise, which leads to higher communicative success.

1 Introduction

Speakers tend to adapt their language use to the perceived knowledge, information, and linguistic abilities of their interlocutors (Isaacs and Clark, 1987; Clark, 1996; Pickering and Garrod, 2004). When adults speak with children, for example, they use simplified expressions to ensure children are able to understand (Saxton, 2009); when computational linguists give a talk at a cognitive science conference, they (hopefully) avoid making extensive use of NLP jargon, as that would prevent their audience from following through the presentation. Successful adaptation to the conceptual knowledge of conversational partners requires the ability to represent and reason about others' mental states

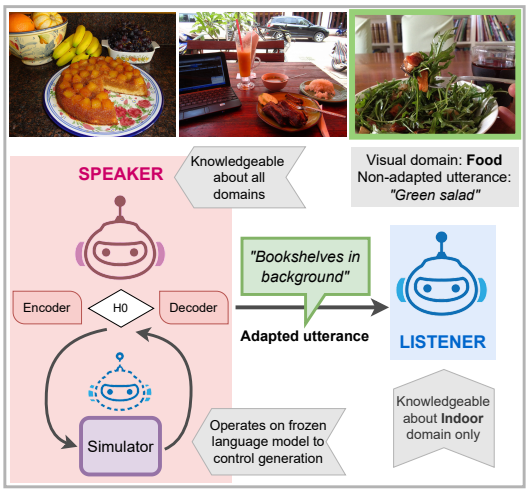

Figure 1: An illustration of our knowledge-asymmetric setup where an expert *Speaker* interacts with a less knowledgeable *Listener*. The *Speaker* tailors its utterance about an image from the **food** domain for a *Listener* who only knows about the **indoor** domain. The speaker's *Simulator* module inspired by Theory of Mind guides this adaptation. The adapted utterance exploits indoor terms ('*bookshelves*') without referring to food.

(Tomasello, 2005), a socio-cognitive ability typically referred to as Theory of Mind (ToM; Premack and Woodruff, 1978). Yet, speakers do not always resort to explicitly modelling the knowledge of their dialogue partner: due to different cognitive costs and pressures, they sometimes plan their utterances egocentrically, i.e., only taking into account their own knowledge and abilities (Keysar, 2007).

In this paper, we model a communicative situation where the interlocutors have *asymmetric language abilities*: a proficient speaker interacts with a listener characterised by limited semantic knowledge to complete a reference game, as illustrated in Fig. 1. Our goal is to mimic a scenario in which, for example, a high school physics professor can make complex atomic models understandable to young students by using terminology familiar to them, such as culinary terminology to explain Thomson's

[*]Shared first authorship.

Findings of the Association for Computational Linguistics: ACL 2023, pages 4198–4217
July 9-14, 2023 ©2023 Association for Computational Linguistics

'plum pudding model'. We focus on the speaker's Referring Expression Generation (REG; Reiter and Dale, 1997; Krahmer and van Deemter, 2012) in a multimodal dialogue setting and use REG models equipped with visual perception to generate discriminative image descriptions within a set of related image candidates. Several psycholinguistic theories have proposed that language production is interwoven with comprehension via 'forward prediction'—i.e., producing an utterance involves predicting how a comprehender would understand it (e.g., Pickering and Garrod, 2013; Roelofs, 2020). Inspired by this idea, we equip our speaker model with a *simulator*, i.e., a module that 'simulates' whether a listener would be able to identify the target referent. Based on this predicted behaviour (i.e., the expected effect of the planned utterance), the simulator modifies the generation plan on the fly to increase communicative success.

These are the main contributions of our study:[1]

- We model adaptation between agents with asymmetric knowledge, using a referential task as case study, where agents communicate in natural language about realistic images (in contrast to related work using synthetic data—see §2).
- We propose a novel simulation-based approach and test it in two settings: (1) a *self-aware* setting where the speaker predicts how a generic listener (with the same knowledge as the speaker) would resolve a planned utterance, and (2) an *audience-aware* setting where the speaker learns—from the behaviour of a listener with restricted semantic knowledge—to form representations of the listeners' knowledge (Clark, 1985; Isaacs and Clark, 1987) and predict their responses.
- We exploit the simulator's representations in an innovative way: by leveraging a *plug-and-play* approach originally introduced for controllable text generation (Dathathri et al., 2020), which steers language production at the decoding stage without altering the underlying language model.
- We show that our approach leads to increased resolution accuracy; in particular, our audience-aware speaker is able to adapt its utterances effectively when referring to a target within a visual domain unknown to the listener.
- We provide an in-depth analysis of the patterns present in the adapted utterances and the model's production strategies underpinning our results.

[1]Code and models available at https://github.com/nicofirst1/speaker-adaptation

2 Related Work

2.1 Pragmatic Reference Generation

Speakers tend to design their referring expressions to be pragmatically informative, i.e., discriminative from the listener's perspective. Most approaches to pragmatic reference expression generation (REG) have considered scenarios where we can assume a shared set of linguistic conventions between speakers and addressees (common domain and training data). The Rational Speech Act framework (RSA; Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; Goodman and Frank, 2016) has become a popular option for characterising such settings, with REG models that reason probabilistically about their interlocutors' interpretation via recursively defined speaker and listener models (Andreas and Klein, 2016; Monroe et al., 2017; Cohn-Gordon et al., 2018; Zarrieß and Schlangen, 2019; Fried et al., 2021), possibly taking into account information accumulated during interaction (Hawkins et al., 2020). There also exist joint speaker-listener models that are not recursive in the RSA sense. In these models, speakers can become listener-aware at inference time thanks to enhanced decoding algorithms (Vedantam et al., 2017) or they can learn to generate discriminative utterances at training time, for example via altered supervised training objectives (Mao et al., 2016) or auxiliary reinforcement learning (RL) modules (Yu et al., 2017), including approaches where the RL rewards are determined by the reference resolution success of a listener model (Lazaridou et al., 2020).

Our model, too, produces audience-aware discriminative image descriptions through an auxiliary module that captures the listener's perspective. However, in contrast to the above studies, the setting we investigate has two distinct key features: (1) we model situations with *knowledge asymmetry* between the dialogue participants, and (2) we experiment with *plug-and-play controlled generation* methods that result in temporary updates to the speaker's language model—rather than steering generation via recursive probabilistic reasoning. We review work related to these two aspects next.

2.2 Knowledge Asymmetry & Referring Tasks

What if the speaker and the listener have access to differing semantic knowledge? It is well known that speakers are able to adapt to less proficient

addressees (Isaacs and Clark, 1987). Janarthanam and Lemon (2010) were one of the first to address adaptation in dialogue systems with asymmetric knowledge. They modelled REG for technical domains where users may not know the jargon, using RL to learn a REG policy from a user simulation. More recently, Ohashi and Higashinaka (2022) focus on generating utterances in task-oriented dialogue with users that have limited vocabulary. They exploit the natural language understanding module of the system (representing user understanding) to set up a reward function, which is then used to finetune the NLG module via RL.

In the context of visually grounded referring tasks, Bao et al. (2022) focus on a scenario where the listener has comprehension difficulties and model adaptation by reweighing the probability of candidate referring utterances as a function of their likelihood to be successfully interpreted by the listener. Similarly, Liu et al. (2016) apply ToM-based listener modelling, where the speaker generates multiple candidate utterances and ranks them with the help of the ToM listener. Generating and ranking multiple utterances, however, is an inefficient production mechanism. For these reasons, others have tried to condition the speaker model prior to utterance generation, mainly with external modules. Corona Rodriguez et al. (2019) model interactions where the listener has an impaired perceptual system and implement this conditioning through an external policy network that takes as input listener embeddings. While Zhu et al. (2021) propose a ToM module that tracks the listener's understanding via meta-learning for few-shot coordination in a setup where listeners understand different languages. Singh et al. (2023) train an attention-based adapter layer in a reward-based manner as part of a multi-agent referential game where the speaker aims to generate utterances that would be understood by one listener, but not the other. Finally, Greco et al. (2023) have a setup that is the most similar to ours, where Expert speakers adapt to Layman listeners. But unlike our *plug-and-play* approach, the authors follow the RSA framework in developing audience-aware models that are updated through interaction.

2.3 Adaptive Controlled Generation

Most of the approaches to adaptation we have reviewed apply RL to the speaker model or finetune its language model through interaction. As a result, the speaker is not able to retain its original knowledge, which might cause catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999). With the advent of large pretrained language models, a plethora of new methods for controlled text generation have been proposed, including prefix-tuning (Li and Liang, 2021; Ben-David et al., 2022), prompting (Brown et al., 2020), adapters (Houlsby et al., 2019; Pfeiffer et al., 2020a,b), and energy-based constraints (Qin et al., 2022). Visual prefixes and prompts (Alayrac et al., 2022) have also been used to condition generation, especially without training the full language model.

We argue that this recent line of research offers promising alternative frameworks for adaptive REG. In particular, we investigate a solution to adaptation inspired by the *plug-and-play* approach to controlled text generation (PPLM; Dathathri et al., 2020; Pascual et al., 2021), which has been used to steer large pretrained language models towards generating texts with certain features (e.g., positive/negative sentiment or a given vocabulary distribution). In Dathathri et al. (2020), latent representations are updated at inference time with the help of a classifier while keeping the model parameters unchanged. Building on this idea, we propose a modular approach to REG adaptation in asymmetric knowledge settings where a module trained to predict the listener's behaviour—similar to the 'prediction net' in the machine ToM model by Rabinowitz et al. (2018b)—is exploited to control generation on the fly.

3 Problem Formulation

We provide an abstract overview of the problem we address and our approach. Details on the data and the experimental pipeline are given in §4 and §5.

Scenario Our setup is a classic referential game: two artificial agents, a speaker and a listener, share a visual context involving multiple images. The speaker produces an utterance to refer to one of the images (the target) and the listener attempts to identify the referent given that utterance. In particular, we model a scenario with *knowledge asymmetry*, where the speaker is more knowledgeable than the listener. We hypothesise that, in such a setup, for communication to be successful, the speaker will need to adapt its utterances to the listener's representational space and language. To make this possible, we endow the speaker with a simulation module and an adaptation mechanism.

Simulation We provide the speaker with a module that simulates how a listener would process a planned utterance. We assume that, by having interacted with listeners in the past, the speaker has learned a model of certain listener types (e.g., a prototypical idea of what a 3-year-old would understand). We operationalise this by pretraining several instances of the simulator, one per listener type, to predict how a listener is likely to resolve a referring utterance. We compare three settings:

- *Baseline*: No simulation takes place.
- *Self-aware*: The simulator is trained to predict how a listener with the same knowledge as the speaker would resolve an utterance. This is equivalent to a pragmatic speaker who reasons about the effect of its utterances on a generic listener (see §2.1), but in our approach at test time the listener's interpretations are predicted rather than directly observed. Our proposal is also inspired by human production models based on 'self-monitoring' (Levelt, 1993; Roelofs, 2020).
- *Audience-aware*: The simulator is trained to predict how a listener with a subset of the speaker's knowledge, i.e., a single domain, would resolve an utterance. Thus, the speaker learns a model—a theory of mind—of a less knowledgeable listener type that allows the speaker to make predictions about the listener's behaviour. When performing the referential task, we assume that the speaker knows the type of the listener beforehand, i.e., which simulator needs to be engaged (similarly to knowing that we are addressing a 3-year-old, for example).

Adaptation Rather than finetuning the speaker's language model, we exploit the pretrained simulators to control utterance generation on the fly via a monitoring loop. The simulator checks whether planned utterances would be effective; if that is not the case, a loss is backpropagated to update the initial hidden state of the speaker's decoder and a new utterance is generated. Our hypothesis is that such a mechanism will lead to referring utterances that are adapted to the listener's knowledge.

4 Data

As a basis for our experiments, we use the PhotoBook dataset (PB; Haber et al., 2019), a collection of task-oriented visually grounded English dialogues between pairs of participants who communicate via written chat. In a PhotoBook game, two participants see their own private sets ('photobooks') of real-life images belonging to the same visual domain. The goal of the interaction is for them to find out which images they have in common. This elicits referring utterances such as *"I have a little boy holding a phone to a teddy bear"*, where participants refer to an image that their dialogue partner needs to identify among six similar images. Our focus is on the generation of such referring utterances, leaving aside the dialogue context for simplicity. We use the dataset of referring utterances automatically extracted from the PB dialogues by Takmaz et al. (2020), which includes 41,340 utterances paired with their target image and the other five images in the visual context.

We choose PhotoBook because its visual contexts consist of realistic images and feature multiple challenging distractors, all selected from the visual domain of the target image. The original images are taken from the Microsoft COCO dataset (Lin et al., 2014) and belong to 30 different visual domains (e.g., *'person-umbrella'*, *'car-motorcycle'*). To model speaker adaptation to different semantic domains, we split the dataset of PB referring utterances according to the visual domain of each game. We cluster the image domains as a function of the similarity between their vocabulary vectors, constructed by counting word frequencies in the referring utterances belonging to a given domain. We obtain a set of 5 macro-domains (*appliances*, *food*, *indoor*, *outdoor*, *vehicles*), selected so that the domain vocabularies have minimal overlap. For each cluster of visual domains, we extract the corresponding referring utterance and visual context. We then randomly split these into training (70%), validation (15%), and test set (15%). We also merge the 5 domain-specific datasets into an *'all-domains'* dataset to be used to train domain-general models as described in §5. See summary in Table 1.

5 Experimental Pipeline

As described in §3, our experimental pipeline includes two agents—a speaker and a listener—implemented as a generative language model instantiating the speaker, a discriminative model instantiating the listener, and a third model, a simulator used by the speaker to assess the forward effect of its planned utterance on the listener. The language model and the discriminator model are adapted from those by Takmaz et al. (2020), and the simulator model is built on the discriminator's

Domain	Prop	N	$\|V\|$	Images	Specific	Overlap
Appliances	9.4%	4,310	1,271	36	29.5%	23.2% (*Ind*)
Food	12.4%	5,682	1,646	36	43.3%	22.9% (*App*)
Indoor	26.4%	12,088	2,477	96	44.3%	26.0% (*Out*)
Outdoor	35.9%	16,427	2,858	108	47.0%	26.2% (*Veh*)
Vehicles	15.8%	7,234	1,738	48	36.0%	26.2% (*Out*)
All	100%	45,741	6,038	324	-	-

Table 1: Statistics of the domain-specific datasets: # of utterances (N) and % within the entire dataset (Prop), vocabulary size ($|V|$), # of unique images (Images), % of domain-specific vocabulary (Specific), and max. lexical overlap with another domain (Overlap). The max. overlap is between *outdoor* and *vehicles*. Example shared words are *'left'*, *'black'*, *'driving'*, and *'glasses'*.

architecture with additional components. We train these models from scratch to have full control over the linguistic and visual knowledge of the agents and their degree of asymmetry. We use ResNet-152 to encode the images (He et al., 2016). See Appendix A for more information about the training schemes and hyperparameters.

5.1 Generative Language Model

The speaker is a visually conditioned language model that generates an utterance describing a target image within a visual context. The model follows an encoder-decoder architecture consisting of a visual encoder that represents the visual context along with the target image, and a decoder for language generation. The decoder generates a referring utterance via nucleus sampling (Holtzman et al., 2020), also paying attention to the encoder output at every time step. See Appendix A.1 for more details about the model architecture.

We train the visually conditioned language model from scratch using the training set including all domains in PB and optimize the model with respect to Cross Entropy Loss using Adam (Kingma and Ba, 2015). We select the best model based on its performance on a set of natural language generation metrics on the validation set. The weights of the trained speaker are then frozen and used as the core language generation model in all our experiments identically.

Performance The speaker's language model obtains reasonable scores in terms of classic natural language generation metrics:[2] 23.8 BLEU-2, 32.9 ROUGE, 44.1 CIDEr, and 57.7 BERTScore F1 (Papineni et al., 2002; Lin, 2004; Vedantam et al., 2015; Zhang et al., 2020). All scores are averages across 4 seeds on the test set. For details, see Appendix B.1.

5.2 Discriminator

Our listener is a discriminator model that receives six images in the visual context plus an utterance, and is tasked with identifying the target image that the utterance refers to. To encode the utterance, we use word embeddings trained from scratch to make sure no knowledge leaks from any pretraining. The model combines the visual context and the utterance to produce a multimodal context vector. The listener identifies the target image by comparing this multimodal context vector to the representations of each candidate image via dot-product and selecting the image with the highest score. See Appendix A.2 for the detailed description of the model architecture.

We train one listener model per domain in Table 1.[3] The models are optimized with Cross Entropy loss using the Adam optimizer. The best models are selected based on resolution accuracy on the validation set. We keep these domain-specific listener models frozen in the rest of the study. See Appendix A.2 for further details.

Performance We distinguish between *in-domain* (IND) accuracy—i.e., the resolution accuracy achieved on the test set of the domain on which the listener has been trained—and *out-of-domain* (OOD) accuracy—accuracy on domains the listener has not been exposed to (e.g., the accuracy on images from the *vehicles* domain of a listener exclusively trained on the *food* domain). Our listeners are truly domain specific: they are able to identify the target image with an average accuracy of 83.08% in IND, while their OOD accuracy is 19.05% on average—barely above a random baseline (16.67%). See Appendix B.2 for the full results broken down per domain.

5.3 Simulator

As explained in §3, the speaker is endowed with a simulator module. The simulator receives inputs in two parallel streams. In one stream, it receives the visual context $\mathbf{v}$ coupled with the speaker's planned utterance u_t, and in the second stream, the visual context along with the language model's initial hidden state h_0. The motivation behind this

[2] Comparable to those obtained by Takmaz et al. (2020) with their 'Ref' model.

[3] We also train a general listener model on all domains which is only used to train the self-aware simulator; see §5.3.

architectural choice is related to the plug-and-play approach at the core of our proposal. The first stream is inspired by previous work on ToM (e.g., Rabinowitz et al., 2018a): its main input is the same as what a listener would receive, an utterance. However, to control generation on the fly, we need to modify the language model's internal representations. Thus, the main reason for the second stream is technical: the gradients from the simulator's loss cannot flow back to the language model's hidden states if the input to the simulator is text due to the non-differentiability of the *argmax* operation.[4] The second stream uses a combination of linear layers and standardization to compute the dot product between h_0 and $\mathbf{v}$. The outcomes of the two streams are multiplied to obtain the final representation that is compared to the candidate images.

We train one audience-aware simulator per domain-specific listener and one self-aware general simulator with Cross Entropy loss using the AdamW optimizer (Loshchilov and Hutter, 2017). The training set sizes of both types of simulators are the same, with the target behaviour being different. In the simulation of a general listener, the simulator predicts the behaviour of a listener that was exposed to all domains as the speaker, contrary to one domain in the domain-specific case. We choose the best simulator per listener type based on the simulators' prediction accuracies (more details in Appendix A.3). The simulators are then frozen in the rest of the pipeline.

Performance The self-aware simulator achieves an accuracy of 70% when predicting the behaviour of a general listener. The audience-aware simulators predict the behaviour of domain-specific listeners with an average accuracy of 78.20% for IND samples, and 72.78% for OOD samples.[5] The drop in accuracy from IND samples to OOD samples could be due to difficulties in ascertaining the reactions of a listener on OOD data. See details of the results in Appendix B.3.

6 Audience-Aware Adaptation

In our framework, adaptation takes place at inference time building on our pretrained, frozen models for the language model, the discriminators and simulators described in §5. We first explain our adaptation mechanism (§6.1) and then report the results obtained (§6.2).

6.1 Adaptation Mechanism

Algorithm 1 describes the adaptation mechanism sketched in §3, which exploits the simulator to iteratively monitor the generation outcomes of the speaker. Given the visual context $\mathbf{v}$, the initial hidden state of the speaker's decoder h_0 and the currently planned utterance u_t, the simulator makes a prediction for the listener's selection.[6] We calculate the Cross Entropy loss between the simulator's prediction and the true target. We use the gradients flowing back from this loss to update h_0 with the Adam optimizer. That is, adaptation is performed by backpropagating the loss to modify only the initial hidden state of the speaker's decoder. Based on the updated h_0, the language model generates a new utterance to be reviewed by the simulator. The mechanism stops when: either (1) the simulator predicts that the listener will choose the gold target image; or (2) when the maximum number of adaptation steps is reached (st_{adp}). At each step, we reset the random seed to ensure that the changes in the sampling of the words are only attributable to the updates to h_0, showing the effects of adaptation directly without being confounded by the stochastic nature of sampling.

Algorithm 1: Adaptation Mechanism

Input: st_{adp} : maximum number of adaptation steps
lr_{adp} : learning rate for adaptation
$seed$: random seed
Data: h_0 : speaker's initial hidden state
$\mathbf{v}$: visual context
t_g : true target

$i \leftarrow 0$
while $i \leq st_{adp}$ **do**
  $set_seed(seed)$
  $u_t = Speaker(\mathbf{v}, h_0)$
  $o_{sim} = Simulator(\mathbf{v}, u_t, h_0)$
  $t_{sim} = \arg\max(o_{sim})$
  if $t_{sim} == t_g$ **then**
    break
  $loss = CrossEntropy(o_{sim}, t_g)$
  $h_0 = backprop(loss, h_0, lr_{adp})$
  $i \mathrel{+}= 1$
$t_l = Listener(\mathbf{v}, u_t)$

[4] We observed that using the Gumbel-Softmax trick (Jang et al., 2017) led to unstable behaviour.

[5] Possibly because the general knowledge space is bigger, it could also be more difficult to model a general listener than a domain-specific listener with a limited knowledge space.

[6] To avoid excessive language drift and help regularize utterance generation, at inference time we condition h_0 with the previous gold utterance referring to the target image in the current dialogue (if it exists), as done by Takmaz et al. (2020). This resonates with precautions taken in other plug-and-play approaches against text degeneration (Dathathri et al., 2020).

6.2 Results

We evaluate whether our approach leads to increased communicative success, quantified in terms of listener resolution accuracy. We report the results for the three settings described in §3. For each of the three modules involved in these settings, we provide an evaluation card (Hupkes et al., 2022) to clarify the nature of our generalisation tests in Appendix C.

Baseline Table 2 provides a breakdown of resolution accuracies per type of domain-specific listener in the setting without simulation; Table 3 shows the averages. Not surprisingly, the results obtained with generated utterances are lower than those reported in §5.2. However the patterns are the same: when the speaker agent refers to an image within a domain known to the listener (IND), the average resolution accuracy is 52.30%; communication however breaks down in out-of-domain instances, where the average OOD score is 19.06%, close to random choice.

	app	food	indoor	outdoor	vehi
appliances	57.61	20.10	19.92	21.27	15.98
food	19.11	54.29	18.60	18.85	18.85
indoor	22.71	19.65	53.62	20.82	16.77
outdoor	15.08	21.46	19.62	52.93	17.69
vehicles	16.36	16.17	17.41	20.13	43.08

Table 2: Resolution accuracy in the Baseline setting. Rows indicate the listener domain and columns the evaluation domain. Shaded cells show IND accuracy. Averages across 5 seeds. Full table with sd's in App. B.2.

Self-aware adaptation As shown in Table 3, with the added capability to simulate and adapt to a generic listener, we observe an increase in IND resolution accuracy (52.30% vs. 65.09%). Yet, this setting does not help to bridge the knowledge gap between speaker and listener: when the input is OOD for a domain-specific listener, adaptation with a general simulator does not lead to higher communicative success (19.06% vs. 19.11%).

	Baseline	Self-aware	Audience-aware
OOD	19.06 ± 0.47	19.11 ± 1.12	26.74 ± 1.48
IND	52.30 ± 1.10	65.09 ± 1.98	71.77 ± 2.16

Table 3: Average resolution accuracy for our 3 settings in OOD and IND. Results on the test set over 5 runs.

Audience-aware adaptation When the speaker adapts its utterances by predicting the behaviour of a domain-specific listener, we see a significant increase in both IND and OOD (Table 3). This indicates that audience-aware adaptation helps in knowledge asymmetric scenarios, including in IND situations where the agents communicate about a domain known to the listener (65.09% vs. 71.77%). More importantly, while there is certainly room for improvement, the speaker is able to generate utterances that can more often be resolved in OOD (19% vs. 26.74%).

7 Analysis

Our experiments show that simulation-based adaptation leads to more successful communication. In this section, we analyse the speaker model and its generated utterances to understand which neural processing mechanisms and which production strategies are behind our main results.

7.1 Probing for Domain Information

We begin with an analysis of the neural representations of the speaker model in the *audience-aware* setting. We focus on h_0, the first hidden state of the LSTM decoder. This is the output of the visual encoder on which the simulator module intervenes in order to adapt the speaker's utterance plan. Because h_0 is the result of encoding a target image (within a visual context), we expect it to carry information about the semantic domain of the image. If it was not able to differentiate visual domains, it would be very unlikely to successfully adapt to domain-specific listeners. We test this hypothesis using diagnostic probing (Adi et al., 2017; Conneau et al., 2018; Hupkes et al., 2018). We train a logistic regression classifier on a 70% of hidden states h_0 collected from the speaker when at test time, and then we assess whether it can predict the image domain corresponding to the remaining 30% of the hidden states. As expected, the probing classifier is able to do so with perfect precision and recall (both equal 1.0) across the 5 visual domains. Using the same approach, we test whether the domain of the *listener* – rather than the image domain – is also encoded in h_0.[7] Our hypothesis is that this should not be the case: before the simulator kicks in, the speaker model has no information on

[7]We train a logistic regression classifier on the 70% split of the h_0 but this time using as label the domain of the listener. We then evaluate whether the classifier can predict the listener domain in the 30% probing test set.

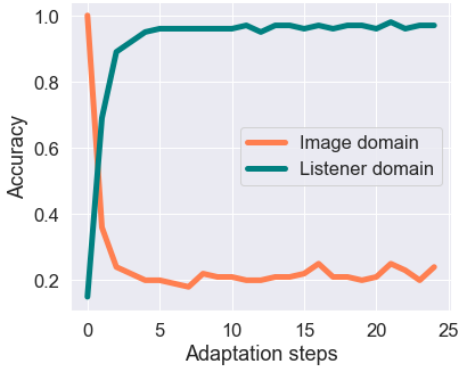

Figure 2: Probing accuracy for image domain and listener domain predictions over adaptation steps. The 0-th step corresponds to the non-adapted h_0.

the listener's domain-specific knowledge. Probing accuracy scores vary between 0.13 and 0.16 across domains (the random baseline is 0.17), indicating that indeed the speaker's hidden state does not carry listener information before adaptation.

As the simulator activates, the original h_0 is updated for a maximum of st_{adp} adaptation steps. We now look at the updated hidden states $h_0^1, \ldots, h_0^{st_{adp}}$ and test whether their encoding of the image and the listener domain changes with adaptation. First, we use the probing classifier previously trained to predict image domains from h_0 to test the adapted hidden states. We find that the encoding of the image domain deteriorates with domain-specific adaptation (Figure 2). Then, we probe $h_0^1, h_0^2, \ldots$ for listener information and we show that the listener's domain can be predicted almost perfectly from the adapted h_0 after only three adaptation steps (Figure 2).[8] Taken together, these observations indicate that the neural processing mechanism that leads to more successful interaction is one by which information about the semantic domain of the visual context is replaced by information on the domain of the listener – and one which only requires a few gradient updates.

7.2 The Speaker's Adapted Vocabulary

We analyse macro-level properties of the corpus of adapted utterances as compared to the utterances generated in the simulator-less baseline setting. We compute type-utterance ratio and type-token ratio over adaptation steps to monitor the relative size and the variety of the vocabulary as the speaker uses its simulator module. As Figure 3 shows, after an initial drop for the first 1-3 adaptation steps, type-utterance ratio and type-token ratio increase substantially with respect to the non-adapted utterances (and to the gold referring utterances). The speaker vocabulary becomes much more diverse. What remains rather stable throughout adaptation, instead, is the unigram part-of-speech distribution (Figure 7 in Appendix E). While, after the first adaptation step, the difference in POS usage is notable (e.g., less punctuation, more nouns), only proper nouns and determiners show substantial changes in relative proportions, with proper nouns increasing and determiners decreasing over time.

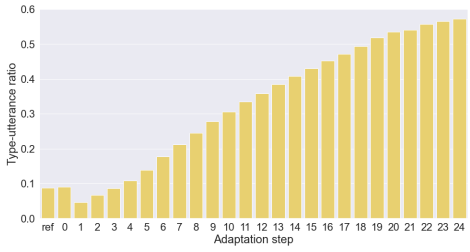

Figure 3: Type-utterance ratio across adaptation steps (type-token ratio in Fig. 6, App. E). Human gold utterances (*ref*) and non-adapted utterances (0) also shown.

7.3 Adaptation Strategies

The trends observed so far characterise the effect of adaptation across steps but they do not differentiate between successful and unsuccessful adaptation. In Figure 4, we split adapted utterances (the ones actually generated by the speaker when it believed its utterance would be successful) according to whether they lead to a correct listener guess. We observe that more successful utterances contain words with lower age of acquisition[9] (AoA, $t = -28.88$, $p < 0.001$), they show a lower rate of lexical choice from the target image vocabulary ($t = -28.76$, $p < 0.001$), and a higher rate of words from the listener vocabulary ($t = 5.88$, $p < 0.001$). The average AoA in an utterance increases with adaptation steps (see Fig. 8 in Appendix E), suggesting that the excessive abstractness of the descriptions may be behind the limited gains we observe with adaptation.

[8] For this analysis, we train and test one probing classifier for each adaptation step. Using the classifier trained on h_0 would not make sense as we showed that it is not possible to extract listener information from non-adapted representations.

[9] Age of Acquisition is a psycholinguistic measure expressing the age at which a word is typically learned. We use the ratings by Kuperman et al. (2012); they range from 0 to 25.

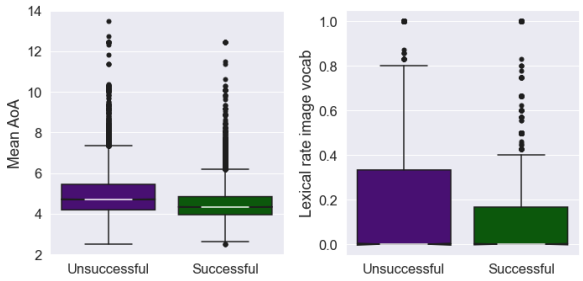

Figure 4: Factors affecting the success of an adapted utterance, age of acquisition (left) and % of words in an utterance belonging to the target image domain (right).

Target:FOOD. Listener domain:INDOOR
Gold (not adapted): *green salad with a person holding up a portion with fork?*
Generated (not adapted): *I have one more maybe round you think that has a lime green shaped greens, a salad?*
Adapted: *must bookshelves in the salad?*

Target:OUTDOOR. Listener domain:FOOD
Gold (not adapted): *I have the pink food truck again ... white shirt lady*
Generated (not adapted): *girl at black phone, red truck, brown hair, pink*
Adapted: *pink donuts*

Figure 5: Examples showing how audience-aware adaptation changes the generated utterances. For simplicity, we only show the target images and not the whole visual contexts. We report the final adapted utterances when the adaptation mechanism stops because the simulator predicts that the listener will select the correct image.

7.4 Qualitative Inspection

In Figure 5, we provide examples of adapted sentences from the test set to demonstrate how the audience-aware adaptation strategies affect the lexical choices made by the language model. In the top example, the image domain is 'food'; however, the listener was trained on the 'indoor' domain. We see that the speaker moves away from generating detailed mentions of food to including a word related to the listener's own domain, *bookshelves*. In the bottom example where the listener has only been exposed to the 'food' domain and the image domain is 'outdoor', the model avoids mentioning the *truck*. Instead, it produces an utterance containing a prominent color in the image, i.e., pink, and some visible entities that belong to the listener's domain, namely, *donuts*. These observations suggest that the model exploits various adaptation strategies.

In the whole set of adapted utterances, we observe comprehensible sentences; however, there is also a large number of less fluent, unnatural ones. As we do not use pretrained large language models, sometimes, the speaker's initial utterances themselves are not fluent. The dynamics of adaptation may further exacerbate this situation and lead the language model towards generating unnatural utterances. Such utterances may not be understood by human listeners; yet, they could make sense to artificial listeners. In order to ensure that the adapted utterances are comprehensible to humans, further precautions may be needed, such as incentivizing the generative model to keep the adapted utterances grammatical and fluent, possibly with the aid of human feedback.

8 Conclusion

We focused on a standard reference game—a speaker produces an utterance, and a listener uses it to pick the referent from a visual context. However, our setup is *asymmetric*—the speaker has general semantic knowledge, while the listener has little knowledge of all domains but one (e.g., *food*). Such a setting is a perfect scenario for studying adaptation, i.e., the common process in human communication by which a speaker tunes its language to that of a listener to achieve communicative success. We modeled this mechanism using a *plug-and-play* approach to controllable text generation: the speaker's output is conditioned on the effect of the planned utterance on the listener, as predicted by an internal simulator. Our results show that *speaking the language of a listener* increases communicative success. Through adaptation, the speaker's language becomes less tied to the input domain and more tied to the listener's vocabulary, revealing that audience-aware adaptation can be realized without irreversible changes to generation models.

Our approach and findings pave the way for pragmatic models that can account for different communicative scenarios. Future work may study adaptation to other dimensions such as age group or sociocultural background. Moreover, adaptation could be explored in multiple 'directions'—in our setup, only the speaker adapts. We also simplify the setup by abstracting away the online process that leads to the simulation of the listeners. It would be beneficial to allow the simulators to learn to predict listener behaviour during interaction in an online manner. Finally, our approach could be applied to other and possibly more complex communicative tasks, perhaps in conjunction with a mechanism leveraging human feedback via reinforcement learning.

Limitations

Although we use data from dialogues, we do not model collaborative reference, i.e., we do not model continual mutual adaptation. Instead, we focus on the speaker's adaptation to the listener in a single turn, which is certainly a simplified setup. Furthermore, our plug-and-play approach still requires the training of simulators per listener type. However, as we keep the speaker and listener models frozen and use the output obtained from them to train the simulators, this allows us to reduce the required amounts of training. We train the models from scratch using PhotoBook data and do not make use of state-of-the-art large pretrained vision-and-language models that are nowadays commonly based on Transformers, which could be considered a limitation. We opted for this setup as it is more aligned with our research questions, allowing us to control the domain-specificity of the models. We also acknowledge the imbalance in the set sizes of the domains, as well as the possible lexical and visual overlaps in the samples across domains. The overlaps may facilitate the adaptation of certain sentences from one domain to another (asymmetry is not controlled in a fine-grained manner), and this is not uncommon in human communication.

Ethics Statement

We are using neither large pretrained language models that have been found to be prone to bias issues nor uncurated data scraped from the internet that would open up myriads of problems. Still, there could be some bias in the PhotoBook data that should be investigated: players might have used offensive or undesirable language in describing images. Therefore, deploying these speakers and listeners directly is not advisable. Our research focusing on the adaptation of a speaker to their audience is done with the aim of improving communicative success within scenarios with knowledge asymmetry following human capabilities of self-monitoring and Theory of Mind. It is possible that adaptation to a specific listener could exacerbate possible biases if the training set of a given listener happens to include more bias. However, the reverse is also the case, where adaptation to underrepresented user groups could be beneficial.

Acknowledgements

We would like to thank Arabella Sinclair for her valuable input regarding the conception of this project, Gabriele Gennaro for his contribution to early implementations of the models, and the members of the Dialogue Modelling Group for their feedback. This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455). Nicolo' Brandizzi carried out the research while visiting the Dialogue Modelling Group in Amsterdam. His contribution was partially supported by the PNRR MUR project PE0000013-FAIR.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

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*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Yuwei Bao, Sayan Ghosh, and Joyce Chai. 2022. Learning to mediate disparities towards pragmatic communication. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2829–2842, Dublin, Ireland. Association for Computational Linguistics.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. PADA: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, 10:414–433.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Herbert H Clark. 1985. Language use and language users. *Handbook of social psychology (3rd ed.)*, pages 179–231.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Rodolfo Corona Rodriguez, Stephan Alaniz, and Zeynep Akata. 2019. Modeling Conceptual Understanding in Image Reference Games. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations*.

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

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Daniel Fried, Justin Chiu, and Dan Klein. 2021. Reference-centric models for grounded collaborative dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2130–2147, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Noah D Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Claudio Greco, Diksha Bagade, Dieu-Thu Le, and Raffaella Bernardi. 2023. She adapts to her student: An expert pragmatic speaker tailoring her referring expressions to the layman listener. *Frontiers in Artificial Intelligence*, 6.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910.

Robert Hawkins, Minae Kwon, Dorsa Sadigh, and Noah Goodman. 2020. Continual Adaptation for Efficient Machine Communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 408–419, Online. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2022. State-of-the-art generalisation research in NLP: A taxonomy and review. *CoRR*.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Ellen A Isaacs and Herbert H Clark. 1987. References in conversation between experts and novices. *Journal of experimental psychology: general*, 116(1):26.

Srinivasan Janarthanam and Oliver Lemon. 2010. Learning to adapt to unknown users: referring expression generation in spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 69–78.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Boaz Keysar. 2007. Communication and miscommunication: The role of egocentric processes.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Angeliki Lazaridou, Anna Potapenko, and Olivier 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*, pages 7663–7674, Online. Association for Computational Linguistics.

Willem J.M. Levelt. 1993. *Speaking: From intention to articulation*. MIT press.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European conference on computer vision*, pages 740–755. Springer.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Will Monroe, Robert X.D. 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.

Atsumoto Ohashi and Ryuichiro Higashinaka. 2022. Adaptive natural language generation for task-oriented dialogue via reinforcement learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 242–252, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Martin J Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018a. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew M. Botvinick. 2018b. Machine theory of mind. *CoRR*, abs/1802.07740.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Ardi Roelofs. 2020. Self-monitoring in speaking: In defense of a comprehension-based account. *Journal of Cognition*, 3(1).

Matthew Saxton. 2009. The inevitability of child directed speech. In *Language acquisition*, pages 62–86. Springer.

Aaditya K Singh, David Ding, Andrew Saxe, Felix Hill, and Andrew Lampinen. 2023. Know your audience: specializing grounded language models with listener subtraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3884–3911, Dubrovnik, Croatia. Association for Computational Linguistics.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3521–3529.

Sina Zarrieß and David Schlangen. 2019. Know what you don't know: Modeling a pragmatic speaker that refers to objects of unknown categories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 654–659, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind. In *International Conference on Machine Learning*, pages 12901–12911. PMLR.

Appendix

A Training Details

We provide the details of the setups of the generative language model in §A.1, the discriminators in §A.2, the simulators in §A.3 and the adaptation mechanism in §A.4. We use Python version 3.9.0 and PyTorch version 1.11.0 in the development and testing of all our models. In Table 4, we report the hyperparameters used in the training of our final models.

	LM	Disc	Simulator
Learning rate	0.0001	0.0001	0.0004
Batch size	3	64	32
Dropout	0.3	0.2	0
Attention dim	512	512	1024
Embed dim	1024	768	1024
Hidden dim	512	512	1024
Patience	30	30	5

Table 4: Hyperparameters used for training the generative language model, discriminator and simulator models.

A.1 Generative Language Model

In addition to the main hyperparameters listed in Table 4, the language model requires several additional parameters. In nucleus sampling, we set the p value for *top-p* to 0.9 and sample from a vocabulary that consists of the words in the training splits of all 5 domains. The maximum length of the generated utterances is set to 30. The model is initialized and trained with 4 different seeds, which yield similar performances. We use an early stopping patience of 30 epochs based on the validation set scores.[10]

Regarding the architectural details of the visually-conditioned language model, in the visual encoder we feed both the standardized target image vector and the concatenation of the six images in the full visual context into a linear layer followed by the ReLU non-linearity. We then concatenate the ensuing representations of the target image with the visual context and once more apply a linear layer followed by a ReLU non-linearity to obtain the final visual context, $\mathbf{v}$. This visual context is used to initialize a bidirectional LSTM encoder that takes as input the previous utterance referring to the target image in the current dialogue, if exists (see footnote 6), otherwise a special token indicating the absence of such an utterance. The final forward and backward hidden states of this encoder are concatenated, go through a linear layer and $tanh$ non-linearity. The output is then set as the initial hidden state h_0 of the LSTM decoder (Hochreiter and Schmidhuber, 1997).

[10]We use the 'nlg-eval' library at https://github.com/Maluuba/nlg-eval to obtain scores for the common NLG metrics and also use BERTScore version 0.3.11 provided at https://github.com/Tiiiger/bert_score.

A.2 Discriminators

In these models instantiating the listeners, the word embeddings go through a dropout layer and a linear layer followed by the Leaky-ReLU non-linearity, after which standardization is applied. The visual context is processed in the same way as in the generative language model. Each word representation is concatenated with the representation of the visual context. The resulting vectors go through a linear layer and ReLU. Finally, we apply attention over these vectors to obtain the attention-weighted multimodal context vector. It is this context vector that is compared to the representations of candidate images via dot product.

We use the same set of hyperparameters for each domain as shown in Table 4. The domain-specific listener models were selected based on their accuracy on the in-domain validation set. We report accuracy and MRR on the in-domain and out-of-domain test sets in Table 6.

OOD word masking Our listeners are initialized with the same vocabulary comprising all the words in the training data. However, the domain-specific listeners only learn the words that exist in their own training sets. Therefore, if the speaker generates an OOD word for a domain-specific listener, in order not to further confound the effects of adaptation on the listeners, we mask the word with the <unk> vector. This vector is the same across all domains.

A.3 Simulator

We select the simulator models based on their accuracy in predicting the behaviour of the listener models on the validation set. The simulator models are trained using the AdamW optimizer (Loshchilov and Hutter, 2017) with a weight decay of 0.0001, and a plateau learning scheduler with a patience of 2, a factor of 0.5, a threshold of 0.5.

A.4 Adaptation Mechanism

We optimize the values of the number of adaptation steps and the learning rate for the adaptation mechanism. We perform 2 hyperparameter sweeps using the Weight & Biases (WandB) platform (Biewald, 2020), evaluating a range of values. We find a positive correlation between both hyperparameters and adaptation accuracy, with Pearson's correlation coefficients of 0.71 for the learning rate, and 0.66 for the number of steps.

B Additional Results

Here, we provide additional results yielded by our models for the speaker in §B.1, the listener in §B.2, the simulator in §B.3 and for the adaptation mechanism in §B.4.

B.1 Speaker Results

We provide the detailed results of the speaker model on the test set in Table 5 with the averages and standard deviations over 4 runs.

B.2 Listener Results

Table 6 reports the domain-specific listener performances on IND and OOD gold data. We observe that the domain-specific listeners perform well in in-domain settings and perform close to the random baseline in OOD settings.

Table 7 presents the domain-specific listener accuracies on speaker-generated input. Especially in IND settings, we see lower scores as compared to use of the gold data, presumably because the listener models were trained on gold data.

B.3 Simulator Results

The detailed outcomes of the simulator models are reported in Table 8. Here, we also report the results for the subset where the *listener* made a correct prediction (Pos) vs. it made an incorrect prediction (Neg). The simulators are better able to capture the correct listener behaviour, possibly because during the training of simulators, in-domain data provides a clear picture of listener's correct behaviour.

B.4 Adaptation Results

In Table 9, we provide the test set results of the adaptation pipeline, broken down into domains and for IND and OOD inputs separately. The outcomes show that adaptation has effects in both IND and OOD settings, increasing resolution accuracies over speaker-generated utterances.

C Evaluation Cards

For each of the three main modules in our experiments, we provide an evaluation card to clarify the nature of our generalisation tests.[11] See Table 10 for the generator, Table 11 for the simulator, and Table 12 for the listener. We also register our work in the GenBench evolving survey of generalisation in NLP (Hupkes et al., 2022).[12]

D Additional Experiments

Here, we provide details on additional experiments we performed in our adaptation pipeline.

In our adaptation mechanism, one of the stopping conditions is that the simulator predicts that the listener will be able to guess the referent. We also explored continuing adaptation until the *listener* itself correctly guesses the referent. We report the results in Table 13, which reveal that using this stopping condition would yield higher results since the utterances are adapted until the actual listener makes a correct guess, mimicking an online interaction setup.

E Additional Analyses

We note that we measure the type-utterance ratio for each step (i.e., the vocabulary size divided by the number of utterances available for that step), rather than the vocabulary size, because different steps correspond to different numbers of utterances: adaptation stops when the simulator module predicts the target image.

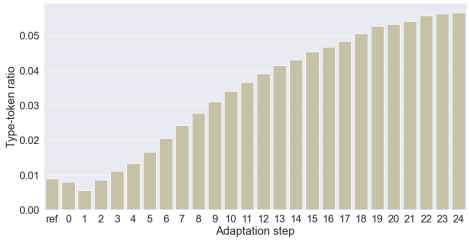

Figure 6: Type-token ratio across adaptation steps. Human gold utterances (*ref*) and non-adapted utterances (0) also shown.

Figure 7 shows unigram part-of-speech distribution across adaptation steps for the in-domain and out-of-domain conditions.

We also measure the domain-specificity of utterances over steps, both in terms of the target image domain and of the listener domain, as the per-

[11] https://genbench.org/eval_cards
[12] https://genbench.org/references

BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge
40.06 ± 1.60	23.81 ± 1.51	14.09 ± 1.20	8.46 ± 0.89	32.92 ± 0.93
CIDEr	**BertScore - Recall**	**BertScore-Recall - F1**	**BertScore - Precision**	
44.07 ± 1.68	58.91 ± 0.19	57.7 ± 0.12	57.9 ± 0.16	

Table 5: Speaker results on the test set as measured by common natural language generation evaluation metrics.

Domain	Epoch	IND		OOD	
		Accuracy	MRR	Accuracy	MRR
Appliances	23	84.12 ± 0.33	90.27 ± 0.10	20.28 ± 0.23	44.07 ± 0.11
Food	21	85.40 ± 0.28	91.20 ± 0.20	17.72 ± 0.18	42.42 ± 0.06
Indoor	14	82.94 ± 0.13	89.32 ± 0.09	19.14 ± 0.09	43.46 ± 0.06
Outdoor	19	83.96 ± 0.23	90.01 ± 0.14	19.64 ± 0.07	43.52 ± 0.06
Vehicles	17	78.99 ± 0.35	86.81 ± 0.14	18.46 ± 0.28	42.36 ± 0.20

Table 6: Listener performance on gold utterances. Accuracy and MRR for the in-domain (IND) and out-of-domain (OOD) samples given to listeners trained on specific domains (indicated under the 'Domain' column).

Listener domain	Data domain				
	appliances	food	indoor	outdoor	vehicles
appliances	57.61 ± 1.38	20.10 ± 0.63	19.92 ± 0.47	21.27 ± 0.83	15.98 ± 0.82
food	19.11 ± 1.70	54.29 ± 1.06	18.60 ± 0.84	18.85 ± 0.49	18.85 ± 0.49
indoor	22.71 ± 1.30	19.65 ± 1.77	53.62 ± 0.79	20.82 ± 1.05	16.77 ± 0.79
outdoor	15.08 ± 1.04	21.46 ± 0.70	19.62 ± 0.69	52.93 ± 1.11	17.69 ± 0.97
vehicles	16.36 ± 1.55	16.17 ± 0.81	17.41 ± 0.64	20.13 ± 0.59	43.08 ± 1.16

Table 7: Listener accuracies on speaker-generated data. Each row indicates the domain a listener was trained on and the columns indicate the domain of the input samples. Results over 5 seeds.

centage of domain-specific words in an utterance. We consider as domain-specific the words that appear *only* in interactions about a certain domain. The speaker, throughout adaptation, produces more words belonging to both the image and the listener domain (Figure 9) and thus less domain-agnostic words. We saw that, over adaptation steps, the decoder hidden state forgets image domain information in favour of the listener domain. This does not translate into no longer producing words from the image domain, suggesting that the speaker may be focusing more on the specific image than on its semantic domain.

Figure 8 shows mean utterance age of acquisition rating (Kuperman et al., 2012) over steps.

Simulator	Setting	Avg	Pos	Neg
all domains	–	69.97 ± 0.79	85.15 ± 1.39	54.73 ± 0.76
domain specific	IND	78.20 ± 1.26	88.09 ± 1.98	67.36 ± 2.96
	OOD	72.78 ± 0.56	73.67 ± 1.69	72.58 ± 0.71

Table 8: Simulator's accuracy in predicting the behaviour of a listener knowledgeable about *all domains* (as the speaker) and a listener with *domain-specific* knowledge for IND and OOD samples. 'Avg' is the overall accuracy, 'Pos' and 'Neg' are the percentages of correct predictions for the samples where the listener picked the correct (Pos) and the incorrect image (Neg).

	OOD			IND		
	Golden	**Speaker**	**Adapted**	**Golden**	**Speaker**	**Adapted**
appliances	20.21	19.30	27.74	84.21	57.21	74.28
indoor	18.50	19.53	28.34	83.22	52.94	69.62
food	17.06	18.31	26.26	85.61	55.54	78.15
outdoor	18.89	18.54	26.21	84.38	52.83	73.04
vehicles	18.25	17.35	25.16	78.67	42.09	63.75

Table 9: Test results for the audience-aware adaptation pipeline, 5 seeds for each domain.

Motivation			
Practical □ △ ◯	*Cognitive* □ △ ◯	*Intrinsic*	*Fairness*
Generalisation type			
Compositional	*Structural*	*Cross Task*	*Cross Language*	*Cross Domain* □ △ ◯	*Robustness*
Shift type			
Covariate	*Label* □ △ ◯	*Full*	*No shift*
Shift source			
Naturally occurring	*Partitioned natural*	*Generated shift*	*Fully generated* □ △ ◯
Shift locus			
Train–test	*Finetune train–test*	*Pretrain–train*	*Pretrain–test* □ △ ◯

Table 10: **Generator**'s evaluation card for the three main setups: baseline □, self-aware adaptation △, and audience-aware adaptation ◯

Motivation			
Practical △ ◯	*Cognitive* △ ◯	*Intrinsic*	*Fairness*
Generalisation type			
Compositional	*Structural*	*Cross Task*	*Cross Language*	*Cross Domain* △ ◯	*Robustness*
Shift type			
Covariate	*Label* △	*Full*	*No shift* ◯
Shift source			
Naturally occurring	*Partitioned natural*	*Generated shift*	*Fully generated* △ ◯
Shift locus			
Train–test	*Finetune train–test*	*Pretrain–train*	*Pretrain–test* △ ◯

Table 11: **Simulator**'s evaluation card for the two setups in which it is used (i.e., baseline setup excluded): self-aware adaptation △, and audience-aware adaptation ◯.

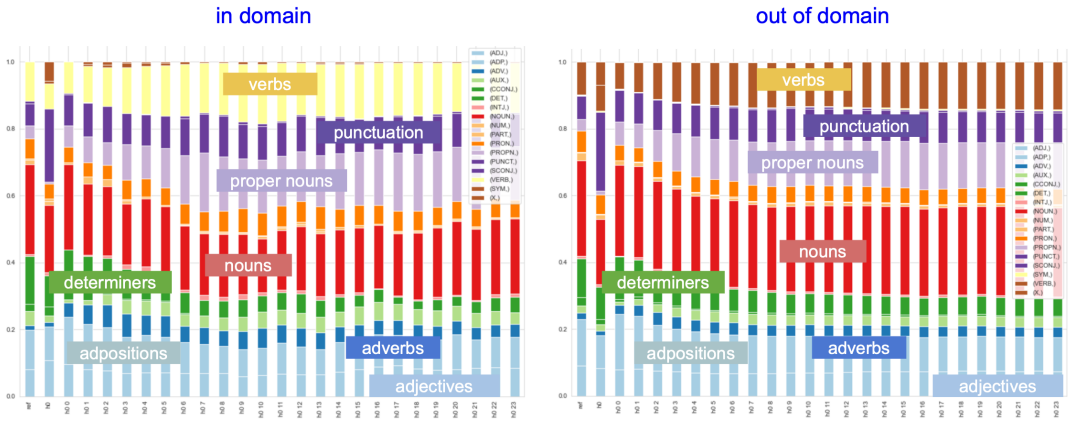

Figure 7: Unigram POS distribution across adaptation steps.

Motivation			
Practical □ △ ◯	*Cognitive* □ △ ◯	*Intrinsic*	*Fairness*
Generalisation type			
Compositional	*Structural*	*Cross Task*	*Cross Language*
Shift type			
Covariate □ △ ◯	*Label*	*Full*	*No shift* □ △ ◯
Shift source			
Naturally occurring	*Partitioned natural* □ △ ◯	*Generated shift*	*Fully generated* □ △ ◯
Shift locus			
Train–test	*Finetune train–test*	*Pretrain–train*	*Pretrain–test* □ △ ◯

Table 12: **Listener**'s evaluation card for the three main setups: baseline □, self-aware adaptation △, and audience-aware adaptation ◯. In out-of-domain settings (OOD), the type of shift is covariate. In in-domain settings (IND), there is no shift between the training and test.

Target domain	Golden	Speaker	Adapted
appliances	16.85	20.04	**38.89**
food	85.57	55.26	**91.74**
indoor	18.69	18.47	**39.49**
outdoor	19.03	18.33	**37.96**
vehicles	13.75	16.63	**35.43**

Table 13: Listener accuracy using the listener stopping condition in the adaptation mechanism.

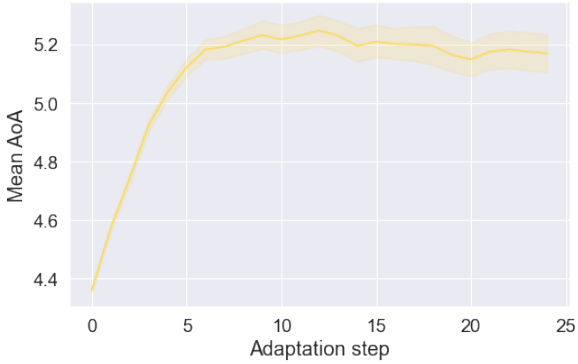

Figure 8: Mean utterance Age of Acquisition over adaptation steps. Step 0 corresponds to the non-adapted utterance.

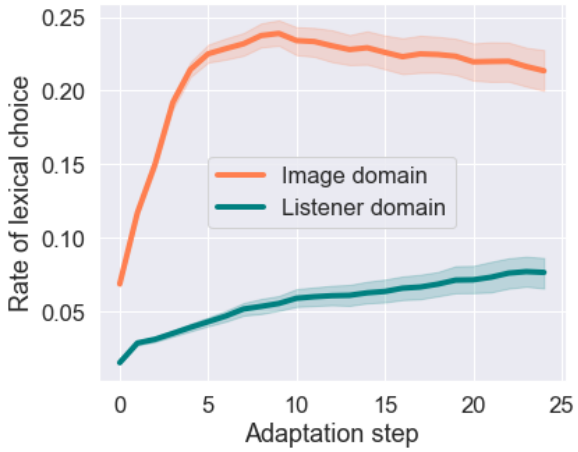

Figure 9: Rate of lexical choice from image and listener domain-specific vocabularies.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We discuss them in the Limitations section.

☑ A2. Did you discuss any potential risks of your work?
We discuss them in our Ethics Statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
We summarize them both in the abstract and the introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We trained models as described in Section 5 and 6 and generated outputs from them.

☑ B1. Did you cite the creators of artifacts you used?
We provide citations for the dataset in Section 4 and the models we build on in Section 5.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We used publicly available data.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The dataset we used did not contain information that could de-anonymize the participants.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Reported in Section 4.

C ☑ Did you run computational experiments?

We describe the computational experiments in Sections 5 and 6 and more details in Appendix.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Experimental setup is discussed in the main paper sections 3,5 and 6. Hyperparameter search and the selected values are described in the Appendix.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We report multiple runs for each model type, and report the mean and standard deviations of these runs.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We report the the main libraries we use and their versions in the Appendix.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.