

# Don't Say *That*!

## Making Inconsistent Dialogue Unlikely with Unlikelihood Training

Margaret Li<sup>1</sup>, Stephen Roller<sup>1</sup>, Ilia Kulikov<sup>2\*</sup>, Sean Welleck<sup>2\*</sup>

Y-Lan Boureau<sup>1</sup>, Kyunghyun Cho<sup>1,2</sup>, Jason Weston<sup>1,2</sup>

<sup>1</sup>Facebook AI Research

<sup>2</sup>New York University

{margaretli, roller, ylan, kyunghyuncho, jase}@fb.com

wellecks@nyu.edu, kulikov@cs.nyu.edu

### Abstract

Generative dialogue models currently suffer from a number of problems which standard maximum likelihood training does not address. They tend to produce generations that (i) rely too much on copying from the context, (ii) contain repetitions within utterances, (iii) overuse frequent words, and (iv) at a deeper level, contain logical flaws. In this work we show how all of these problems can be addressed by extending the recently introduced unlikelihood loss (Welleck et al., 2019a) to these cases. We show that appropriate loss functions which regularize generated outputs to match human distributions are effective for the first three issues. For the last important general issue, we show applying unlikelihood to collected data of *what a model should not do* is effective for improving logical consistency, potentially paving the way to generative models with greater reasoning ability. We demonstrate the efficacy of our approach across several dialogue tasks.

## 1 Introduction

Open-ended tasks such as dialogue reveal a number of issues with current neural text generation methods. In more strongly grounded tasks such as machine translation and image captioning, current encoder-decoder architectures provide strong performance, where mostly word-level decisions are often taken correctly by the model. However, critical failings are exposed in less constrained generation: reliance on repetitive copying and overuse of frequent words, and an inability to maintain logical coherence. The former shows the learning objective is faulty in that it cannot match simple statistics of the training data, while the latter touches more to the heart of artificial intelligence:

Sentence:

I love basketball. It's  
awesome. I really dislike

Completions:

8.3% **basketball**  
7.7% **it**  
6.5% **the**  
4.0% **sports**

Figure 1: GPT-2 345M model completions can show lack of coherence, e.g. direct contradictions.

these models do not understand what they are saying. For example, Figure 1 shows how the 345M-parameter GPT2 model (Radford et al., 2019) can give high probability to contradictory generations.

In this work, we show how the recently introduced unlikelihood objective (Welleck et al., 2019a) can be generalized to remedy these problems. Unlikelihood is a technique developed for removal of repetition in language model completions, and works by adding an extra term to the objective that forces repetitions to have low probability, alleviating the degenerative problems highlighted in Holtzman et al. (2019). In fact, unlikelihood can be seen as a much more general framework, as we will see.

We first generalize unlikelihood to a different domain: dialogue, where we measure statistics of the training distribution in terms of contextual copies, within-utterance repeats, and vocabulary usage. We then develop loss functions that control these statistics, providing improved metrics on several tasks. Secondly, we show how the same tools can be used to address deeper semantic issues in such models. By leveraging existing natural language inference (NLI) data (Welleck et al., 2019b) as supervision against poor quality generations, we train models that assign low probability to generating incoherent and contradictory text. Overall, our approach yields more consistent dialogue models across several axes, and provides a

\*Work done while at Facebook AI Research (FAIR).

promising framework for further advances.

Code and pre-trained models will be made available.<sup>†</sup>

## 2 Dialogue Unlikelihood Training

**Dialogue Generation** Dialogue generation consists in predicting an utterance  $\mathbf{y} = (y_1, \dots, y_{|\mathbf{y}|})$  given a context  $\mathbf{x} = \{s_1, \dots, s_k, u_1, \dots, u_t\}$  that consists of initial context sentences  $s_{1:k}$  (e.g., scenario, knowledge, personas, etc.) followed by dialogue history utterances  $u_{1:t}$  from speakers who take consecutive turns.

**Likelihood Training** Given a dataset  $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}$  derived from a collection of human-human interactions, the standard approach to generative training for dialogue tasks is maximum likelihood estimation (MLE), that minimizes:

$$\mathcal{L}_{\text{MLE}}^{(i)}(p_\theta, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = - \sum_{t=1}^{|\mathbf{y}^{(i)}|} \log p_\theta(y_t^{(i)} | \mathbf{x}^{(i)}, y_{<t}^{(i)}),$$

where  $\mathbf{x}^{(i)}$  is a gold context (dialogue history and initial context sentences) and  $\mathbf{y}^{(i)}$  is a gold next-utterance, and  $y_t^{(i)}$  is the  $t$ -th token of  $\mathbf{y}^{(i)}$ .

Likelihood-based (greedy or beam) decoding applied after training a model with this objective yields sequences with statistics that do not match the original human training sequence distribution.

**Unlikelihood Training** To control for such distribution mismatches, we employ the unlikelihood loss (Welleck et al., 2019a), generalizing it to our setting, and developing a particular form of the loss function for each type of mismatch.

The general form of the unlikelihood loss penalizes a set of tokens  $\mathcal{C}_t$  at each time-step,  $\mathcal{L}_{\text{UL}}^{(i)}(p_\theta, \mathcal{C}_{1:T}, \mathbf{x}, \mathbf{y}) =$

$$- \sum_{t=1}^{|\mathbf{y}|} \sum_{y_c \in \mathcal{C}_t} \beta(y_c) \log(1 - p_\theta(y_c | \mathbf{x}, y_{<t})),$$

where  $\mathcal{C}_t \subseteq \mathcal{V}$  is a subset of the vocabulary, and  $\beta(y_c)$  is a candidate-dependent scale that controls how much the candidate token should be penalized. The overall objective in unlikelihood training then consists of mixing the likelihood and unlikelihood losses,

$$\mathcal{L}_{\text{ULE}}^{(i)} = \mathcal{L}_{\text{MLE}}^{(i)} + \alpha \mathcal{L}_{\text{UL}}^{(i)}, \quad (1)$$

<sup>†</sup>[https://parl.ai/projects/dialogue\\_unlikelihood/](https://parl.ai/projects/dialogue_unlikelihood/)

where  $\alpha \in \mathbb{R}$  is the mixing hyper-parameter.

Likelihood tries to model the overall sequence probability distribution, while unlikelihood corrects for known biases. It does this via the set of *negative candidates*  $\mathcal{C}_t$  calculated at each step  $t$ , where we are free to select candidate generation functions depending on the biases to be mitigated. Likelihood pushes *up* the probability of a gold token  $y_t^{(i)}$  while unlikelihood pushes *down* the probability of negative candidate tokens  $y_c \in \mathcal{C}_t$ .

In Welleck et al. (2019a) the context  $\mathbf{x}$  consists of a ground-truth sequence ( $\mathbf{x} = \mathbf{x}^{(i)}$ ), the target  $\mathbf{y}$  is either a ground-truth sequence ( $\mathbf{y} = \mathbf{y}^{(i)}$ ) or a model-generated sequence ( $\mathbf{y} = \hat{\mathbf{y}}$ ), and the per-token scale parameter  $\beta(y_c)$  is 1.

In this paper, we demonstrate how unlikelihood can be used as a general framework by applying it to the dialogue domain. We show how varying the contexts  $\mathbf{x}$ , targets  $\mathbf{y}$ , candidates  $\mathcal{C}$  and scaling  $\beta$  can be used to improve the coherence and language modeling quality of dialogue models. To do this, we now consider the different biases we wish to mitigate, and construct a specific unlikelihood loss for each in turn.

### 2.1 Repetition and Copying

Generative dialogue models are known to both (i) rely too much on copying existing context knowledge or dialogue history; and (ii) repeat themselves within individual utterances. To address this with unlikelihood, we define two types of negative candidate tokens which either appear in a repeating n-gram from the context or from the generated label itself,

$$\mathcal{C}_t^{\text{context-copy}} = \begin{cases} \{y_t\} & y_t \in \text{repeat context n-gram} \\ \emptyset & \text{otherwise,} \end{cases}$$

$$\mathcal{C}_t^{\text{label-repeat}} = \begin{cases} \{y_t\} & y_t \in \text{repeating label n-gram} \\ \emptyset & \text{otherwise,} \end{cases}$$

where  $y_t$  is a token in a repeating context n-gram when  $y_t$  is part of an n-gram that already appeared in the context tokens  $\mathbf{x}$ , and is in a repeating label n-gram when  $y_t$  is part of an n-gram that already appeared in  $y_{<t}$ . Given a ground-truth context  $\mathbf{x}^{(i)}$ , we apply these two forms of unlikelihood to a model-generated sequence  $\hat{\mathbf{y}}^{(i)}$ . In summary, we either apply the per-example loss

$$\mathcal{L}_{\text{UL}}^{(i)}(p_\theta, \mathcal{C}_{1:|\mathbf{y}|}^{\text{context-copy}}, \mathbf{x}^{(i)}, \hat{\mathbf{y}}^{(i)})$$

for controlling context copies, or

$$\mathcal{L}_{\text{UL}}^{(i)}(p_{\theta}, \mathcal{C}_{1:|y|}^{\text{label-repeat}}, \mathbf{x}^{(i)}, \hat{\mathbf{y}}^{(i)}).$$

for controlling label repeats. We also consider mixing the two losses to mitigate both issues.

## 2.2 Vocabulary Usage

Neural sequence models trained with maximum likelihood generate sequences with token distributions that differ from those of human text (Dinan et al., 2020; Holtzman et al., 2019). In particular, these models tend to produce high frequency tokens too often and low frequency tokens too rarely, where frequency is defined by the human token distribution.

We address this with unlikelihood by penalizing tokens according to the mismatch between the model and ground-truth unigram distributions. Specifically, we first maintain an empirical estimate of the model’s unigram distribution  $p_{\text{model}}(y_t)$  and the human distribution  $p_*(y_t)$ :

$$p_{\text{model}}(y_t) = \frac{\text{count}(y_t)}{|Y|},$$

where  $Y$  is a collection of token predictions on a subset of training data  $\mathcal{D}'$  (e.g. the preceding  $k = 256$  batches), and  $\text{count}(y_t)$  is the number of occurrences of  $y_t$  in  $Y$ . This is computed using model sequences ( $\mathbf{y} = \hat{\mathbf{y}}$ ), defining  $Y$  as the collection of all tokens in all  $\hat{\mathbf{y}}$ .

We wish to *push down* the probability of tokens appearing too often, i.e. when  $p_{\text{model}}(y_t) > p_*(y_t)$ . For the unlikelihood loss, each step’s candidate is thus the current token,  $\mathcal{C}_t^{\text{identity}} = \{y_t\}$ , and each token’s unlikelihood loss is scaled according to the mismatch between the approximated model and human distributions,

$$\beta(y_c) = p_{\text{model}}(y_c) \log \left( \frac{p_{\text{model}}(y_c)}{p_*(y_c)} \right).$$

The unlikelihood loss for a token  $y_c$  is non-zero when the token occurs more often in the model’s estimated unigram distribution. In summary, the resulting per-example loss is

$$\mathcal{L}_{\text{UL}}^{(i)}(p_{\theta}, \mathcal{C}_{1:|y|}^{\text{identity}}, \mathbf{x}^{(i)}, \mathbf{y})$$

where  $\mathbf{y}$  is a model-generated sequence.

## 2.3 Contradictions

Neural generation models appear fluent, especially when pre-trained on large datasets, but are still poor at understanding the language they produce. That is, they can produce logically or factually inaccurate, or contradicting statements (Welleck et al., 2019b; Zhang et al., 2018; Hayashi et al., 2019; Petroni et al., 2019). Here, we show how the unlikelihood objective can be used to train such models to assign low probability to inconsistent and contradictory utterances.

To do so, we assume the existence of training data of both positive *and* negative examples of coherent behavior. There is a raft of recent large-scale, high quality data that can be massaged into this form, from natural language inference (NLI) tasks (Bowman et al., 2015; Williams et al., 2018; Welleck et al., 2019b) to commonsense reasoning tasks (Zellers et al., 2019; Qin et al., 2019). Two collections of data can be derived from the labels of such a supervised task:

$$\mathcal{D}^+ = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)+})\}, \quad \mathcal{D}^- = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)-})\},$$

where  $\mathcal{D}^+$  is coherent behavior, e.g. neutral or entailing data in NLI, and  $\mathcal{D}^-$  is incoherent behavior, e.g. contradictions. In general, many forms of this type of data can be collected, not just NLI, and it is also not necessary for the contexts  $\mathbf{x}^{(i)}$  to overlap as we have written here.

Standard likelihood training can then be performed on coherent data  $\mathcal{D}^+$ , while the unlikelihood objective is applied to  $\mathcal{D}^-$  as we wish to *push down* the probability of generating the incoherent response  $\mathbf{y}^-$  given a context  $\mathbf{x}$ . That is, given an incoherent pair  $(\mathbf{x}, \mathbf{y}^-)$  we use the loss

$$\mathcal{L}_{\text{UL}}(p_{\theta}, \mathcal{C}_{1:|y|}^{\text{identity}}, \mathbf{x}, \mathbf{y}^-),$$

where we penalize each token in the target ( $\mathcal{C}_t^{\text{identity}} = \{y_t^-\}$ ). Hence, the loss makes generating the contradicting sentences less likely.

## 3 Related Work

Our work provides new applications of unlikelihood training (Welleck et al., 2019a), showing that unlikelihood offers a general framework for improving generative models, and in particular dialogue models. Outside of that work, the use of negative training in dialogue retrieval, rather than generation, has been previously extensively studied, see e.g. (Humeau et al., 2019; Nugmanova

et al., 2019). In the area of generative dialogue, a number of works have focused on improving the standard likelihood training approach. Closer to our work is that of He and Glass (2019) which developed the approach of negative training to prevent generic and malicious responses in dialogue models. In terms of improving repetition and specificity, a recent alternative approach is that of control (Fan et al., 2018; Fidler and Goldberg, 2017; Ghazvininejad et al., 2017; See et al., 2019). Nucleus sampling (Holtzman et al., 2019) can help to remove generic or repetitive utterances at the expense of accuracy, but was shown to be inferior to beam blocking, which in turn was shown to be inferior to unlikelihood in Welleck et al. (2019a).

In terms of dialogue coherence, Welleck et al. (2019b) showed that retrieval, but not generative models, could be improved with NLI as a rescorer, while Yang et al. (2018) multi-tasked with NLI. The work of Gabriel et al. (2019) has also studied improving narrative flow with a discriminative rescorer, but in that case for generated language. In our work, the improvements are tightly integrated into the training of the model itself.

## 4 Experiments

In all of our experiments we employ a large pre-trained seq2seq Transformer (Vaswani et al., 2017) as our base model, which we then fine-tune for particular tasks with the objectives outlined in Section 2 and specified in each experiment below. Following previous work (Humeau et al., 2019), we pre-train our model on dialogue data, using a previously existing Reddit dataset extracted and obtained by a third party and made available on pushshift.io, training to generate a comment conditioned on the full thread leading up to the comment, spanning  $\sim 2200M$  training examples. Our Transformer model consists of an 8 layer encoder, 8 layer decoder with 512-dimensional embeddings and 16 attention heads, and is based on the ParlAI implementation of Miller et al. (2017). The model was trained with a batch size of 3072 sequences for approximately 3M updates using a learning rate of  $5e-4$ , and an inverse square root scheduler. This pre-training took approximately two weeks using 64 NVIDIA V100s.

### 4.1 Repetition and Copying

We use the ConvAI2 persona-based dialogue (Zhang et al., 2018), Wizard of Wikipedia

Model	PPL	F1	Repetition	
			Context	Label
Human	-	-	.0223	.0004
MLE Baseline	11.4	.199	.1131	.0210
UL (Context only)	11.8	.194	.0330	.0069
UL (Label only)	11.4	.203	.0984	.0005
UL (Context & Label)	11.9	.193	.0352	.0023

Table 1: Evaluation on the ConvAI2 task valid set (test set is hidden), comparing standard likelihood (MLE) with context and label repetition unlikelihood loss training. The repetition types can be decreased depending on which type of unlikelihood loss is used, with minimal changes in perplexity and F1.

Model	PPL	F1	Repetition	
			Context	Label
Human	-	-	.160	.001
MLE Baseline	8.3	.368	.441	.014
UL (Context only)	8.8	.346	.229	.037
UL (Label only)	8.3	.371	.426	.001
UL (Context + Label)	8.5	.358	.313	.009

Table 2: Evaluation on the Wizard of Wikipedia test set, comparing standard likelihood (MLE) with context and label repetition unlikelihood loss training. The repetition types can be decreased depending on the type of unlikelihood loss used, while minimally impacting F1.

knowledge-grounded dialogue (Dinan et al., 2019) and ELI5 long-form question answering (Fan et al., 2019) datasets to evaluate the effect of using unlikelihood to reduce copying and repetition in model generated utterances. On each dataset, we fine-tune the pre-trained pushshift.io Reddit model, then evaluate by generating next-utterances for dialogue contexts from the test set (or validation in ConvAI2, as the test set is hidden). We use greedy decoding in our main experiments for simplicity and scalability, but we also obtained similar results with beam search, shown in Appendix A.

To measure label repetition in a sequence  $y$ , we use the portion of duplicate n-grams:

$$1.0 - \frac{|\text{unique n-grams}(y)|}{|\text{n-grams}(y)|},$$

and report the metric averaged over the examples. Label repetition increases from zero as the model generates more repeated n-grams. To measure context repetition, we measure the fraction of gen-



Model	PPL	F1	Repetition	
			Context	Label
Human	-	-	.009	.010
MLE Baseline	21.0	.130	.033	.617
UL (Context only)	21.4	.163	.008	.322
UL (Label only)	21.4	.183	.015	.055
UL (Context + Label)	21.8	.184	.009	.078

Table 3: Evaluation on the ELI5 task test set, comparing standard likelihood (MLE) with context and label repetition unlikely loss training. The repetition types can be decreased depending on which type of unlikely loss is used, while improving F1.

erated n-grams that appear in the original context:

$$\frac{|n\text{-grams}(\mathbf{y}) \cap n\text{-grams}(\mathbf{x})|}{|n\text{-grams}(\mathbf{y})|},$$

and report the metric averaged over the examples. Context repetition increases when the model ‘copies’ n-grams from the context. To quantify language modeling quality, we use standard perplexity and F1 metrics.

We use the pre-trained model fine-tuned with MLE as the baseline, and compare it against the pre-trained model fine-tuned with copy and repetition unlikely loss (§2.1).

**Results** Results for ConvAI2 are shown in Table 1. We see that training unlikely loss using only-contexts or only-labels reduces their corresponding metrics dramatically compared to the MLE baseline. Training with both context- and label-repetition unlikely loss reduced both context repetitions (by 69%, .0352 vs. .1131) and label repetitions (by 89%, .0023 vs .0210) compared to the MLE baseline, much closer to human levels, while keeping perplexity essentially constant.

Comparatively, the Wizard of Wikipedia MLE baseline experiences a much larger problem with context repetition, due to its tendency to copy grounded knowledge verbatim (Table 2).

Results for ELI5, shown in Table 3, show that it has an especially large problem with label repetition, and that label-unlikely loss is able to reduce the repetitions by 91% (.055 vs .617), while significantly boosting F1 (.130 to .182).

Figures 2 and 3 show perplexity as a function of label and context repeats respectively using unlikely loss on ELI5. The parameter  $\alpha$  can clearly control repeats smoothly, with only very high values resulting in increased perplexity.

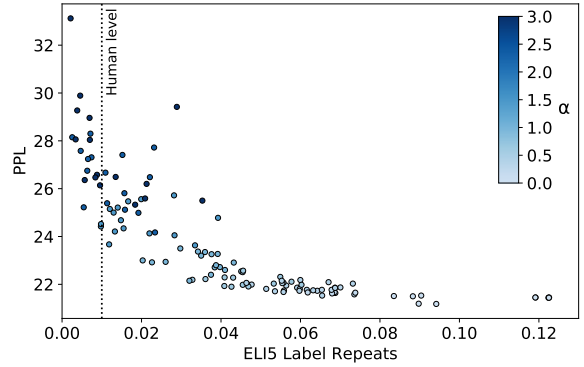


Figure 2: ELI5: Perplexity vs. label repeats as a function of  $\alpha$  in the label unlikely objective.

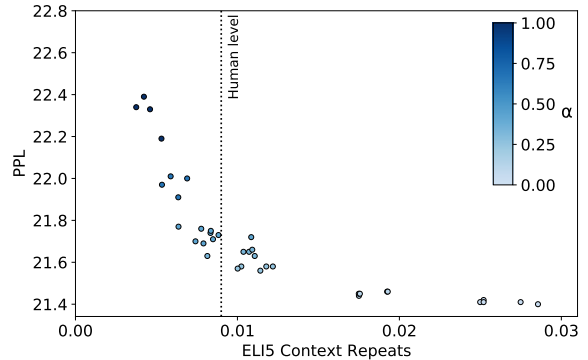


Figure 3: ELI5: Perplexity vs. context repeats as a function of  $\alpha$  in the context unlikely objective.

**Human Evaluation** Finally, we perform a human evaluation using the same pairwise evaluation scheme as (Fan et al., 2019) performed on ELI5, comparing the MLE baseline to UL (Label only) which asks: *Which response answers the question better?* The evaluators are asked to consider both the readability and accuracy of the answer. Results are given in Figure 4 (left), showing a statistically significant improvement over the baseline (150 trials, two tailed binomial test,  $p < 0.01$ ). Further details are given in Appendix C.

## 4.2 Vocabulary Usage

We evaluate the ability of vocabulary unlikely loss (§2.2) to reduce the mismatch between model and human token distributions.

We use the ConvAI2 dataset, where our baseline is again trained using maximum likelihood. Starting with the baseline model, we then fine-tune several models using vocab unlikely loss at logarithmically interpolated values of  $\alpha \in [1, 1000]$ .

We partition the vocabulary into ‘frequent’, ‘medium’, ‘rare’, and ‘rarest’ using the human

unigram distribution computed with the ConvAI2 training set, corresponding to the sorted token sets whose cumulative mass accounts for the top 40%, the next 30%, the next 20% and the final 10% of usage, respectively. We evaluate a model by generating utterances given contexts from the ConvAI2 validation set, and compute the fraction of tokens within each class.

**Results** Figure 5 shows how the vocabulary distribution obtained after unlikelihood training is affected by the choice of mixing hyperparameter  $\alpha$  (Eq. 1): it can smoothly transition between the human training distribution and the MLE trained distribution (‘Baseline’), which is far from the human one.

Table 4 compares the MLE baseline with unlikelihood with increasing  $\alpha$  values in terms of distribution and F1 score. The vocabulary unlikelihood fine-tuning shifts probability mass from the over-represented frequent words towards under-represented medium and rare words, with the effect strengthening as  $\alpha$  increases. At a small cost to perplexity and F1, the unlikelihood tuning reduced the overuse of common tokens by 9 points, matching the human rate, while improving the production of rare tokens by 3 percentage points.

**Human Evaluation** Finally, we perform a human evaluation using the ACUTE-EVAL framework (Li et al., 2019), comparing the MLE baseline to UL for various  $\alpha$ . First, 252 human-bot conversations (8 turns each) are collected, and then models are compared pairwise by asking the question: *Who would you prefer to talk to for a long conversation?* For these experiments we compare with both methods generating using beam with context blocking of trigrams. Results are given in Figure 4 (right), showing a statistically significant improvement over the baseline according to humans (two tailed binomial test,  $p < 0.01$ ). Further details are given in Appendix C.

### 4.3 Contradictions

We use the dialogue natural language inference (NLI) task of Welleck et al. (2019b) to obtain labeled non-contradicting and contradicting dialogue sentence pairs to use in unlikelihood training (§2.3). Dialogue NLI contains utterances labeled as entailing (E), neutral (N) or contradiction (C), given a premise that is either a persona sentence (an initial context sentence describing a dialogue agent’s personality) or another dialogue utterance

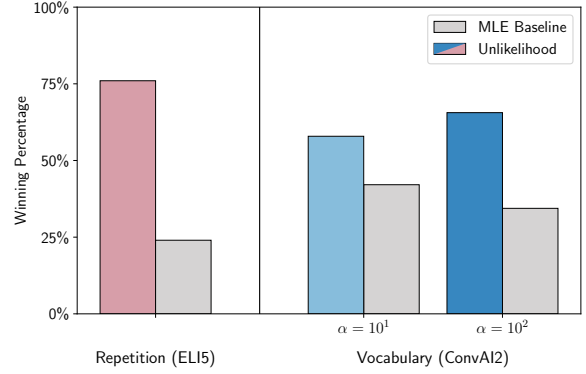


Figure 4: Human evaluation experiments for label unlikelihood on ELI5 (left), and vocabulary unlikelihood on ConvAI2 for two values of  $\alpha$  (right). Unlikelihood significantly outperforms the MLE baselines.

Model	PPL	F1	Token frequency classes			
			Freq	Med	Rare	Rarest
Human	-	-	.400	.300	.200	.100
MLE Baseline	11.4	.199	.491	.282	.157	.068
UL, $\alpha = 10^0$	11.4	.200	.483	.289	.163	.063
UL, $\alpha = 10^1$	11.9	.201	.459	.328	.154	.058
UL, $\alpha = 10^2$	12.5	.190	.430	.335	.163	.071
UL, $\alpha = 10^3$	14.4	.174	.399	.339	.188	.073

Table 4: Unlikelihood loss applied to vocabulary distributions. Stronger  $\alpha$  terms greatly shift probability mass from the most Frequent words to Medium and Rare words, at a small cost to PPL and F1. Frequent, medium, rare and rarest token classes are defined as the sets of tokens whose cumulative masses account for the top 40%, the next 30%, the next 20% and final 10% of tokens empirically generated by humans, respectively.

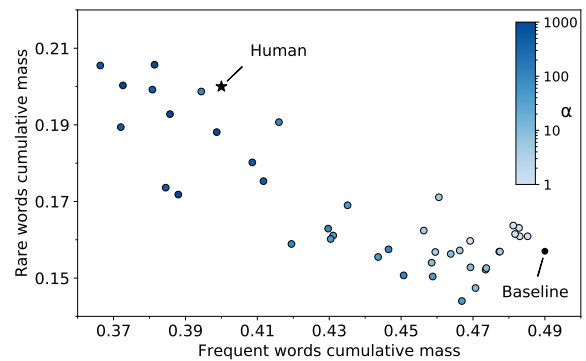


Figure 5: Vocabulary control with unlikelihood training: more probability mass is transferred from Frequent words to Rare words as we increase the  $\alpha$  weighting parameter. The maximum likelihood baseline is far from the human distribution.

from the Persona-Chat dialogue task (Zhang et al., 2018). We show examples from Dialogue NLI in

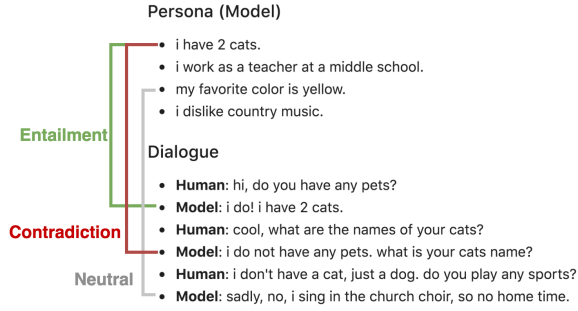


Figure 6: Dialogue NLI from (Welleck et al., 2019b).

	Train	Test	Valid
Entailment	95k	4613	4959
Triple-Entailment	105k	5285	5481
Neutral	110k	5500	5700
Negatives	110k	5500	5700

Table 5: Dialogue NLI two utterance generation task dataset statistics.

Figure 6. The original data consists of sentence pairs  $(s_1, s_2)$  along with a label (E, N, or C), and was constructed by developing a schema and employing crowdworkers to label utterances with relation triples. The labels are then inferred from the triple representation.

We first transform the original classification dataset into a form useful for unlikelihood training of a generative dialogue model. We consider two setups: (i) a two utterance generation task; and (ii) a full dialogue generation task.

**Two Utterance Generation Task** We adapt the initial dialogue NLI dataset by using entailing and neutral training sentence pairs as plausible positive utterances, and contradicting pairs as negatives. That is, if a pair  $(s_1, s_2)$  from Dialogue NLI has label E or N, the example  $(\mathbf{x}, \mathbf{y}) = (s_1, s_2)$  is added to  $\mathcal{D}^+$ , otherwise (label C) it is added to  $\mathcal{D}^-$ .

We consider two types of entailment: entailing sentence pairs that appear together in a dialogue in the original Persona-Chat dataset and are therefore natural (‘entailment’), and those that only entail via their triple relations (‘triple-entailment’). The latter are more challenging, noisier targets. Evaluation is performed by measuring the test set perplexity over the four target label types, where contradictions should have relatively higher perplexity. We additionally evaluate a selection accuracy task, where for each test example there are two candidate responses: a positive and a negative

(contradicting) statement. The candidate response with the lowest perplexity is considered to be the model’s selection, and we measure the selection success rate. Evaluation is broken down by positive type (entailment, triple-entailment, neutral). Dataset statistics are given in Table 5.

**Full Dialogue Task** To evaluate in a more realistic setup that involves full dialogue rather than a single utterance, we take full Persona-Chat dialogues (Zhang et al., 2018) similar to Figure 6, and map back the dialogue NLI data to provide positive and negative continuations of the dialogue. We consider continuations as either triple entailing utterances, neutral utterances or contradictions – where the relation triple is used to match the existing persona or dialogue turns by the same speaker to induce the label. That is, an example  $(\mathbf{x}, \mathbf{y})$  consists of a dialogue history  $\mathbf{x} = \{p_1, \dots, p_k, u_1, \dots, u_t\}$  and utterance  $\mathbf{y} = s_2$ , where  $(s_1, s_2)$  is a sentence pair from Dialogue NLI, and at least one sentence in  $\mathbf{x}$  has the same relation triple as  $s_1$ . When the pair  $(s_1, s_2)$  is labeled as E or N in Dialogue NLI, the example  $(\mathbf{x}, \mathbf{y})$  is added to  $\mathcal{D}^+$ , and otherwise it is added to  $\mathcal{D}^-$ .

**Results** Our MLE baseline obtains a perplexity of 11.4, in line with current best systems on this task (Lewis et al., 2019). Unfortunately, despite being good on such standard metrics, our baseline models fail at our coherence task. As seen in Table 6 for the two utterance task, the perplexity of contradicting utterances (12.5) is on average *lower* than for neutral (36.7) or triple-entailing utterances (17.5), although it is higher than entailing utterances. We believe this is due to contradicting utterances having high word overlap with the premise utterance, coupled with an inability to judge incoherence. Viewed as a selection task between utterances, picking the utterance with the lowest perplexity, this means the selection rates of non-contradicting utterances are very low, e.g. picking neutral utterances over contradicting utterances only 18% of the time. Even fully entailing utterances are only picked 73% of the time. Similar results are found on the full dialogue task as well, see Table 7.

Unlikelihood training brings large improvements in coherence metrics, whilst minimally impacting overall dialogue perplexity. After applying unlikelihood, perplexity for contradicting utterances has a clear signature, with very large av-

Data + Model	Selection Accuracy			Perplexity				
	Entail	Tr.-E	Neutral	Entail	Tr.-E	Neutral	Contradict	ConvAI2
MLE Baseline	72%	41%	18%	8.54	17.5	36.7	12.5	11.4
UL (Dialogue NLI)	96%	85%	78%	9.1	26.6	39.4	248.9	11.9

Table 6: Test evaluation on the Dialogue NLI two utterance generation task, comparing standard likelihood (MLE) models trained on pushshift.io Reddit and ConvAI2 with unlikelihood loss NLI training. Results are broken down according to whether the premise and positive candidate are entailing, triple-entailing, or neutral (Entail, Tr.-E, Neutral). Selection Accuracy measures how often the model assigns lower perplexity to the positive candidate than to the negative candidate in the pair. Top two rows: for standard maximum likelihood models, the perplexity of contradicting utterances is *lower* compared to neutral or triple-entailing utterances (albeit higher compared to entailing utterances), showing partial failure at the coherence task. Bottom row: NLI Unlikelihood training yields large improvements on all coherence metrics, while minimally increasing overall perplexity.

Data + Model	Selection Accuracy (vs. Neg)		Perplexity			
	Triple-Entail	Neutral	Triple-Entail	Neutral	Contradict	ConvAI2
MLE Baseline	66.5%	36.8%	23.3	45.1	35.9	11.4
UL (Dialogue NLI)	89.0%	69.8%	21.5	40.3	63.5	11.8

Table 7: Test evaluation on the Full Dialogue NLI generation task. NLI unlikelihood training improves coherence metrics compared to likelihood (MLE) training. For UL, the triple-entailing or neutral candidates are assigned relatively lower perplexity compared to contradicting candidates, with higher selection accuracy for coherent labels.

Premise	Hypothesis	$\mathcal{L}_{MLE}$ PPL	$\mathcal{L}_{UL}$ PPL
Yes, I love watching baseball and basketball. I do not like running though.	(C) I love running.	25.5	226.9
	(E) I despise running.	29.9	9.4
Yes, I love watching baseball and basketball. I do like running though.	(E) I love running.	26.2	3.1
	(C) I despise running.	42.8	247.1
We did too but working in real estate for 12 years . sucked up a lot of time	(E) I have been working as a real estate agent for the past 12 years.	3.9	3.8
	(C) We did too but working in real estate for fifteen years sucked up a lot of time.	3.1	17.6

Figure 7: Example perplexities of a baseline maximum likelihood model ( $\mathcal{L}_{MLE}$ ) and our unlikelihood trained model ( $\mathcal{L}_{UL}$ ) when generating the provided hypotheses, given the premise. The maximum likelihood trained model assigns high probability (low perplexity) to contradictory generations, while unlikelihood does not.

erage values compared to entailing or neutral utterances, e.g. 248.9 vs. 9.1 for contradict vs. entail on the two utterance task. This converts to corresponding large increases in selection accuracy across all types on both tasks, e.g., an increase from 18% to 78% on neutral statements on the two utterance task, and from 37.4% to 69.8% on the full dialogue task.

Some example model predictions are given in Figure 7, comparing the MLE baseline and unlikelihood model perplexities of generating the given hypotheses. The likelihood model cannot differentiate between contradicting and entailing statements easily, while there are large perplexity differences for the unlikelihood model in these cases.

## 5 Conclusion

Generating consistent and coherent human-like dialogue is a core goal of natural language research. We studied several aspects that contribute to that goal, defined metrics to measure them, and proposed algorithms that improve them, mitigating some of the failings of maximum likelihood training, the current dominant approach. Our method defines objective functions under the umbrella of unlikelihood: during training, we wish to make inconsistent dialogue unlikely by lowering the probability of such events occurring. This makes generative models repeat themselves less, copy the context less, and use more rare words from the vocabulary – closer to matching human statistics. Further, utilizing supervised datasets with labeled



coherent and incoherent utterances and applying unlikelihood yields measurably improved levels of coherence with respect to the aspect measured, in this case contradiction. Future work could apply this same technique with other supervised data, e.g. correcting causal or commonsense reasoning errors (Zellers et al., 2019; Qin et al., 2019).

## References

- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020. The second conversational intelligence challenge (ConvAI2). In *The NeurIPS '18 Competition*, pages 187–208, Cham. Springer International Publishing.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. [Wizard of wikipedia: Knowledge-powered conversational agents](#). In *Proceedings of the International Conference on Learning Representations*.
- Angela Fan, David Grangier, and Michael Auli. 2018. [Controllable abstractive summarization](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54. Association for Computational Linguistics.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [EL15: Long form question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- Jessica Fidler and Yoav Goldberg. 2017. [Controlling linguistic style aspects in neural language generation](#). In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.
- Saadia Gabriel, Antoine Bosselut, Ari Holtzman, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2019. [Co-operative generator-discriminator networks for abstractive summarization with narrative flow](#). *arXiv preprint arXiv:1907.01272*.
- Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. [Hafez: an interactive poetry generation system](#). In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, Canada. Association for Computational Linguistics.
- Hiroaki Hayashi, Zecong Hu, Chenyan Xiong, and Graham Neubig. 2019. [Latent relation language models](#). *arXiv preprint arXiv:1908.07690*.
- Tianxing He and James Glass. 2019. [Negative training for neural dialogue response generation](#). *arXiv preprint arXiv:1903.02134*.
- Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. [The curious case of neural text degeneration](#). *arXiv preprint arXiv:1904.09751*.
- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. [Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring](#). *arXiv preprint arXiv:1905.01969*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). *arXiv preprint arXiv:1910.13461*.
- Margaret Li, Jason Weston, and Stephen Roller. 2019. [ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons](#). In *Proceedings of the NeurIPS Workshop on Conversational AI*.
- Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. [ParlAI: A dialog research software platform](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.
- Aigul Nugmanova, Andrei Smirnov, Galina Lavrentyeva, and Irina Chernykh. 2019. [Strategy of the negative sampling for training retrieval-based dialogue systems](#). In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 844–848. IEEE.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. [Language models as knowledge bases?](#) *arXiv preprint arXiv:1909.01066*.
- Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. [Counterfactual story reasoning and generation](#). 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 5042–5052, Hong Kong, China. Association for Computational Linguistics.

- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI Blog*, 1(8).
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. [What makes a good conversation? how controllable attributes affect human judgments](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019a. [Neural text generation with unlikelihood training](#). *arXiv preprint arXiv:1908.04319*.
- Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019b. [Dialogue natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. [Learning semantic textual similarity from conversations](#). In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, Melbourne, Australia. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [Hellaswag: Can a machine really finish your sentence?](#) *arXiv preprint arXiv:1905.07830*.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Model	PPL	F1	Repetition	
			Context	Label
Human	-	-	.160	.0006
MLE Baseline	8.3	.373	.582	.002
UL (Context only)	8.8	.345	.270	.001
UL (Label only)	8.3	.371	.645	.000
UL (Context + Label)	8.5	.358	.445	.003

Table 8: Evaluation on the Wizard of Wikipedia task test set, comparing standard likelihood (MLE) with repetition unlikelyhood loss training, where both methods use beam search (beam size of 5).

## A Repetition Control with Beam Search

The experiments on repetition and copying in the main paper were carried out with greedy decoding for simplicity. In this section we show that similar results hold with beam decoding as well. Using a beam size of 5, we take the same 4 models from Table 2 and compute metrics with beam instead. The results are given in Table 8 which show similar trends to before, except the baseline model using beam tends to suffer more from repetition, which is a known result (Holtzman et al., 2019). Note that we simply evaluated the same unlikelyhood models as before, but we expect that better results could be obtained by performing sequence level unlikelyhood training with beam search in the training loop, as well as choosing hyperparameters specifically with this kind of decoding being used to measure validation performance.

## B Nucleus Sampling for Vocabulary control

Table 9 compares the MLE baseline, unlikelyhood with increasing  $\alpha$  values, and Nucleus sampling (Holtzman et al., 2019) with hyperparameter  $p$  in terms of distribution and F1 score. The vocabulary unlikelyhood fine-tuning shifts probability mass from the over-represented frequent words towards under-represented medium and rare words, with the effect strengthening as  $\alpha$  increases. At a small cost to perplexity and F1, the unlikelyhood tuning reduced the overuse of common tokens by 9 points, matching the human rate, while improving the production of rare tokens by 3 percentage points.

Nucleus sampling is a popular method that can also produce generations closer to the human vocabulary distribution. It does this by sampling from the model’s probability distribution rather

Model	PPL	F1	Token frequency classes			
			Freq	Med	Rare	Rarest
Human	-	-	.400	.300	.200	.100
MLE Baseline	11.4	.199	.491	.282	.157	.068
Nucleus $p = 0.3$	11.4	.180	.452	.315	.168	.064
Nucleus $p = 0.4$	11.4	.171	.440	.320	.172	.068
Nucleus $p = 0.5$	11.4	.160	.425	.322	.180	.072
Nucleus $p = 0.6$	11.4	.151	.411	.318	.192	.078
Nucleus $p = 1.0$	11.4	.141	.394	.302	.201	.101
UL, $\alpha = 10^0$	11.4	.200	.483	.289	.163	.063
UL, $\alpha = 10^1$	11.9	.201	.459	.328	.154	.058
UL, $\alpha = 10^2$	12.5	.190	.430	.335	.163	.071
UL, $\alpha = 10^3$	14.4	.174	.399	.339	.188	.073

Table 9: Unlikelyhood loss applied to vocabulary distributions. Stronger  $\alpha$  terms greatly shift probability mass from the most Frequent words to Medium and Rare words, at a small cost to PPL and F1. Frequent, medium, rare and rarest token classes are defined as the sets of tokens whose cumulative masses account for the top 40%, the next 30%, the next 20% and final 10% of tokens empirically generated by humans, respectively. Nucleus sampling can also produce a distribution close to human with parameter  $p$  close to 1, but with larger losses in F1.

than using beam search, where the sampler restricts to the smallest set of tokens with total mass above a threshold  $p \in [0, 1]$ . Small values of  $p$  are similar to greedy sampling. Increasing  $p$  yields distributions closer to human, but with large losses in F1 score, e.g.  $p = 0.5$  has a similar distribution to unlikelyhood with  $\alpha = 10^2$  but the F1 scores are 0.160 vs. 0.190. This can be understood because maximizing likelihood during decoding yields better token accuracy than sampling (Welleck et al., 2019a), so the unlikelyhood training approach to both use likelihood decoding *and* match the human distribution can obtain the best of both worlds.

## C Human Evaluation

**Description of ConvAI2 vocabulary setup** We follow (Li et al., 2019) and perform a pairwise comparison with full-length model conversations. We first collected 252 model-human conversations with each of the models (MLE baseline, and weights for  $\alpha$  of Unlikelyhood, examples in 8). We then set up a pairwise-comparison using the software of (Li et al., 2019), using the same question (“Who would you prefer to talk to for a long conversation?”) and use the exact same quality control question (a baseline greedy model without repetition control, versus a human). We collected ap-

proximately 200 preferences per model comparison and filtered annotators who failed quality control.

**Description of ELI5 repetition setup** We follow (Fan et al., 2019) and perform a pairwise evaluation where human annotators were asked “*which response answers the question better?*” A screenshot of the UI is shown in Figure 9. Human evaluators were asked to rate a total of 5 questions, two of which were quality control annotations. The quality control examples contained the real human responses, along with model predictions: one question contained a baseline model, and one contained an unlikelihood model. Annotators which did not pick humans in quality controls were removed from the final setups. We collected 200 annotations comparing the baseline and the unlikelihood model.

**Results** Evaluation results from all evaluated matchups are shown in Figure 10. We find our repetition-controlled ELI5 model significantly outperforms the MLE baseline. We find that two of the vocabulary repetition significantly outperform the MLE baseline. We compute significance with a two-tailed binomial test ( $p < .01$ ).

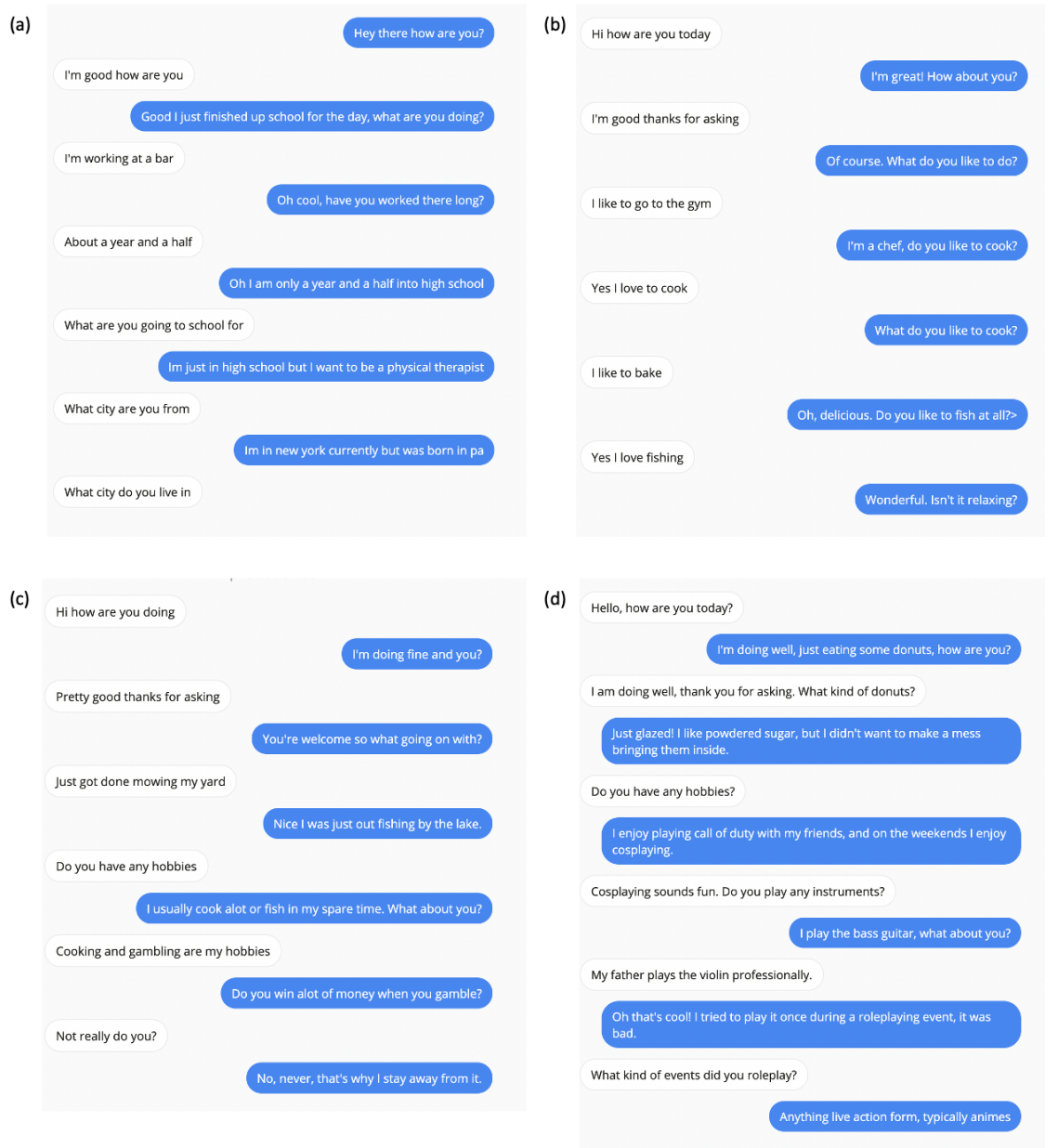


Figure 8: Examples of model-human conversations collected during human evaluation of the vocab unlikelyhood models. Human utterances are in blue bubbles, model utterances are in white. Conversations (a) and (b) are from the baseline. Conversations (c) and (d) are from the  $\alpha = 10^2$  model and more frequently employ rarer words.



## Which response answers the question better?

You are currently at question 1 / 5

You will see a question and two different answers. Choose which answer you think **better answers the question**.

Consider the **readability of the answer** and how **accurately** the question is answered.

You **must** also provide a very brief (a few words or a sentence) justification for your choice. Failure to do so could result in your hits being rejected.

You may need to scroll down to see the full answers.

**Question:**

Why is it so easy to fall asleep watching the TV but radio or simply waiting to fall asleep don't have the same effect?

**Answers:**

It's because you're not watching the tv. You're watching the radio.

I think it's because you're distracted by the tv and the radio.

Which better answers the question, considering both readability and accuracy?

☐ Answer 1 better answers the question
 ☐ Answer 2 better answers the question

Please provide a brief justification for your choice (a few words or a sentence)

Please enter here...

Figure 9: Screenshot of the Human Evaluator UI.

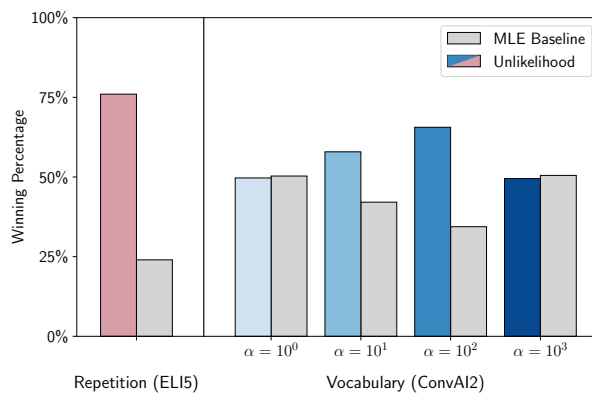


Figure 10: Complete Human Evaluation results. Human evaluators do not significantly prefer the  $\alpha = 10^0$  and  $\alpha = 10^3$  models over the baseline model.