# Probing the Robustness of Trained Metrics for Conversational Dialogue Systems

**Anonymous EMNLP submission**

## Abstract

This paper introduces an adversarial method to stress-test trained metrics for the evaluation of conversational dialogue systems. The method leverages Reinforcement Learning to find response strategies that elicit optimal scores from the trained metrics. We apply our method to test recently proposed trained metrics. We find that they all are susceptible to give high scores to responses generated by rather simple and obviously flawed strategies that our method converges on. For instance, simply copying parts of the conversation context to form a response yields competitive scores or even outperforms responses written by humans.

## 1 Introduction

One major issue in developing conversational dialogue systems is the large efforts required for evaluation. This hinders rapid developments in this field because frequent evaluations are not possible or very expensive. The goal is to create automated methods for evaluating to increase the efficiency. Unfortunately, methods such as BLEU (Papineni et al., 2002) have been shown to not be applicable to conversational dialogue systems (Liu et al., 2016). Following this observation, in recent years the trend towards training methods for evaluating dialogue systems emerged (Lowe et al., 2017; Deriu and Cieliebak, 2019; Mehri and Eskenazi, 2020; Deriu et al., 2020). The models are trained to take as input a pair of context and candidate response, and output a numerical score that rates the candidate for the given context. These systems achieve high correlations to human judgments, which is very promising. Unfortunately, these systems have been shown to suffer from instabilities. (Sai et al., 2019) showed that small perturbations to the candidate response already confuse the trained metric. In this work. we go one step further: we propose a method that automatically finds strategies that elicit very high scores from the trained metric, while being of obvious low quality. Our method can be applied to automatically test the robustness of trained metrics against adversarial strategies that exploit certain weaknesses of the trained metric.

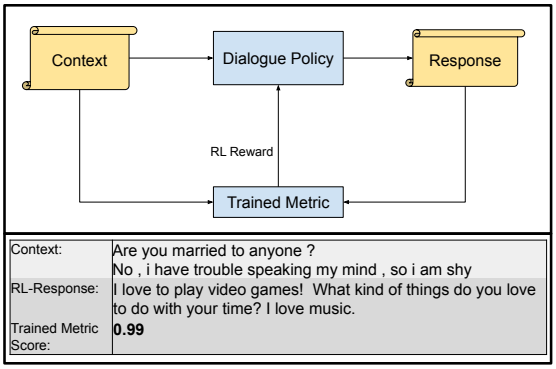


Figure 1: Overview of the process. It takes a context and an response generated by a dialogue policy and computes a score based on the trained metric. The score is then used as a reward to update the policy. In this example, the policy converges to a fixed response, which achieves an almost perfect score, although it is clearly a low-quality response. The policy always returns this response, regardless of the context, and the trained metric always scores it perfectly.

Our method uses a trained metric as a reward in a Reinforcement Learning setting, where we fine-tune a dialogue system to maximise the reward. Using this approach, the dialogue system converges towards a degenerate strategy that gets high rewards from the trained metric. It converges to three different degenerate types of strategies to which the policy converges in our experiments: the *Parrot*, the *Fixed Response*, and the *Pattern*. For each dataset and metric, an adversarial response is found, which belongs to one of the three strategy types. The responses generated form these strategies then achieve high scores on the metric. Even more, in most cases the scores are higher than the scores achieved by human written responses. Figure 1 shows the pipeline. The dialogue policy

receives a reward signal from the trained metric. Over time, the policy converges to a fixed response, which objectively does not match the context but gets a near perfect score on the trained metric. We release the code [1].

## 2 Related Work

**Trained Metrics.** In recent years the field of trained metrics gained traction after word-overlap methods have been shown to be unreliable (Liu et al., 2016). The first of these metrics is ADEM (Lowe et al., 2017), which takes as input a context, a reference, and the candidate response and returns a score. The main issue with ADEM is the reliance on references and annotated data (i.e. human ratings of responses), which are costly to obtain, and it needs to be redone for each domain. RUBER (Tao et al., 2018) extended ADEM by removing the reliance on annotated data for training. However, it still relies on a reference during inference. AutoJudge (Deriu and Cieliebak, 2019) removed the reliance on references, which allows the evaluation of multi-turn behaviour of the dialogue system. However, AutoJudge still leverages annotated data for training. USR (Mehri and Eskenazi, 2020) is a trained metrics which does not rely on either annotated data or any reference. It is trained in a completely unsupervised manner while still achieving high correlation to human judgement (0.4 Spearman Correlation). Similarly, MAUDE (Sinha et al., 2020) is trained as an unreferenced metric built to handle online evaluation of dialogue systems.

**Robustness of Trained Metrics.** There is not yet much research on the robustness of trained metrics. Sai et al. (2019) evaluated the robustness of ADEM by corrupting the context in different ways. They show that by just removing punctuation, the scores of ADEM change, and in 64% of cases are superior to the scores given for the same response without removed punctuation. Other corruption mechanism yielded similar results. Yeh et al. (2021) compared a large variety of automated metrics for dialogue system evaluation by comparing e.g. turn- and dialogue-level correlation with human judgements and studying the impact of the dialogue length. They find that no single metric is robust against all alternations but see potential in ensembling different metrics. Novikova et al. (2017) investigate automated metrics in the task-oriented NLG domain and find that the metrics do not sufficiently reflect human ratings.

[1] URL Placeholder

**Algorithm 1:** Advantage Actor-Critic Algorithm, where $\pi_\theta$ denotes the policy, $c$ denotes the context, $r$ the response generated by the policy, and $s$ denotes the score by the automated metric, i.e., the reward.

**while** *training* **do**
  sample $c$ from pool of contexts;
  $r = \pi_\theta(c)$ generate response;
  $s = R(c, r)$ compute reward;
  fit action-value function $Q_\sigma$ i.e., $\mathcal{L}(\sigma) = \frac{1}{2}\sum_i \left\| R(c,r) + Q_(c', r') - Q_\sigma(c,r) \right\|$; compute the advantage $A(r,c) = R(r,c) - Q(c,r) + Q(c',r')$;
  $\theta = \theta + \alpha \bigtriangledown J_{RL}(\theta)$ fit policy;
**end**

## 3 Method

Our method applies a trained metric as a reward signal $R(c, r)$ to update a dialogue system $\pi(c)$ in a reinforcement learning setting, where $c$ denotes the context and $r$ the response. The dialogue system is trained by generating a response for a context, which is then scored by the automated metric. The dialogue system is then updated using the score as the reward. This process is repeated for different contexts. We use the Actor-Critic framework to optimize the policy (Sutton et al., 1999). See Algorithm 1 for an overview. The policy gradient is defined as $\bigtriangledown J_{RL}(\theta) = \bigtriangledown_\theta log\ \pi_\theta(r|c) * A(r, c)$, where $\pi_\theta(r|c)$ defines the probability of the generated response for the given context, and $A(c, r)$ the advantage function.

The learned policy depends on the reward function, i.e., the automated metric. If the reward function is susceptible to adversarial attacks, the policy most likely will generate an objectively suboptimal solution, which is rated highly be the automated metric. Conversely, we expect the policy to improve the dialogue systems's responses if the automated metric is robust against adversarial examples.

## 4 Experimental Setup

### 4.1 Datasets

We perform the evaluation on three widely-used used datasets in the dialogue modelling domain. Namely, Dailydialog (Li et al., 2017), Empathetic Dialogues (Rashkin et al., 2019), and PersonaChat (Zhang et al., 2018).

| Metric | Strategy | Response |
|---|---|---|
| | | PersonaChat |
| ATT | Fixed | yea!!! 1!! 2!! 3!! * * * fucking fucking fucking * * [ [ [ fucking * fucking * |
| BLM | Fixed | that sounds like a lot of fun. what do you like to do in your spare time? |
| MAUDE | Fixed | What kind of work do you have? What do you like to do in your free time? |
| USR FULL | Parrot | - |
| USR MLM | Fixed | i am a stay at home mom and i am trying to figure out what i want to do with my life |
| USR RET | Fixed | I love to be a musician. I love music. What kind of music do you listen to as a music lover |
| | | Dailydialog |
| ATT | Fixed | ! freaking out! one of these days! * * one * * freaking * * out! * even * * damn * * even damn |
| BLM | Fixed | that would be great! what do you do for a living, if you don't mind me asking? |
| MAUDE | Fixed | I hope it works out for you. What kind of car did you get? |
| USR FULL | Pattern | i'm not sure if i'd like to [copy context tokens]. i'll let you know if i do. |
| USR MLM | Fixed | i am not sure if i am going to be able to go out of my way to get to know each other or not. |
| USR RET | Parrot | - |
| | | Empathetic Dialogues |
| ATT | Fixed | I know right? I felt SO SO ASHAmed of myself. I felt so embar assed. |
| BLM | Fixed | I'm so sorry to hear that. What happened, if you don't mind me asking? |
| MAUDE | Fixed | I wish I could go back in time and be a kid again. I miss those days. |
| USR FULL | Pattern | i don't think it's [ random context noun]. i'm sorry to hear that. what do you mean by that? |
| USR MLM | Fixed | I don't know what I'm going to do if it doesn't work out. I'm not sure what to do. |
| USR RET | Parrot | - |

Table 1: The strategies achieved for each metric and domain.

### 4.2 Metrics

We use a variety of different state-of-the-art automated metrics that were developed for evaluating conversational dialogue systems without reference, i.e., so-called unreferenced metrics. These are metrics where no reference is needed, they just use the context and response to determine the score. They can be represented as a function $s = R(c, r)$, which rate the response $r$ for a given context $c$.

We selected state-of-the-art trained metrics which achieve good correlations to human judgements to evaluate our approach. Namely, USR (Mehri and Eskenazi, 2020), ATT (Gao et al., 2021), and MAUDE (Sinha et al., 2020). Additionally, we added the Blender language model score (BlenderLM) (Roller et al., 2020). For the ATT [2], MAUDE [3], and BlenderLM metric [4], we use the out-of-the-box models provided by the respective authors. For the USR metric, we perform a custom training on each dataset. Furthermore, we report the USR-retireval (*USR Ret*), USR-masked-language-model *USR MLM*, and the USR-regression *USR Full* scores. Note that the *USR Full* is a combination of the *USR Ret* and *USR MLM* metric. More details can be found in Appendix A.

[2] https://github.com/golsun/AdversarialTuringTest

[3] https://github.com/facebookresearch/online_dialog_eval

[4] https://huggingface.co/facebook/blenderbot-400M-distill

### 4.3 Strategies

For our approach, we use Blenderbot as our policy (Roller et al., 2020), since it is currently a state-of-the-art conversational dialogue system [5]. For each domain, we use the validation set to perform the reinforcement learning. This is to avoid that the dialogue systems are fine-tuned on already seen data. We use the test set to evaluate the reward over the number of episodes. We perform the reinforcement learning for 15 epochs, where each epoch is composed of 500 updates. We noted from pre-experiments that this is enough for a dialogue system to converge to a degenerate strategy. We track the average reward achieved on the test set after each epoch. Each experiment is repeated 10 times, since we expect the policy to converge to slightly different strategies in different runs. We select the repetition which achieved the highest score (i.e., reward), and use it to determine the strategy. We also experimented with automated strategy detection, see Appendix B.

## 5 Results

The policies typically converge towards one of following three degenerate strategies.

**Parrot.** Here, the policy simply copies parts of the context into the response. Sometimes, it applies slight changes. For instance, it changes the pronouns from "you" to "I".

**Fixed Response.** Here, the policy converges on a fixed response which it returns regardless of the

[5] Note that here we are referring to Blenderbot as dialogue system. BLM is using the Blenderbot LM as metric.

| Dailydialog | | | | | | |
|---|---|---|---|---|---|---|
| | USR RET | USR MLM | USR FULL | ATT | MAUDE | BLM |
| BL | 0.440 | 0.426 | 4.951 | 0.0002 | 0.664 | 0.096 |
| HU | 0.928 | 0.409 | 7.904 | 0.0006 | 0.898 | 0.183 |
| COPY | 0.998 | 0.811 | 9.429 | 0.0002 | 0.921 | 0.233 |
| FIXED | - | **0.505** | - | **0.435** | **0.985** | **0.239** |
| PARROT | **0.998** | - | - | - | - | - |
| PATTERN | - | - | **7.091** | - | - | - |
| Empathetic Dialogues | | | | | | |
| | USR RET | USR MLM | USR FULL | ATT | MAUDE | BLM |
| BL | 0.935 | 0.298 | 7.645 | 0.001 | 0.820 | 0.087 |
| HU | 0.891 | 0.384 | 7.611 | 0.120 | 0.942 | 0.264 |
| COPY | 0.996 | 0.885 | 9.617 | 0.054 | 0.935 | 0.358 |
| FIXED | - | **0.912** | - | **0.731** | **0.976** | **0.333** |
| PARROT | **0.994** | - | - | - | - | - |
| PATTERN | - | - | **7.240** | - | - | - |
| PersonaChat | | | | | | |
| | USR RET | USR MLM | USR FULL | ATT | MAUDE | BLM |
| BL | 0.847 | 0.185 | 6.797 | 0.0006 | 0.844 | 0.070 |
| HU | 0.927 | 0.267 | 7.512 | 0.0024 | 0.951 | 0.153 |
| COPY | 0.925 | 0.794 | 8.933 | 0.0001 | 0.898 | 0.223 |
| FIXED | **0.977** | **0.852** | - | **0.813** | **0.933** | **0.250** |
| PARROT | - | - | **7.542** | - | - | - |
| PATTERN | - | - | - | - | - | - |

Table 2: Scores achieved by humans (HU), Blenderbot (BL) and the degenerate strategies with regard to the different metrics for each domain.

context.

**Pattern.** This is a mix between the *Parrot* and the *Fixed Response*. It creates a fixed template, which is filled with parts of the context.

Table 1 shows the selected responses for each pair of domain and metric. For all metrics except *ATT*, the fixed response is composed of a grammatically correct sentence. Note that these responses are always returned by the fine-tuned dialogue system, regardless of the cotnext.

### 5.1 Scores

Table 2 shows the main results. In almost all cases the degenerated strategy outperforms the vanilla Blenderbot and humans with respect to the automated metric. The most striking example is the *ATT* metric, where the fixed response achieves scores that are by orders of magnitude better than the ones achieved by humans. For both *USR Ret* and *MAUDE*, the scores achieved by the fixed response are almost perfect, i.e. they are close to 1.0, which is the upper bound. Also for *USR MLM*, the scores are significantly higher than the ones achieved by Blenderbot. Interestingly, the *USR FULL* seems to be more immune to the pattern that were found. However, even for *USR FULL*, the parrot strategy beats the humans by a significant margin in the *PersonaChat* domain.

**Copy.** We also display the scores achieved by simply copying the context on each metric, which is inspired by the *Parrot* strategy. The only metric which is immune to the *Copy* strategy is *ATT*. Under all the other metrics, the *Copy* achieves very high scores. In some cases it achieves even better scores than the converged policy. For instance, for the *Dailydialog* domain, it achieves $0.811$ points under the *USR MLM* metric, which is $0.3$ point higher than the converged policy and twice as good as the human score.

## 6 Conclusion

Trained metrics for automatic evaluation of conversational dialogue systems are an attractive remedy for the costly and time-consuming manual evaluation. While high correlation with human judgements seems to validate the metrics regarding their ability to mimic human judging behaviour, our analysis shows that they are susceptible to rather simple adversarial strategies that are easily identified by humans. In fact, all metrics that we used failed to recognize degenerate responses. Our approach is easily adaptable to any newly developed trained metric that takes as input a pair of context and response. There are no known remedies for this problem. Thus, the next open challenge is to find methods that improve the robustness.

## References


Jan Deriu and Mark Cieliebak. 2019. Towards a Metric for Automated Conversational Dialogue System Evaluation and Improvement. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 432–437, Tokyo, Japan. Association for Computational Linguistics.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on Evaluation Methods for Dialogue Systems. *Artificial Intelligence Review*, pages 1–56.

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.

Xiang Gao, Yizhe Zhang, Michel Galley, and Bill Dolan. 2021. An adversarially-learned turing test for dialog generation models. *arXiv preprint arXiv:2104.08231*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. 2019. Re-Evaluating ADEM: A Deeper Look at Scoring Dialogue Responses. In *Proceedings of the thirty-third AAAI Conference on Artificial Intelligence*, volume 33 of *AAAI'19*, pages 6220–6227, Honolulu, Hawaii, USA.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. *arXiv preprint arXiv:2005.00583*.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 1057–1063, Cambridge, MA, USA. MIT Press.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In *Proceedings of the thirty-second AAAI Conference on Artificial Intelligence*, AAAI'18, New Orleans, Louisiana USA.

Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *ArXiv*, abs/2106.03706.

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.

# A Correlation between Human Judgements and Trained Metrics

In this section, we evaluate the metrics with regards to their correlation to human judgments to show that these metrics have reasonable performance. For this, we sample 100 contexts for each domain. For each domain, we use a set of bots to create a response for each context. Furthermore, we add the human response to the pool of responses for each context. Then, we let crowdworkers annotate the responses. We correlate the scores of each metric on the same set of contexts and responses to the human annotations.

## A.1 Domains and Bots

We perform the evaluation on the three datasets from the main paper.

**Dailydialog.** We prepared 5 bots using ParlAI (Miller et al., 2017). We fine-tune a GPT-2 (GPT) model (Radford et al., 2018), a BERT-Rank (BR) model, a sequence-to-sequence model (S2) with attention, and a weakly trained sequence-to-sequence model (DR). We also use the Blender model (Roller et al., 2020), although it was not specifically tuned on Dailydialog.

**Empathetic Dialogues.** We prepared the same pool of models as in Dailydialog.

**PersonaChat.** We mostly reuse the openly available systems of the ConvAI2 challenge (Dinan et al., 2020), namely, Lost in Conversation[6] (LC) and Huggingface (HF) [7] , and KVMemNN (KV). We also add the Blender model, which is also trained in this domain, a custom-trained BERT-Rank model (BR), and a sequence-to-sequence model (S2). Together with the DR model, the pool consists of 7 different dialogue systems.

## A.2 Annotation Process

Since we perform the evaluation on a static-context setting, we also add the human response (i.e., the gold response) to the pool of systems. For evaluation, we use 600 samples for Dailydialog and Empathetic Dialogues each, and 800 samples for the PersonaChat domain. Each sample is composed of a context (sampled from the test set), and a generated response. We annotated the overall quality of each sample on a Likert scale from 0 (bad) to

[6] https://github.com/atselousov/transformer_chatbot

[7] https://github.com/huggingface/transfer-learning-conv-ai

| | DD | ED | PC |
|---|---|---|---|
| USR RET | 0.561 | 0.524 | 0.605 |
| USR MLM | 0.138 | 0.452 | 0.303 |
| USR REG | 0.559 | 0.573 | 0.585 |
| ATT | 0.154 | 0.385 | -0.099 |
| MAUDE | 0.211 | 0.086 | 0.357 |
| BLENDERLM | 0.201 | 0.287 | 0.266 |

Table 3: Correlations of the automated metrics to human judgments. For all runs $p < 0.05$.

2 (good) using Mechanical Turk[8]. Each sample is annotated by three different humans. As the final score, we use the average score of the three annotations. For each metric, we apply the metric to all samples, and then compute the Spearman correlation between the human scores and the scores predicted by the metric.

### A.3 Correlation to Human Judgements

Table 3 shows the correlations of the human judgments to each of the metrics for each domain. For all domains, the *USR* metric performs best, achieving strikingly high correlations to humans. *MAUDE* also achieves good correlation scores on the PersonaChat domain, and *ATT* performs well on the Empathetic Dialogues domain. *BlenderLM* has mediocre performance on all domains equally.

### A.4 Original USR

Note that the *USR Ret* scores are significantly higher than in the original paper (Mehri and Eskenazi, 2020), which is due to the fact that we use more turns to represent the context, whereas the original implementation uses only the previous turn for the context. In the original implementation, *USR Ret* achieves a Spearman correlation of 48.67 on our annotated data. If we train our implementation of *USR Ret* using only one turn to represent the context, we also achieve a Spearman correlation of 40.34, which is comparable to the original. We did not experience a discrepancy on the *USR MLM* model, where the original model achieves the same correlation as ours.

## B Strategy Selection

We observed in our experiments that the dialogue system almost always converges to one of three degenerate strategies. In order to atomize their detection in the experiments, we used a set of heuristics for their identification.

[8] https://www.mturk.com/

### B.1 Heuristics

Since the strategies are very simple, we propose heuristics to detect the policy automatically. This avoids the need for manual inspection of a potentially large amount of log files. For this, we introduce the following measures.

- *Response Frequency.* The percentage of times that the same response is generated for all samples in the test set.
- *Lexical Variety.* The ratio between number of different tokens and the total number of tokens over all responses in the test set.
- *BLEU score.* The BLEU score between the context and the response. This is computed for each pair of context and responses and then averaged over all samples in the test set.
- *Jaccard score.* The Jaccard overlap between the context and response tokens. Analogous to the BLEU score, the Jaccard overlap is computed between each context-and response-pair, and then averaged over all samples in the test set.

These measures can be used to detect the various strategies the policy converges to. For instance, a high *Response Frequency* indicates that the policy converges to a fixed response. A high *BLEU* score and *Jaccard score* indicate that the policy converges to the parrot strategy. A low *Response Frequency*, a low *Lexical Variety* and a moderate *Jaccard score* indicate that the policy converges to a pattern. A pattern is composed of a fixed template where parts are filled with tokens from the context.

### B.2 Application of the Heuristics

For each run, we use these metrics to determine which strategy the policy has converged on. The final strategy is extracted by selecting the best epoch across all 10 runs for each domain. If the *Response Frequency* is larger than 0.7, we extract the most common sentence and use this as our fixed response. If the *BLEU* score is larger than 0.2, we assign the parrot strategy. If the *Response Frequency* is smaller than 0.1, the *Lexical Variety* is smaller than 0.15, and the *Jaccard score* is larger than 0.05, it indicates a pattern emerged. In this case, we manually extract the pattern.

### B.3 Overview

Table 4 shows the measures used to perform the automated strategy selection. The automated strategy

| domain | metric | Avg Reward | Resp Freq | Lex Var | BELU | Jacccard | Strategy Inferred | Strategy Manual | Strategy Final |
|---|---|---|---|---|---|---|---|---|---|
| Persona Chat | ATT | 0.77 | 0.14 | 0 | 0 | 0 | Not Conclusive | Fixed Response | Fixed Response |
| Persona Chat | BLM | 0.41 | 0.01 | 0.11 | 0.03 | 0.06 | Not Conclusive | Fixed Response | Fixed Response |
| Persona Chat | MAUDE | 0.98 | 0.7 | 0.01 | 0 | 0.07 | Fixed Response | | Fixed Response |
| Persona Chat | USR Full | 7.7 | 0 | 0.09 | 0.42 | 0.48 | Parrot | | Parrot |
| Persona Chat | USR MLM | 0.84 | 0.94 | 0.01 | 0.01 | 0.1 | Fixed Response | | Fixed Response |
| Persona Chat | USR Ret | 1 | 0.8 | 0 | 0 | 0.07 | Fixed Response | | Fixed Response |
| Dailydialog | ATT | 0.42 | 0.55 | 0.01 | 0 | 0.01 | Not Conclusive | Fixed Response | Fixed Response |
| Dailydialog | BLM | 0.26 | 0.32 | 0.01 | 0 | 0.05 | Not Conclusive | Fixed Response | Fixed Response |
| Dailydialog | MAUDE | 0.99 | 0.99 | 0 | 0 | 0.06 | Fixed Response | | Fixed Response |
| Dailydialog | USR Full | 7.65 | 0 | 0.11 | 0.08 | 0.15 | Pattern | | Pattern |
| Dailydialog | USR MLM | 0.52 | 1 | 0 | 0 | 0.04 | Fixed Response | | Fixed Response |
| Dailydialog | USR Ret | 0.99 | 0 | 0.19 | 0.21 | 0.31 | Parrot | | Parrot |
| Empathetic Dialogues | ATT | 0.78 | 0.98 | 0 | 0 | 0.04 | Fixed Response | | Fixed Response |
| Empathetic Dialogues | BLM | 0.33 | 0.47 | 0.03 | 0 | 0.05 | Not Conclusive | Fixed Response | Fixed Response |
| Empathetic Dialogues | MAUDE | 0.98 | 0.96 | 0 | 0 | 0.06 | Fixed Response | | Fixed Response |
| Empathetic Dialogues | USR Full | 8.67 | 0.01 | 0.07 | 0.04 | 0.1 | Pattern | | Pattern |
| Empathetic Dialogues | USR MLM | 0.77 | 0.98 | 0 | 0 | 0.06 | Fixed Response | | Fixed Response |
| Empathetic Dialogues | USR Ret | 1 | 0 | 0.17 | 0.33 | 0.44 | Parrot | | Parrot |

Table 4: Scores achieved on the test set during the evaluation.

selection worked in 72% of cases. There are two main cases in which it was not conclusive. First, for the *ATT* metric, where for both the *Dailydialog* and *PersonaChat* domains no clear fixed response arose. However, after manual inspection, we noted that for the *PersonaChat* the policy generated the same tokens in various frequencies and orders. For the *Dailydialog* the most frequent response arose in 55% of cases. Thus, we used this fixed response. The second case is the *BLM* metric. For all the domains we selected the most frequent response, although it appeared in less than 70% of cases.

## C Full Results

Table 5 shows all scores achieved by the dialogue systems on the respective metrics. Furthermore, we also added the average score of the Amazon Mechanical Turk judges, which ranges from (0-2).

## D Technical Explanation

One potential reason why our approach is able to find a degenerate strategy lies in the exploration problem in reinforcement learning. Blender's language model can be interpreted as a policy which performs a sequence of actions, i.e., sampling a sequence of tokens. Thus, the language model loss during standard Blender training can be interpreted as an indicator for how sure the policy is of its actions. A high language model loss indicates that the policy assigns low probability scores to its actions. Conversely, a low language model loss indicates that the policy is sure of it's actions. This could be further investigated by measuring the entropy of the language model. Indeed, in all our experiments, we notice that the language model loss collapses toward a very small value. This indicates that the language model collapsed to a single simple strategy. Figure 2 shows the language model loss over the number of steps. The loss quickly collapses from an average of 4 points to around 0.5 points. At the same time the average reward (orange) rises from 0.78 to 0.92. Similarly, the response frequency rises from 0 to 0.94. In the middle, the loss rises again, which indicates the search for a new strategy. This coincides with a lower response frequency.

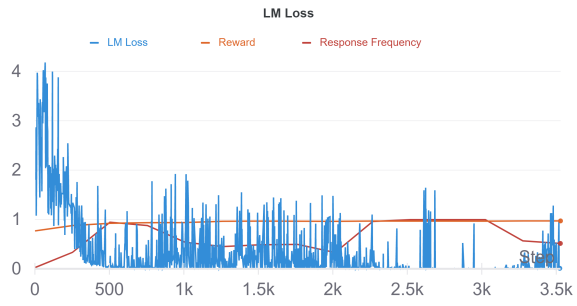


Figure 2: The language model loss (blue), the Average Reward (orange), and the Response Frequency (red) over time.

## E Examples

In Tables 6, 7, and 8, we show examples of the outputs from the fine-tuned Blenderbot model. For each of the five metrics, we show the output to which Blenderbot converged to when using the metric as a reward. Furthermore, we show the score which the respective metric assigns to the generated response. Note that the *Parrot* strategies simply copy the text form the context. For the *Empathetic Dialogues* dataset, the degenerate strategy prepends a "I'm not sure" to the context. For the *PersonaChat*, the degenerate strategy prepends a "i've always wanted to". The *Copy* strategy (see Table 2 in main Paper), ignores these prefaces, and simply copies the context.

Dailydialog

| | AMT | USR RET | USR MLM | USR FULL | ATT | MAUDE | BLM |
|---|---|---|---|---|---|---|---|
| BR | 1.836 | 0.928 | 0.409 | 7.904 | 0.0006 | 0.898 | 0.177 |
| BL | 1.386 | 0.440 | 0.426 | 4.951 | 0.0002 | 0.664 | 0.096 |
| HF | 1.656 | 0.925 | 0.080 | 6.989 | 0.0026 | 0.866 | 0.371 |
| HU | 1.782 | 0.928 | 0.409 | 7.904 | 0.0006 | 0.898 | 0.183 |
| S2 | 1.024 | 0.512 | 0.300 | 5.050 | 0.0003 | 0.895 | 0.183 |
| DR | 0.729 | 0.308 | 0.338 | 3.900 | 0.0001 | 0.891 | 0.204 |
| PARROT | - | **0.998** | *0.811* | *9.429* | *0.0002* | *0.921* | *0.233* |
| FIXED | - | - | **0.505** | - | **0.435** | **0.985** | **0.239** |
| PATTERN | - | - | - | **7.091** | - | - | - |

Empathetic Dialogues

| | AMT | USR RET | USR MLM | USR FULL | ATT | MAUDE | BLM |
|---|---|---|---|---|---|---|---|
| BR | 1.808 | 0.891 | 0.384 | 7.611 | 0.120 | 0.942 | 0.260 |
| BL | 1.640 | 0.935 | 0.298 | 7.645 | 0.001 | 0.820 | 0.087 |
| HF | 1.610 | 0.887 | 0.644 | 8.292 | 0.044 | 0.948 | 0.462 |
| HU | 1.816 | 0.891 | 0.384 | 7.611 | 0.120 | 0.942 | 0.264 |
| S2 | 0.702 | 0.493 | 0.145 | 4.510 | 0.010 | 0.932 | 0.159 |
| DR | 0.822 | 0.354 | 0.182 | 3.759 | 0.001 | 0.936 | 0.199 |
| PARROT | - | **0.996** | *0.8848* | *9.617* | *0.054* | *0.935* | *0.358* |
| FIXED | - | - | **0.912** | - | **0.731** | **0.976** | **0.333** |
| PATTERN | - | - | - | **7.240** | - | - | - |

PersonaChat

| | AMT | USR RET | USR MLM | USR FULL | ATT | MAUDE | BLM |
|---|---|---|---|---|---|---|---|
| BR | 1.350 | 0.725 | 0.211 | 6.120 | 0.0020 | 0.946 | 0.138 |
| BL | 1.507 | 0.847 | 0.185 | 6.797 | 0.0006 | 0.844 | 0.070 |
| HF | 1.480 | 0.794 | 0.272 | 6.707 | 0.0023 | 0.925 | 0.152 |
| HU | 1.623 | 0.927 | 0.267 | 7.512 | 0.0024 | 0.951 | 0.153 |
| KV | 1.147 | 0.538 | 0.217 | 4.982 | 0.0023 | 0.852 | 0.122 |
| LC | 1.572 | 0.879 | 0.103 | 6.769 | 0.0011 | 0.918 | 0.195 |
| S2 | 0.681 | 0.390 | 0.121 | 3.814 | 0.0013 | 0.845 | 0.111 |
| DR | 0.906 | 0.482 | 0.268 | 4.779 | 0.0002 | 0.537 | 0.220 |
| PARROT | - | *0.925* | *0.794* | **8.933** | *0.0001* | *0.898* | *0.223* |
| FIXED | - | **0.977** | **0.852** | - | **0.813** | **0.933** | **0.250** |
| PATTERN | - | - | - | - | - | - | - |

Table 5: Scores achieved by the various dialogue policies with regard to the different metrics. For each metric, we show the average scores achieved by each dialogue system and humans (HU) for each domain. We then show the average score achieved by applying the strategy to which the policy converged to. The AMT metric denotes Amazon Mechanical Turk ratings.

| Turn | Utterance | Score | Strategy |
|---|---|---|---|
| *usr_ret* | | | |
| 1 | Did this just come on suddenly ? | | |
| 2 | I felt OK until an hour or so ago , and then I just doubled over . | | |
| 3 | Have you exercised strenuously or played sports recently ? | | |
| RL-Response | i feel ok until an hour or so ago, and then<br>i just doubled over. have you exercised strenuously or played sports recently | 0.9986 | Parrot |
| *usr_mlm* | | | |
| 1 | Yes . Who is speaking ? | | |
| 2 | This is Kelly speaking . I am calling from London . How are you , Mr . Chang ? | | |
| 3 | Oh , Kelly , I couldn't recognize your voice . Well , I'm fine . | | |
| RL-Response | i am not sure if i am going to be able to<br>go out of my way to get to know each other or not. | 0.4820 | Fixed |
| *usr_full* | | | |
| 1 | You really shouldn ' t write it down . You should memorise it . You can change it to a more convenient number if you like . | | |
| 2 | Ok , I ' Ve entered my PIN . There are several options on the screen . I guess I just choose which service I want , don ' t I ? | | |
| 3 | Yes . Choose the service you need to use . If you choose to withdraw money , the machine will ask you how much you want . | | |
| RL-Response | i'm not sure if i'd like to change my pin<br>to a more convenient number if i like. i'll let you | 7.466 | Pattern |
| *att* | | | |
| 1 | Hello , hot water overflowed from the bathtub . Would you help me ? | | |
| 2 | Would you turn off the tap ? | | |
| RL-Response | hahaha!!! thats so funny!! im laughing so<br>hard!! its so true!! ive been laughing so | 0.1176 | Fixed |
| *maude* | | | |
| 1 | Good morning , Mr Zhang , I wonder whether you have got an answer from your superior . | | |
| 2 | Yes , we had a meting about it yesterday afternoon . | | |
| RL-Response | I hope it works out for you. What kind of car did you get? | 0.9901 | Fixed |
| *blender* | | | |
| 1 | There is a Matisse exhibition at the Museum of Modern Art . Larry and I will go . | | |
| 2 | Did you buy tickets already ? | | |
| 3 | No . Do we need tickets ? | | |
| RL-Response | what do you like to do in your spare time? do you have any hobbies? | 0.2078 | Fixed |

Table 6: Randomly sampled responses for the Dailydialog dataset. We randomly chose a context, and let the fine-tuned Blenderbot create a response, which at this point is degenerate. We then let the metric which was used as a reward score rate the generated response.

| Turn | Utterance | Score | Strategy |
|---|---|---|---|
| *usr_ret* | | | |
| 1 | My boyfriend's brothers got into a fist fight one night and he jumped in the middle. | | |
| 2 | That sucks. Did anyone get hurt? | | |
| 3 | Some things got broke and there were some bruises, but it was really scary to watch that. | | |
| RL-Response | i'm sure your boyfriend's brothers got into a fist fight one night and he jumped in the middle. did anyone get hurt? | 0.9982 | Parrot |
| *usr_mlm* | | | |
| 1 | Financially, things will turn around for my family soon! | | |
| 2 | I hope and wish so. In this tough economic climate many are finding it difficult to cope with financial responsibilities | | |
| 3 | Yes, and we had a layoff of five months. I believe in a brighter day, though, so I know it will improve. | | |
| RL-Response | I don't know what I'm going to do if it doesn't work out. I'm not sure what to do. | 0.7689 | Fixed |
| *usr_full* | | | |
| 1 | I had a major project and felt very ready to present it to everyone at the office | | |
| 2 | How'd it go | | |
| 3 | It went well, I think I gained confidence by being so prepared | | |
| RL-Response | i don't think i gained confidence by being so prepared. i'm sorry to hear that. what do you mean by that? | 9.068 | Pattern |
| *att* | | | |
| 1 | I once got in a really bad accident on my way home from work and it really upset me | | |
| 2 | Wow! Were you hurt? Was anyone else hurt? | | |
| 3 | luckily the only problem I had was my neck was sore for a few weeks but nobody else was hurt | | |
| RL-Response | I know right? I felt SO SO ASHAmed of myself. I felt so embar assed. | 0.8138 | Fixed |
| *maude* | | | |
| 1 | like who on earth writes var int value = 1, this is so ridiculous i am so annoyed at my brother, after everything i tought him, he doesn't even know the basics | | |
| RL-Response | I wish I could go back in time and be a kid again. I miss those days. | 0.9946 | Fixed |
| *blender* | | | |
| 1 | My friend left my front door open when she was house-sitting for me. My cat snuck out and was killed. | | |
| RL-Response | i feel like i know what i'm going to do, but i don't know what to do with myself. | 0.1299 | Fixed |

Table 7: Randomly sampled responses for the Empathetic Dialogues dataset. We randomly chose a context, and let the fine-tuned Blenderbot create a response, which at this point is degenerate. We then let the metric which was used as a reward score rate the generated response.

| Turn | Utterance | Score | Strategy |
|---|---|---|---|
| *usr_ret* | | | |
| 1 | nemo , i love him already . its a tea mix , green and black . | | |
| 2 | the tea sounds cool . is there honey in it ? | | |
| 3 | i add honey to it , yes . i drink it while running museum tours . | | |
| RL-Response | I love to be a musician. I love music. What kind of music do you listen to as a music lover | 0.9997 | Fixed |
| *usr_mlm* | | | |
| 1 | hi i am a musician with big dream | | |
| RL-Response | i am a stay at home mom and i am trying to figure out what i want to do with my life | 0.8854 | Fixed |
| *usr_full* | | | |
| 1 | it is tough and smelly . 25 people and 3 toilets . | | |
| 2 | that would be awful . you really should move . | | |
| 3 | i cant yet . need to save money , spend it too quickly . | | |
| RL-Response | i've always wanted to move 25 people and 3 toilets that would be awful you really should move i cant yet need to save money | 8.498 | Parrot |
| *att* | | | |
| 1 | me too ! i aspire to become a physical therapist and run a 5k ! | | |
| 2 | awesome . i just started a new job as a life coach | | |
| 3 | that is great ! i get straight as so maybe you could coach me in social skills . | | |
| RL-Response | yea!!! 1!! 2!! 3!! * * * [ * * fucking * fucking fucking * [ fucking * * | 0.7417 | Fixed |
| *maude* | | | |
| 1 | how are you doing today | | |
| 2 | hello , how are you today ? | | |
| 3 | i am doing really well | | |
| RL-Response | What are you up to today? What do you like to do in your free time? | 0.9923 | Fixed |
| *blender* | | | |
| 1 | haha agreed ! or a lot . either way . i try to eat well though | | |
| 2 | me too . should i trade my mini van in for a sports car . | | |
| 3 | yes . who does not love a sports car . | | |
| RL-Response | what kind of mini van do you have? i have a corvette and a camaro | 0.1970 | Fixed |

Table 8: Randomly sampled responses for the PersonaChat dataset. We randomly chose a context, and let the fine-tuned Blenderbot create a response, which at this point is degenerate. We then let the metric which was used as a reward score rate the generated response.