A Follow-up Study on Evaluation Metrics Using Follow-up Utterances

Toshiki Kawamoto[1] **Yuki Okano**[1] **Takato Yamazaki**[2]
Toshinori Sato[2] **Kotaro Funakoshi**[1] **Manabu Okumura**[1]
[1] Tokyo Institute of Technology [2] LINE Corporation
{kawamoto, okano, funakoshi, oku}@lr.pi.titech.ac.jp
{takato.yamazaki, toshinori.sato}@linecorp.com

Abstract

As the human evaluation of dialogs is costly, reliable automatic reference-free evaluation methods are important. In this paper, we focus on the FED and FULL automatic unsupervised reference-free evaluation metrics, which evaluate dialogs by using the likelihood of manually-designed follow-up utterances, and reportedly show considerable performance. In our experiment using English and Japanese dialog competition datasets, FED and FULL did not correlate well with the human evaluation. However, when a subset of follow-up utterances was chosen for each dataset, FED showed strong correlations with the subsets. The obtained results suggest that selecting the optimal follow-up utterances is crucial and depends on the target domain and language.

1 Introduction

Performance improvements of pre-trained language models have enabled more and more excellent dialog systems (Xu et al., 2022; Shuster et al., 2022; Adiwardana et al., 2020; Thoppilan et al., 2022) to be released every year, but it is common to evaluate the superiority of dialog systems by humans (Ji et al., 2022). However, as human evaluations are time-consuming and costly, it would be more efficient to use automatic evaluation in the development phase. The current study aims to identify an automatic evaluation metric that strongly correlates with human evaluations.

There are two types of metrics for automatic evaluation of dialogs: those that require reference responses (reference-based metrics, e.g., BLEU (Papineni et al., 2002), FBD (Xiang et al., 2021), etc.) and those that do not (reference-free metrics, e.g., perplexity, USR (Mehri and Eskenazi, 2020b), FED (Mehri and Eskenazi, 2020a), DynaEval (Zhang et al., 2021), etc.) Readers can refer to Table 1 of Yeh et al. (2021) for a comprehensive comparison of evaluation metrics.

Reference-based evaluation is commonly used for translation and summarization tasks. However, since a wide range of responses can be appropriate for a given input, it is less suitable for dialog evaluation. Moreover, it is costly, often more than human evaluations. Reference-free metrics should alleviate these issues because they do not require preparing references beforehand, enabling a broader range of possible responses beyond just a few references.

Among reference-free metrics, we focus on FED, which is a fully unsupervised metric that has been reported to have a strong correlation with human evaluations (Yeh et al., 2021; Ji et al., 2022), and FULL (De Bruyn et al., 2022), which is an improved version of FED. FED and FULL calculate likelihoods of multiple follow-up utterances by a language model to evaluate dialogs (see §2 for details). To the best of our knowledge, other reference-free metrics require supervised training. The need for supervised training makes it difficult to apply the metrics to dialogs in different domains and languages.

In this paper, we examine whether FED and FULL can be used both in English and Japanese using dialog system competition datasets. To begin with the conclusion, we have found that both FED and FULL are domain-sensitive and do not correlate as expected with human evaluations in the examined datasets. However, we have also found that FED can be optimized rather easily to correlate well with human evaluations by choosing a subset of follow-up utterances for both languages. Predicting rankings of dialog systems on the basis of the automatic evaluation scores, the method which uses a subset of follow-up utterances showed strong ranking correlations in all datasets. These results suggest that the proposed five follow-up utterances of FULL are simply a particular adaptation to the tested dataset, i.e., the FED dataset (Mehri and Eskenazi, 2020a) and the optimal follow-up utterances depends on the target domain and language.

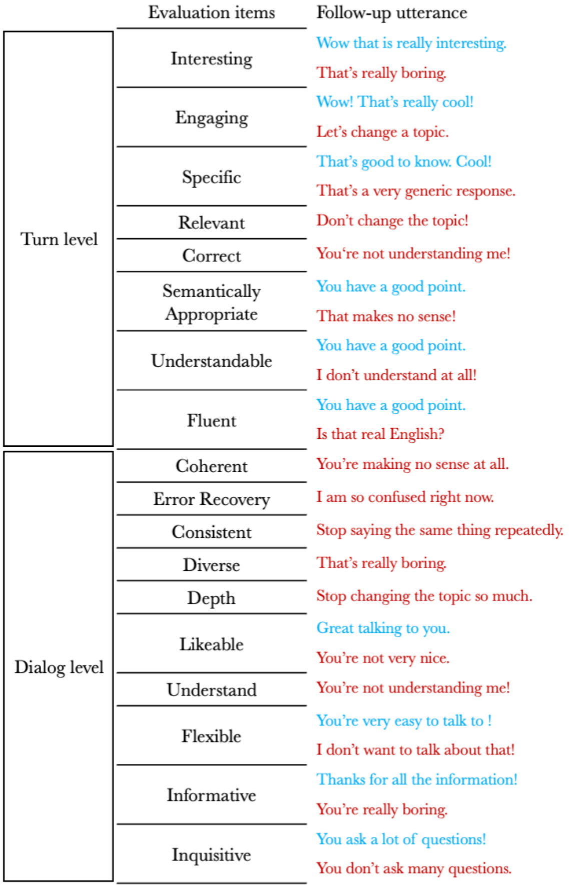
2 FED and FULL Evaluation Metrics

FED Given a dialog history, FED (Mehri and Eskenazi, 2020a) measures the log-likelihood of follow-up utterances from multiple perspectives, called 'evaluation items', such as *Interesting* and *Engaging* with respect to the dialog. FED has two levels of items, turn and dialog, with eight and ten items for the turn and dialog levels, respectively, for a total of 18 items. Each item has an average of 4.05 follow-up utterances. Figure 1 shows the complete evaluation items and examples of follow-up utterances. In the turn level, a follow-up utterance (such as, "You have a good point") is added after an utterance of the speaker to be evaluated in a dialog, and the log-likelihood is calculated over both the dialog and added utterance. In the dialog level, a follow-up utterance is added to the end of the dialog, and the log-likelihood is calculated similarly.

The follow-up utterances are divided into two types: positive and negative. For example, the positive utterances in the *Interesting* item are those that increase the likelihood when the previous utterance is interesting, such as "Wow, that is really interesting". The negative utterances are those that increase the likelihood when the previous utterance is not interesting, such as "That's really boring." The FED score is obtained by subtracting a mean log-likelihood of negative utterances from the one of positive utterances.

To compute the log-likelihood, FED uses the publicly available DialoGPT (Zhang et al., 2020). Since no additional data preparation or model training is required, the evaluation can be performed relatively easily compared with other automatic dialog evaluation metrics (Li et al., 2021; Zhang et al., 2021). Ji et al. (2022) reports that the correlation between FED and manual evaluation was 0.59 in their experiment to measure the correlation between automatic and human evaluations.

FULL FULL (De Bruyn et al., 2022) is an improved version of FED. It differs from FED in three ways: (1) FED calculates the log-likelihood including dialog history and follow-up utterances, while FULL calculates the conditional log-likelihood of follow-up utterances given dialog history, (2) FULL only uses a selection of five follow-up utterances, and (3) FULL uses the Blender language model (Roller et al., 2021) to calculate likelihood in accordance with a comparison among multiple language models. The correlation coefficient of FULL with human evaluations in the FED dataset is reported to be 0.69 while that of FED and DynaEval yield 0.32 and 0.55 respectively (De Bruyn et al., 2022).



	Evaluation items	Follow-up utterance
Turn level	Interesting	Wow that is really interesting. That's really boring.
	Engaging	Wow! That's really cool! Let's change a topic.
	Specific	That's good to know. Cool! That's a very generic response.
	Relevant	Don't change the topic!
	Correct	You're not understanding me!
	Semantically Appropriate	You have a good point. That makes no sense!
	Understandable	You have a good point. I don't understand at all!
	Fluent	You have a good point. Is that real English?
Dialog level	Coherent	You're making no sense at all.
	Error Recovery	I am so confused right now.
	Consistent	Stop saying the same thing repeatedly.
	Diverse	That's really boring.
	Depth	Stop changing the topic so much.
	Likeable	Great talking to you. You're not very nice.
	Understand	You're not understanding me!
	Flexible	You're very easy to talk to ! I don't want to talk about that!
	Informative	Thanks for all the information! You're really boring.
	Inquisitive	You ask a lot of questions! You don't ask many questions.

Figure 1: The 18 FED evaluation items and corresponding example follow-up utterances. Positive and negative follow-up utterances are shown in blue and red, respectively. The evaluation items with only negative utterances indicate that no positive utterances are defined.

Issues in FED and FULL As previously described, FED uses 18 evaluation items, but the positive follow-up utterances of the three items, *Semantically Appropriate*, *Understandable*, and *Fluent*, are the same. There are other evaluation items for which the follow-up utterances are the same, and it is doubtful whether all the evaluation items are measured as independent items. Therefore, we hypothesize that using a number of the 18 evaluation items may be better than using all of them in dialog evaluation as FULL does. In addition, FULL uses five follow-up utterances that were highly correlated in the FED dataset, which

Table 1: Dataset sizes of ConvAI2 and DC3.

	ConvAI2	DC3-Opn	DC3-Situ
# of dialogs	568	239	296
# of teams	7	5	6
ave. # of dialogs per team	81.14	47.80	49.33
ave. # of utterances per dialog	11.45	30.00	30.00

may not be appropriate for another dataset. Moreover, the five selected utterances are all negative follow-up utterances, which may lead to a biased evaluation. Other than that, the effectiveness of using the conditional log-likelihood (FULL) instead of the unconditional log-likelihood (FED) is not empirically and sufficiently clear. And since the log-likelihood is calculated using a pretrained language model (Bommasani et al., 2021) such as GPT (Radford et al., 2019), there is a possibility that the follow-up text does not have to be an utterance, just putting the name of evaluation items might be sufficient. In the next section, we report our experiments to address these issues with two languages (English and Japanese) and two different domains (open chit-chats and situated conversations).

3 Experiments

We compare the scores of each automatic evaluation method to those of the human evaluation from two dialog competition datasets in terms of correlation. As we perform feature (item) selection in our experiment, each comparison is done in a two-fold cross-validation (TFCV) manner. Thus all reported results are the averages of two evaluation rounds. We also evaluate the methods in terms of the capability of predicting rankings in the competitions.

3.1 Datasets

We conducted our experiments on two datasets for the respective languages (English and Japanese). For the English dataset, we used ConvAI2[1] (Dinan et al., 2020). For the Japanese dataset, we used Dialog System Live Competition 3[2] (Higashinaka et al., 2020) (hereinafter referred to as DC3 for short). Both datasets are from performance competitions of dialog systems and contain human-rated dialogs between systems and humans. DC3 has the Open and Situation tracks (Opn and Situ for short). Table 1 shows the basic statistics of the datasets.

ConvAI2 This dataset focuses on open chit-chat about people's interests. We applied filters to exclude low-quality dialogs. Dialogs that met all of the following conditions are used.

- The minimum number of turns in the dialog is 7.
- The maximum number of turns is 15.
- The maximum ratio of speakers is 2.
- The maximum number of consecutive system and human utterances are both 2.
- The speaker speaks at least once.

Dataset size of Table 1 is the remaining dialogs after the filtering and the ranks of the systems (bots) were determined in accordance with the average values of human ratings for the one after the filtering.

DC3 This dataset consists of two tracks. The Open track (DC3-Opn) covers open-topic chit-chat and the Situation track (DC3-Situ) covers conversations in a specified situation (the system has to decline a request from his senior (the user) to serve as an alumni party organizer in a socially appropriate manner). We use the preliminary-round data of both tracks, which have about 50 human evaluations for each system.

3.2 Methods

We compare the following six FED-based methods to the original FED and FULL.

FED-Cond A method that applies the conditional log-likelihood to FED considering the difference between FED and FULL.

FED-Cond-Pos A FED-Cond variant that uses only the positive follow-up utterances.

FED-Cond-Neg A FED-Cond variant that uses only the negative follow-up utterances.

FED-Cond-Tag A method using the name of the evaluation items instead of the utterances as the follow-up utterances of FED-Cond.

FED-Selected A method that uses only the follow-up utterances in the selected FED items. We selected an optimal item set using sequential feature selection in each TVCF round. Specifically, we used the training data to correlate all combinations

[1] https://github.com/DeepPavlovAdmin/convai/tree/master/2018

[2] https://dialog-system-live-competition.github.io/dslc3/data.html

Table 2: Results of evaluation score correlations.

	ConvAI2	DC3-Opn	DC3-Situ
FED	0.229	-0.283	0.278
FULL	0.040	-0.018	0.279
FED-Cond	0.091	0.279	0.297
FED-Cond-Pos	-0.084	0.302	-0.010
FED-Cond-Neg	0.086	-0.296	0.258
FED-Cond-Tag	-0.079	0.040	0.001
FED-Selected	0.266	0.485	0.249
FED-Cond-Selected	**0.275**	**0.585**	**0.315**

that used one or more of the 18 evaluation items with the human evaluation, and used the set of evaluation items that obtained the highest correlation among them.

FED-Cond-Selected (FED-C-S for short) The combination of FED-Cond and FED-Selected.

3.3 Settings

The correlation between human and automatic evaluation for each dialog is measured as a Spearman's rank correlation coefficient.

We have to select a language model to calculate the likelihood. For ConvAI2, we used DialoGPT (762M) and Blender (400M) for FED and FULL, respectively, as originally proposed. For our six comparison methods in §3.2, we chose Blender for ConvAI2 because Blender scored better than DialoGPT in FULL experiment. For DC3, we used the Japanese GPT model[3] with all methods. The follow-up utterances were manually translated into Japanese by the first author.

3.4 Results of evaluation score correlations

Table 2 shows the averages of Spearman's rank correlation coefficients.

While FED showed positive correlations with human evaluations on ConvAI2 and DC3-Situ, it showed a negative correlation on DC3-Opn. FULL showed a positive correlation on DC3-Situ but showed almost no correlations on ConvAI2 and DC3-Opn. These results suggest that neither FED nor FULL is universally applicable to any dialog evaluation.

FED-Cond, which applied the conditional likelihood to FED, obtained stronger correlations than FED on DC3, but a much weaker correlation on ConvAI2. However, it showed stronger correlations than FULL for all the datasets. These results indicate that in terms of the differences between FULL and FED reviewed in §2, (1) the conditional log-likelihood, is only effective for a number of cases, and (2) the restriction to the five follow-up utterances proposed for FULL likely brings a negative impact on datasets other than the FED dataset.

Although FED-Cond showed positive correlations on both DC3-Opn and DC3-Situ, interestingly, the contributed follow-ups seem to be entirely different. FED-Cond-Pos, which uses only positive follow-ups, and FED-Cond-Neg, which uses only negative follow-ups, show inverse results on DC3-Open and DC3-Situ respectively. These results also suggest the non-universal nature of FED and FULL. This point is further supported by the result that FED-Selected, which uses the training data of TFCV to select evaluation items, has many gains from FED on both ConvAI2 and DC3-Opn.

FED-Cond-Tag showed almost no correlations in all datasets, indicating that in the case of experiments even with the foundation model, it is better to use utterances that reflect the intention of the evaluation item than the name of the evaluation items as a follow-up.

Finally, FED-C-S obtained stronger correlations than the others in all datasets. Although this is contradictory to the degeneration of FED-Cond from FED on ConvAI2, the use of the conditional likelihood seems beneficial as a whole.

3.5 Results of ranking correlations

The original purpose of the competitions was to determine the team's ranking. Therefore, we attempted to rank the teams by averaging the automatic evaluation scores of each team.

The Spearman's rank correlation coefficients between the competition and automatic rankings are shown in Table 3 and the detailed ranking results of each dataset can be found in Tables 4, 5, and 6. The results are indicated by "team label (score)". Red and blue indicate the teams that correctly and incorrectly predicted their positions, respectively. Team labels in each table refer to different teams among tables.

For DC3-Opn and ConvAI2, FED-C-S showed the strongest correlations. For DC3-Situ, FED-C-S (0.77) followed FULL (0.89) but still showed a strong correlation. We note that FULL and FED-C-S made the same ranking order for the top-three teams on DC3-Situ.

[3] https://huggingface.co/rinna/japanese-gpt-1b

Table 3: Results of ranking correlations.

	ConvAI2	DC3-Opn	DC3-Situ
FED	0.71	-0.50	0.37
FULL	-0.04	-0.30	**0.89**
FED-C-S	**0.86**	**1.00**	0.77

Table 4: ConvAI2 ranking results.

Rank	Gold	FED	FULL	FED-C-S
1.	A (3.12)	A (3.71)	F (8.44)	A (-0.46)
2.	B (2.84)	B (3.60)	A (8.36)	B (-0.53)
3.	C (2.58)	F (3.60)	E (8.33)	C (-0.64)
4.	D (2.36)	C (3.59)	B (8.28)	F (-0.72)
5.	E (2.00)	D (3.45)	G (8.22)	E (-0.77)
6.	F (1.95)	G (3.01)	D (8.14)	D (-0.86)
7.	G (1.73)	E (2.97)	C (8.11)	G (-0.86)
Spear.	-	0.71	-0.04	0.86

Table 5: DC3 Open track ranking results.

Rank	Gold	FED	FULL	FED-C-S
1.	A (3.83)	C (1.55)	E (3.55)	A (9.86)
2.	B (3.11)	E (1.34)	B (3.45)	B (9.25)
3.	C (2.64)	B (1.19)	C (3.44)	C (9.14)
4.	D (2.10)	D (1.18)	A (3.43)	D (8.92)
5.	E (1.45)	A (1.09)	D (3.38)	E (8.71)
Spear.	-	-0.50	-0.30	1.00

Table 6: DC3 Situation track ranking results.

Rank	Gold	FED	FULL	FED-C-S
1.	A (4.26)	D (1.49)	A (3.94)	A (8.09)
2.	B (3.92)	A (1.40)	C (3.65)	C (7.48)
3.	C (3.76)	E (1.39)	B (3.61)	B (7.40)
4.	D (3.76)	B (1.33)	E (3.54)	F (7.20)
5.	E (3.63)	C (1.24)	D (3.52)	D (7.07)
6.	F (3.28)	F (1.20)	F (3.47)	E (7.03)
Spear.	-	0.37	0.89	0.77

4 Analysis and Discussion

Seeing correlations in Table 2 and Table 3, FED-C-S obtained the perfect ranking correlation on DC3-Opn with a score correlation of 0.585. On the other hand, ConvAI2 received a relatively high ranking correlation of 0.86 with a score correlation of only 0.275. This suggests that sufficient degree of a score correlation varies with the dataset. In general, we suppose the stronger the correlation, the better the performance. However, it is difficult to know the sufficient degree of a score correlation for the dataset itself.

By selecting items (FED-Selected), we could improve correlations on ConvAI2 and DC3-Opn. However, it slightly degenerated on DC3-Situ. This may indicate the evaluation items in FED are insufficient to assess the strongly situated conversations. Moreover, the selection was done on the item basis. The observed difference between FED-Cond-Pos and FED-Cond-Neg suggests an utterance-basis selection would provide a better result.

In terms of item combinations, on ConvAI2, FED-Cond-Selected (FED-C-S) identified *Interesting*, *Engaging*, *Semantically Appropriate*, *Understandable*, and *Likeable* for one fold and *Interesting* only for another fold. On DC3-Opn, it identified *Specific*, *Relevant*, and *Fluent* for one fold and *Interesting*, *Specific*, *Correct*, and *Fluent* for another fold. On DC3-Situ, it identified *Interesting*, *Specific*, *Correct*, *Fluent*, and *Depth* for one fold and *Specific*, *Relevant*, *Semantically Appropriate*, *Depth*, *Understand* for another fold. Since the maximum number of items used in this study was five, we confirm that it is not necessary to use all 18 items. The number of necessary items is considered to vary depending on the nature and purpose of the dialog to be evaluated, as well as the situational frame. The selected items quite differed from each round. This might be due to the aforementioned redundancy between items.

Human evaluations on a dialog are usually provided as numerical rates in several indexes. Supervising a model to predict the rates requires a considerable number of evaluations. However, the selection of items (or follow-up utterances) could be done with a small number of ranked samples by searching for a selection that maximizes the correlation of two rankings of the samples from human and machine. This will be future work.

5 Conclusion

This paper examined two unsupervised reference-free dialog evaluation metrics of FED and FULL in two different languages. The experimental results showed that both were domain-sensitive and did not correlate as expected with the human evaluation. However, we successfully optimized FED to correlate well with human evaluations by selecting a subset of follow-up utterances by a simple feature selection method in both languages. Such a feature selection would be conducted with a very limited number of ranked dialog samples. As a future direction, we would also like to pursue a method that can select a good subset of followup utterances automatically without human evaluation.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2022. Open-domain dialog evaluation using follow-ups likelihood. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 496–504, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Ryuichiro Higashinaka, Kotaro Funakoshi, Tetsuro Takahashi, Michimasa Inaba, Yuiko Tsunomori, Reina Akama, Mayumi Usami, Yoshiko Kawabata, Masahiro Mizukami, Masato Komuro, and Dolca Tellols. 2020. The dialogue system live competition 3. *SIG-SLUD*, 90:23.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. 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.

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, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. 2021. Assessing dialogue systems with distribution distances. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2192–2198, Online. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5676–5689, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.