# Finetuning of Blenderbot 90M on Daily dialog, ConvAi2 and Wizard of Wikipedia datasets

**Hrishikesh Kanade**

## 1. Introduction

This project aims to finetune Blenderbot 90M [1] on recent datasets and measure the improvement of the model in the general chit-chat domain. The finetuning of the model on all datasets has been completed.

## 2. Problem Formulation

This project focuses on Dialog modeling and conversational AI/ML task. The Blenderbot 90M parameter model was released in 2020 by Facebook. The base model used in this project has been pre-trained on the Blended Skill talk (BST)[5] dataset. The model serves as a good baseline for testing the efficacy of recent conversational datasets.

## 3. Methods

I intend to finetune Blenderbot 90M using the ParlAI toolkit.

## 4. Dataset and Experiments

- Datasets: daily dialog [2], ConvAi2 [3], a subset of Wizard of Wikipedia [4]
- Test data: 500 samples from the test sets
- Hardware environment: Nvidia A100 40GB rented through Google Colab
- Model parameters: 90M parameters
- Training settings: 20 epochs for finetuning and, Adam Optimizer

Experiments: The Model is finetuned on the mentioned datasets and the performance is checked against the test set of Blended Skill Talk dataset and also the test set of the dataset used for finetuning.

The plots of the training metrics and performance will be included.

### 1. Finetuning with Daily dialog dataset:

#### a. Loss over the epochs

| | Base model | Epoch 1 | Epoch 2 | Epoch 3 | Epoch 4 | Epoch 10 | Epoch 20 |
|---|---|---|---|---|---|---|---|
| Daily dialog | 2.9318 | 2.9673 | 2.9143 | 2.8746 | 2.8594 | 2.8514 | 2.4618 |
| Blended skill talk | 2.6355 | 2.7283 | 2.6643 | 2.6900 | 2.7208 | 2.7490 | 2.9414 |

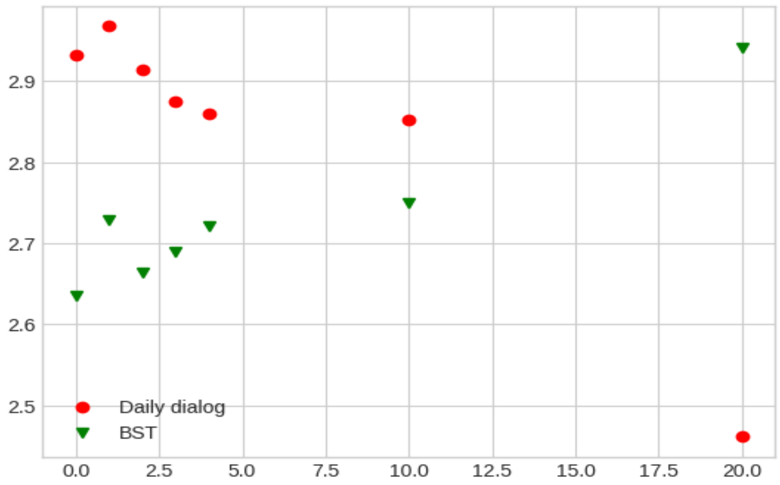


**Fig. 1. Plot of Loss for the Daily dialog and BST datasets over epochs during training**

**b. BLEU score over the epochs**

| | Epoch 1 | Epoch 2 | Epoch 3 | Epoch 4 | Epoch 10 | Epoch 20 |
|---|---|---|---|---|---|---|
| Daily dialog | 0.0953 | 0.1065 | 0.1075 | 0.1047 | 0.1091 | 0.1192 |
| BST | 0.1391 | 0.1385 | 0.1411 | 0.1398 | 0.1392 | 0.1346 |

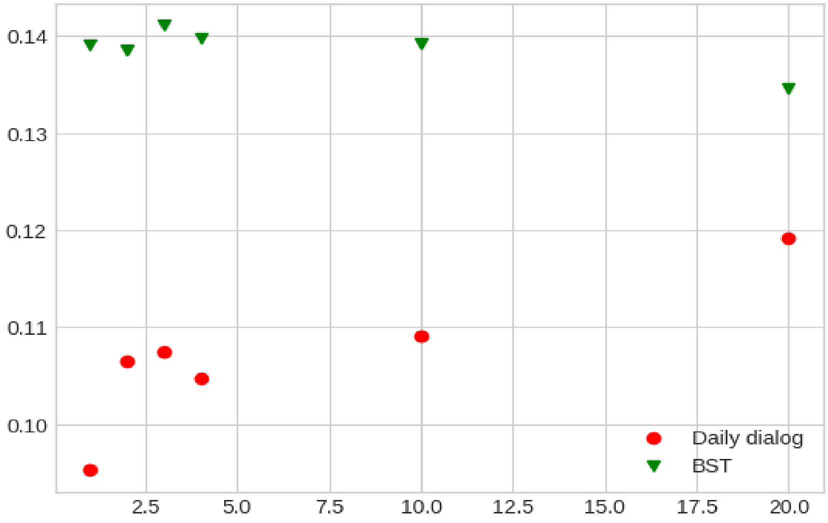


**Fig. 2. Plot of BLEU score over the epochs**

### c. F-1 score over the epochs

| | Epoch 1 | Epoch 2 | Epoch 3 | Epoch 4 | Epoch 10 | Epoch 20 |
|---|---|---|---|---|---|---|
| Daily dialog | 0.1238 | 0.1382 | 0.1396 | 0.1361 | 0.1414 | 0.1552 |
| BST | 0.1658 | 0.1656 | 0.1684 | 0.1661 | 0.1654 | 0.1594 |

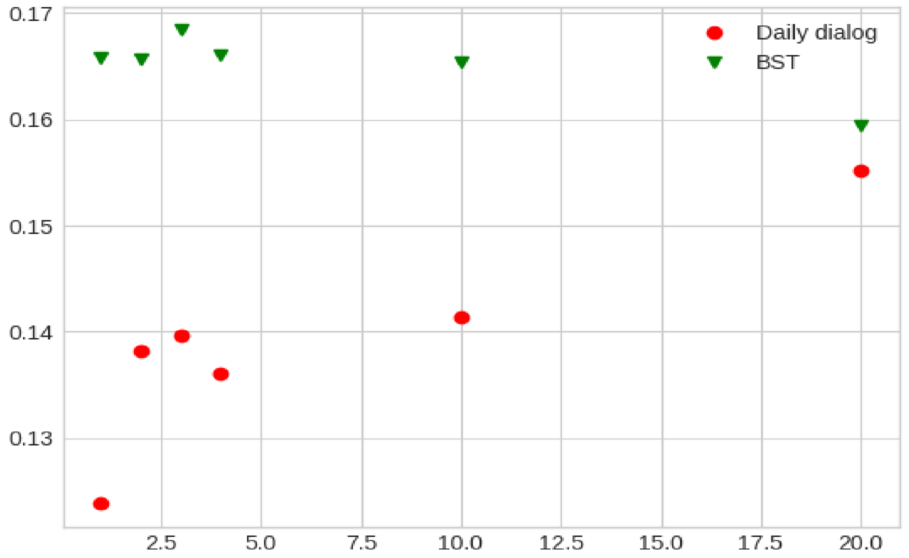


**Fig. 3. Plot of F-1 score over the epochs**

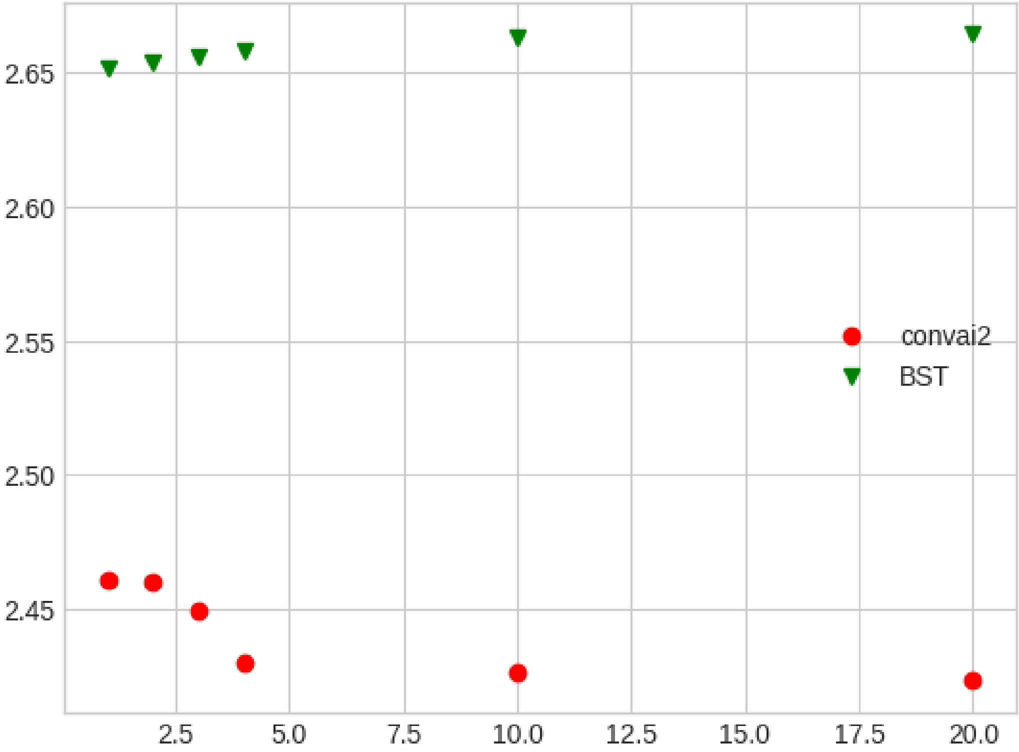
Observations after finetuning with the Daily dialog dataset:

- Over the epochs the Loss on Daily dialog decreases but the loss on BST increases.
- BLEU score doesn’t vary much for both datasets up to 10 epochs.
- F-1 score doesn’t vary much up to 10 epochs.
- The quality of the conversation is acceptable.
- Finetuning for 2 epochs is ideal w.r.t. to loss and BLEU score for both datasets.

## 2. Finetuning with Convai2 dataset:

### a. Loss over the epochs

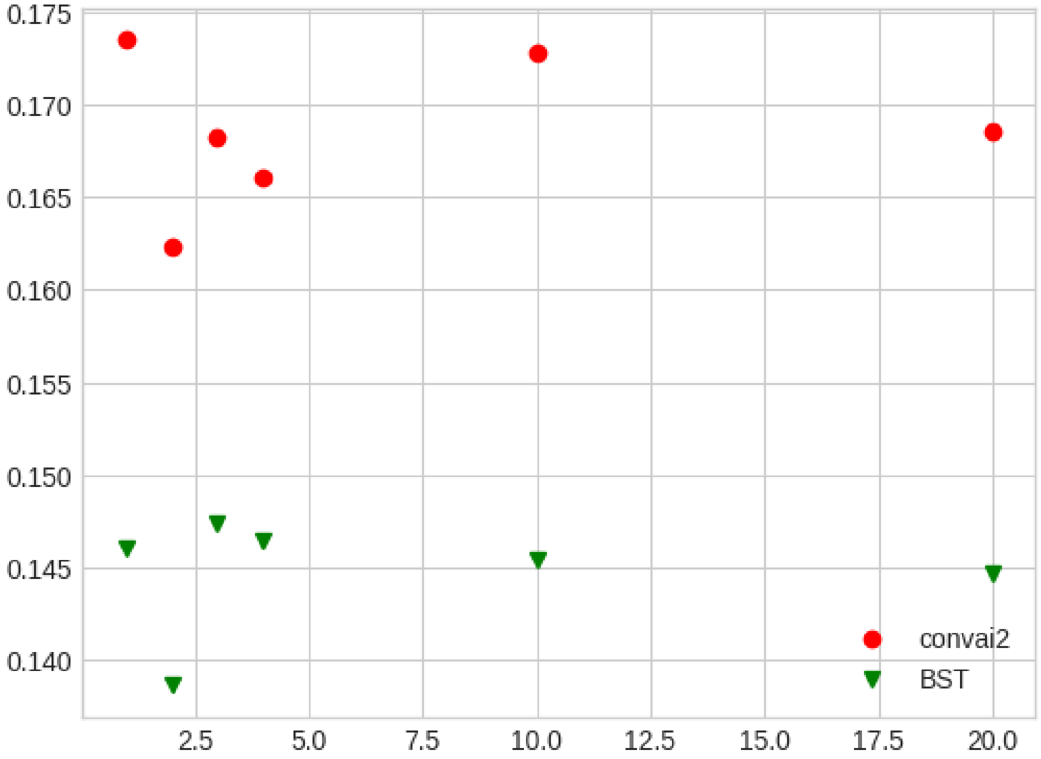
| | Epoch1 | Epoch 2 | Epoch 3 | Epoch 4 | Epoch 10 | Epoch 20 |
|---|---|---|---|---|---|---|
| ConvAI2 | 2.4608 | 2.4602 | 2.4495 | 2.4303 | 2.4265 | 2.4234 |
| BST | 2.6515 | 2.6534 | 2.6554 | 2.6575 | 2.6625 | 2.6639 |



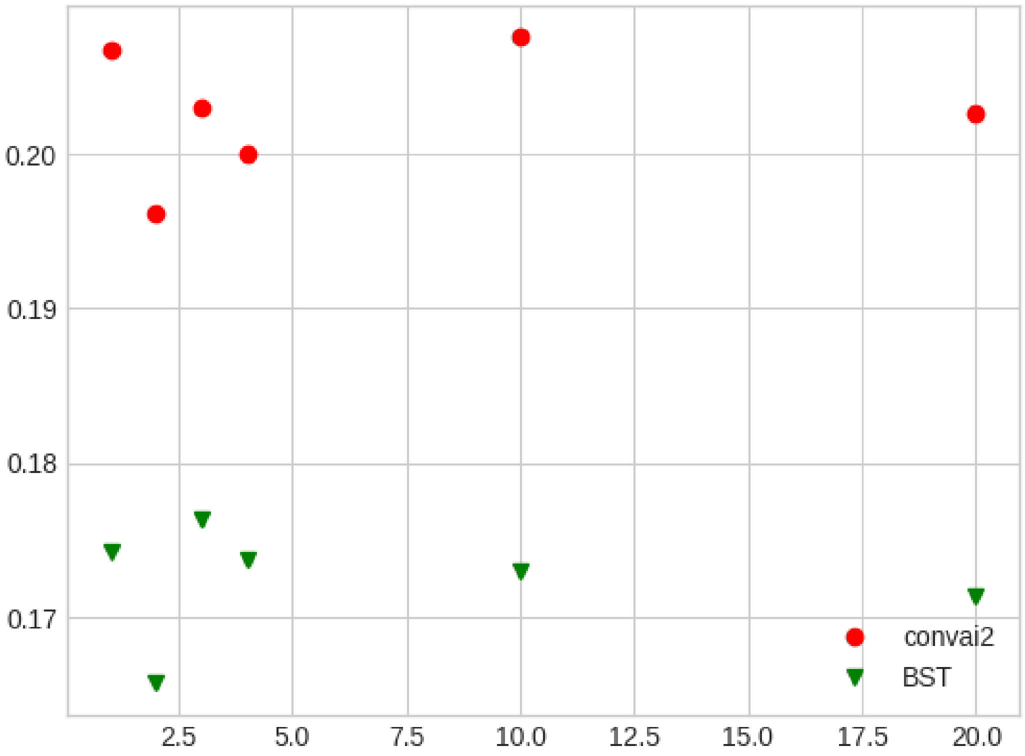
**Loss for ConvAI2 and BST over epochs**

### b. BLEU score over the epochs

| | Epoch 1 | Epoch 2 | Epoch 3 | Epoch 4 | Epoch 10 | Epoch 20 |
|---|---|---|---|---|---|---|
| ConvAI2 | 0.17349 | 0.1623 | 0.1682 | 0.1661 | 0.1728 | 0.1685 |
| BST | 0.14601 | 0.1386 | 0.1473 | 0.1464 | 0.1453 | 0.1446 |



**Bleu score for ConvAI2 and BST over epochs**



**F1 score for ConvAI2 and BST over epochs**

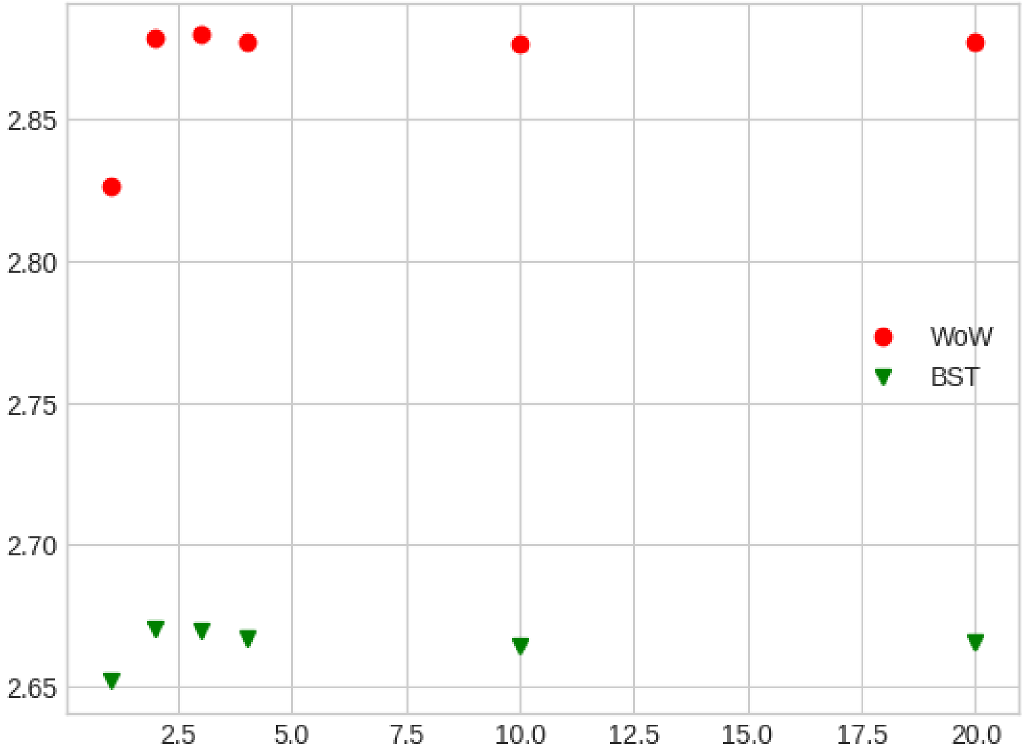
**Observations after finetuning with the ConvAI2 dataset**:

- 1 epoch is ideal for finetuning as the BLEU score for both datasets is high when finetuned for 1 epoch.

## 3. Finetuning with Wizard of Wikipedia dataset:

### a. Loss over the epochs

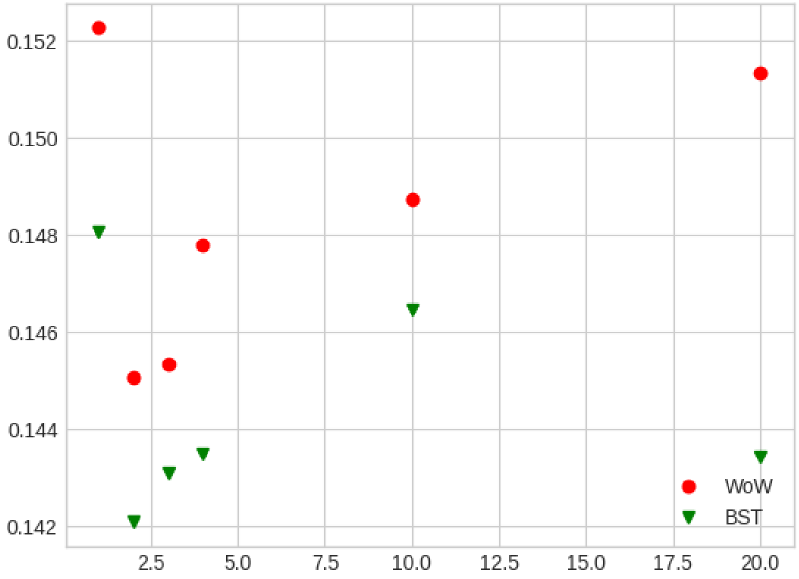
| | Epoch1 | Epoch 2 | Epoch 3 | Epoch 4 | Epoch 10 | Epoch 20 |
|---|---|---|---|---|---|---|
| WoW | 2.8262 | 2.8789 | 2.8798 | 2.8772 | 2.8766 | 2.8772 |
| BST | 2.6518 | 2.6701 | 2.6695 | 2.6667 | 2.6646 | 2.6657 |



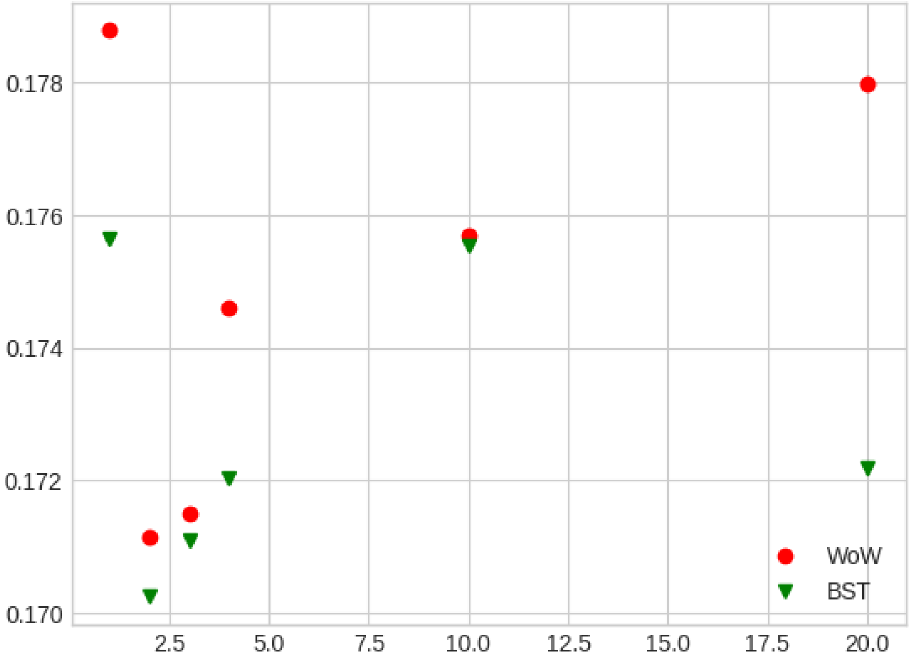
**Loss for Wizard of Wikipedia and BST over epochs**

### b. BLEU score over the epochs

| | Epoch 1 | Epoch 2 | Epoch 3 | Epoch 4 | Epoch 10 | Epoch 20 |
|---|---|---|---|---|---|---|
| WoW | 0.1522 | 0.1450 | 0.1453 | 0.1477 | 0.1487 | 0.1513 |
| BST | 0.1480 | 0.1420 | 0.1431 | 0.1434 | 0.1464 | 0.1434 |



**BLEU score for Wizard of Wikipedia and BST over epochs**



**F-1 score for Wizard of Wikipedia and BST over epochs**

**Observations after finetuning with the Wizard of Wikipedia dataset**:

- 1 epoch is ideal for finetuning as the BLEU score for both datasets is high when finetuned for 1 epoch.

4. **Finetuning with all datasets for previously determined ideal number of epochs:**
    - I finetuned the model for the determined ideal number of epochs that is 1 epoch for Wizard of Wikipedia, 1 epoch for ConvAI2 and 2 epochs for Daily dialog in sequence.

| Dataset | BLEU score on test set |
|---|---|
| Blended Skill Talk | 0.14676 |
| Daily Dialog | 0.1054 |
| Wizard of Wikipedia | 0.1447 |
| ConvAI2 | 0.1627 |

**Observations after finetuning with all datasets for determined epochs**:

- Balanced performance is obtained on all datasets.

## 5. Project Management

a. Hrishikesh Kanade: I am a Ph.D. CS student currently working with Dr. Enrique Dunn. I am working alone on this project.

    i. Milestones:

        **1.** Finetuning with Daily Dialog dataset by November 6th
        2. Finetuning with ConAI2 by November 16th
        3. Finetuning with Wizard of Wikipedia by November 30th
        4. Completion of report and presentation by December 2nd

## 6. Conclusion

- The model was finetuned on subsets of Daily Dialog, Wizard of Wikipedia, and ConvAI2 datasets.
- Metrics such as Loss, BLEU score, and F-1 score were observed to select the number of epochs for finetuning.
- The number of epochs for finetuning with each dataset was chosen such that there is a good balance between performance on the pretraining and finetuning dataset.
- In the final experiment the model was finetuned on all datasets sequentially for the previously determined number of epochs.
- All objectives of the project were met.

## 7. Key references

1. Roller, Stephen, et al. "Recipes for building an open-domain chatbot." *arXiv preprint arXiv:2004.13637* (2020).
2. Rashkin, Hannah, et al. "Towards empathetic open-domain conversation models: A new benchmark and dataset." *arXiv preprint arXiv:1811.00207* (2018).

3. Zhang, Saizheng, et al. "Personalizing dialogue agents: I have a dog, do you have pets too?." *arXiv preprint arXiv:1801.07243* (2018).
4. Dinan, Emily, et al. "Wizard of wikipedia: Knowledge-powered conversational agents." *arXiv preprint arXiv:1811.01241* (2018).
5. Smith, Eric Michael, et al. "Can you put it all together: Evaluating conversational agents' ability to blend skills." *arXiv preprint arXiv:2004.08449* (2020).