

Dear NeurIPS 2022 Workshop RobustSeq Program Chairs and Reviews,

We appreciate your time and effort on reviewing our paper. Your constructive comments are very helpful for improving our paper. The below are the responses to your comments:

[Reviewer 4nKq] “I’m curious if simple non-parametric OOD detection method such as KNN-based score [1] can be better suited than RMD for the tasks?”

Thanks for your suggestion. We compare KNN-based scores [1] with our proposed OOD scores and add the results in Section A.6. There are two hyper-parameters in the KNN-based method, α and k . α is the proportion of training data sampled for nearest neighbor calculation, and k refers to the k -th nearest neighbor. We use the optimal $k = 1000$ and $\alpha = 100$ as suggested by the paper. We also normalize the embedding features since the paper showed the feature normalization is critical for good performance. Table A.8 shows the AUROCs for OOD detection using the KNN-based method and comparing that with MD and RMD methods. As shown in the table, for the input OOD, KNN performs better than MD and RMD in `cnn_dailymail`, but worse than the two for other OOD datasets. For the output OOD, KNN is worse than both MD and RMD. It is possible that the optimal hyper-parameters suggested by the paper may not be the optimal ones for our problem, and a fine-grained hyper-parameter search could achieve better performance. We leave this for future study.

[1] Sun et al., Out-of-Distribution Detection with Deep Nearest Neighbors, ICML 2022.

	cnn_dailymail	newsroom	reddit_tifu	forumsum	samsum
Input OOD					
KNN (alpha=100%, k=1000)	0.887	0.743	0.944	0.961	0.955
MD	0.651	0.799	0.974	0.977	0.995
RMD	0.828	0.930	0.998	0.997	0.999
Output OOD					
KNN (alpha=100%, k=1000)	0.860	0.791	0.948	0.926	0.968
MD	0.944	0.933	0.985	0.973	0.985
RMD	0.958	0.962	0.998	0.993	0.998

[Reviewer 4nKq] “Besides quantitative results, as a generative task, showing some real examples can better demonstrate the generation quality after OOD scores are used.”

Thank you for the suggestion. We added a section A.7 showing a few examples in summarization to demonstrate how well our predicted quality score helps for selective generation.

[Reviewer JoLs] “The Geifman and El-Yaniv paper follows older work from that pair, in addition to Peter Bartlett (the first known selective classification paper to the best of my knowledge is *An optimum character recognition system using decision functions* (from 1957!!))”

The relevant citations are added (Chow 1957, Bartlett and Wegkamp 2008). Thanks for the pointers.

[Reviewer 17UN] “I am not sure if I understand the last part of section 2. After the two RMD scores are computed, how to decide whether each sample is in-domain or out-domain? Seems like contrasting the two RMD scores leads to the decision regarding whether each sample is OOD, but I’m wondering if there is a method to determine the threshold.”

The proposed RMD score is defined as the difference between the distance to the in-domain distribution and that to the general background distribution. So we regard it as a background contrastive score.

To use the RMD score for OOD detection, we can set a threshold based on the precision and recall requirement in real practice. For example, if in practice we want to conservatively eliminate as many OOD examples as possible, then we can choose a threshold at the 95% recall rate.

For method comparison, we use AUROC score between the in-domain test data as negative and the OOD test data as positive sets to evaluate and compare the OOD detection performance. AUROC 1.0 means a perfect separation, and 0.5 means the two are not distinguishable. AUROC is independent of the choice of threshold, so it can be used for fair comparisons among methods.

[Reviewer 17UN] “The idea that identifying and taking care of the out-of-domain samples helps to generate high-quality summarization or translations makes sense, but *how* these identified OOD samples are taken care of is not mentioned much in this paper. I understand that the space is very limited, but since the experiments show how the summarization & translation performances are improved, I would appreciate if there are more elaborations about that.”

We use OOD scores to help predict for which examples the model could generate low-quality outputs and then abstain from those examples. For those abstained examples, we prefer producing a message “Sorry we do not support a function for this input” than showing the bad model output. We call this procedure selective generation. Selective generation will help enable safer deployment of generative language models.