OUT-OF-DISTRIBUTION DETECTION AND SELECTIVE GENERATION FOR CONDITIONAL LANGUAGE MODELS

Jie Ren[1*] **Jiaming Luo**[1] **Yao Zhao**[1] **Kundan Krishna**[2]
Mohammad Saleh[1] **Balaji Lakshminarayanan**[1] **Peter J Liu**[1*]
[1]Google Research [2]Carnegie Mellon University, work done while at Google Research
[*]Correspondence to: {jjren, peterjliu}@google.com

ABSTRACT

Machine learning algorithms typically assume independent and identically distributed samples in training and at test time. Much work has shown that high-performing ML classifiers can degrade significantly and provide overly-confident, wrong classification predictions, particularly for out-of-distribution (OOD) inputs. Conditional language models (CLMs) are predominantly trained to classify the next token in an output sequence, and may suffer even worse degradation on OOD inputs as the prediction is done auto-regressively over many steps. Furthermore, the space of potential low-quality outputs is larger as arbitrary text can be generated and it is important to know when to trust the generated output. We present a highly accurate and lightweight OOD detection method for CLMs, and demonstrate its effectiveness on abstractive summarization and translation. We also show how our method can be used under the common and realistic setting of distribution shift for *selective generation* (analogous to selective prediction for classification) of high-quality outputs, while automatically abstaining from low-quality ones, enabling safer deployment of generative language models.

1 INTRODUCTION

Recent progress in generative language models (Wu et al., 2016a; Radford et al., 2019; Lewis et al., 2020; Raffel et al., 2020; Zhang et al., 2020) has led to quality approaching human-performance on research datasets and has opened up the possibility of their wide deployment beyond the academic setting. In realistic user-facing scenarios such as text summarization and translation, it should be expected that user provided inputs can significantly deviate from the training data distribution. This violates the independent, identically-distributed (IID) assumption commonly used in evaluating machine learning models.

Many have shown that performance of machine learning models can degrade significantly and in surprising ways on OOD inputs (Nguyen et al., 2014; Goodfellow et al., 2014; Ovadia et al., 2019). For example an image classifier may detect cows in images with very high accuracy on its IID test set but confidently fails to detect a cow when paired with an unseen background (Murphy, 2023; Nagarajan et al., 2020). This has led to active research on OOD detection for a variety of domains, including vision and text but focused primarily on classification. Salehi et al. (2021); Bulusu et al. (2020); Ruff et al. (2021) provide comprehensive reviews on this topic.

Conditional language models are typically trained given input sequence $\boldsymbol{x} = x_1 \dots x_L$ to auto-regressively generate the next token in a sequence $\boldsymbol{y} = y_1 \dots y_T$ as a classification over the token-vocabulary V, $p_\theta(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{T} p_\theta(y_t|y_{<t}, \boldsymbol{x})$, $y_t \in V$. Consequently, the perils of out-of-distribution are arguably more severe as (a) errors propagate and magnify through auto-regression, and (b) the space of low-quality outputs is greatly increased as arbitrary text sequences can be generated. Common errors from text generation models include disfluencies (Holtzman et al., 2020) and factual inaccuracies (Goodrich et al., 2019; Maynez et al., 2020). A common failure case we observed in abstractive summarization is for the model to output "All images are copyrighted" as the summary for news articles from a publisher (CNN) different than what it was trained on (BBC) (see Figure A.7).

In this work, we propose OOD detection methods for CLMs using abstractive summarization and translation as case studies. Similar to classification, we show in Section 2.1 that CLMs have untrustworthy likelihood estimation on OOD examples, making perplexity a poor choice for OOD

detection. In Section 2.2, we propose a highly-accurate, simple, and lightweight OOD score based on the model's input and output representations (or embeddings) to detect OOD examples, requiring negligible additional compute beyond the model itself.

While accurate OOD detection enables the conservative option of abstaining from generation on OOD examples, it may be desirable to generate on known near-domain data, e.g. generate summaries for articles from news publishers that differ from our fine-tuning set. Thus the ability to selectively generate where the model is more likely to produce higher-quality outputs, enables safer deployment of conditional language models. We call this procedure *selective generation*, analogous to the commonly used term *selective prediction* in classification (Chow, 1957; Bartlett & Wegkamp, 2008; Geifman & El-Yaniv, 2017). In Section 4, we show that while model perplexity is a reasonable choice for performing selective generation with in-domain examples, combining with our OOD score works much better when the input distribution is shifted.

In summary, our contributions are:

- Propose lightweight and accurate scores derived from a CLM's embeddings for OOD detection, significantly outperforming baselines on abstractive summarization and translation tasks, without the need for a separate detection model.
- Show that model perplexity can be an unreliable signal for quality estimation on OOD examples, but combined with our OOD scores can be used effectively to selectively generate higher-quality outputs while abstaining on lower ones.
- Propose an evaluation framework for OOD detection and selective generation for CLMs, including human quality ratings for summarization.

2 OOD Detection in Conditional Language Models

The maximum softmax probability (MSP), $p(y|\boldsymbol{x})$, $y = \arg\max_{k=1,\ldots,K} p(k|\boldsymbol{x})$ is a simple, commonly used OOD score for K-class classification problem (Hendrycks & Gimpel, 2016; Lakshminarayanan et al., 2017). For CLMs, the perplexity, which is monotonically related to the negative log-likelihood of the output sequence averaged over tokens $-\frac{1}{T}\sum_{t=1}^{T} \log p(y_t|y_{<t}, \boldsymbol{x})$ is a natural OOD score to consider, and analogous to the negative MSP in classification because both are based on softmax probabilities. We first study how well the perplexity performs for OOD detection tasks.

2.1 Perplexity is ill-suited for OOD detection

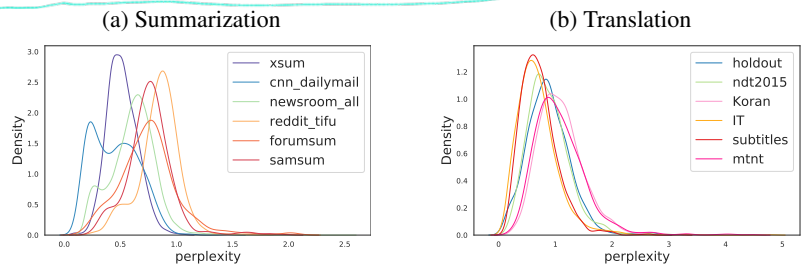


Figure 1: Perplexity scores density of a CLM trained on (a) xsum for summarization, and (b) WMT for translation, evaluated on other datasets/domains. Perplexity is not well suited for OOD detection due to significant overlap between in-domain and OOD scores.

In Figure 1, we compare the distribution of perplexity of (a) a summarization model and (b) a translation model trained on in-domain dataset and evaluated on multiple OOD datasets, respectively. For summarization, a model is trained on xsum and evaluated on other news datasets including cnn_dailymail and newsroom as near-OOD datasets, and forum (forumsum) and dialogue (samsum and reddit_tifu) datasets as far-OOD (see Section 3 for details). The perplexity distributions overlap significantly with each other even though the input documents are significantly different. Furthermore, perplexity assigns cnn_dailymail even lower scores than the in-domain xsum.

For translation, the model is trained on WMT15 dataset and evaluated on other WMT test splits (Bojar et al., 2015), OPUS100 (Aulamo & Tiedemann, 2019), and MTNT (Michel & Neubig, 2018). The in-domain and OOD datasets perplexity densities overlap even more. Overall, these results suggest that perplexity is not well suited for OOD detection.

2.2 DETECTING OOD USING CLM'S EMBEDDINGS

Given a trained conditional language model, we propose using the input and output representations/embeddings computed as part of the inference/generation process to detect OOD examples. In this work, we use Transformer encoder-decoder models and obtain the **input embedding** $\boldsymbol{z}$ by averaging the encoder's final-layer hidden state vectors $\boldsymbol{h}_i \in \mathbb{R}^d$ (d is the hidden dimension) corresponding to the input sequence token x_i. To obtain the **output embedding** $\boldsymbol{w}$ we average the decoder's final-layer hidden state vectors $\boldsymbol{g}_i \in \mathbb{R}^d$ corresponding to the output token y_i. Thus

$$\boldsymbol{z} := \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{h}_i \qquad \boldsymbol{w} := \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{g}_i, \qquad \boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^d$$

where L and T are the input and output sequence lengths respectively. Figure 2 illustrates the idea.

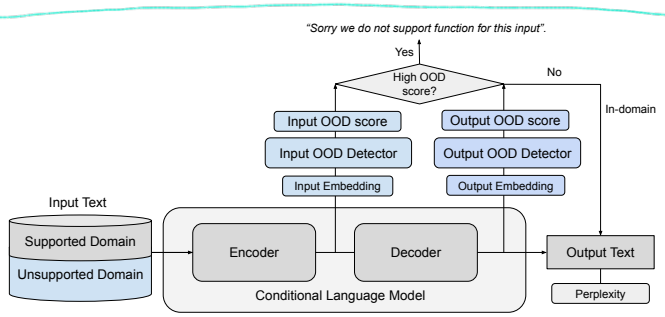


Figure 2: The proposed OOD detector based on input and output embeddings.

Intuitively, if the embedding of a test input or output is far from the embedding distribution of the training data, it is more likely to be OOD. One way of measuring this distance is to fit a Gaussian, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, to the training embeddings and use the *Mahalanobis distance (MD)*:

$$\text{MD}(x; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := (x-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (x-\boldsymbol{\mu}),$$

This has been used for OOD detection using the representations from classification models (Lee et al., 2018) and computing the distances to class-conditional Gaussians.

Unlike classification, which has class labels, in conditional language modeling we have paired input and output text sequences. We fit one Gaussian on the training input embeddings, $\mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{z}}, \boldsymbol{\Sigma}^{\boldsymbol{z}})$, and a second Gaussian on the embeddings of the training ground-truth outputs, $\mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{w}}, \boldsymbol{\Sigma}^{\boldsymbol{w}})$.

For a test input and output embedding pair $(\boldsymbol{z}_{\text{test}}, \boldsymbol{w}_{\text{test}})$, the input MD is computed as

$$\text{MD}_{\text{input}}(\boldsymbol{z}_{\text{test}}) := \text{MD}(\boldsymbol{z}_{\text{test}}; \boldsymbol{\mu}^{\boldsymbol{z}}, \boldsymbol{\Sigma}^{\boldsymbol{z}}) \qquad \textbf{(Input MD OOD score)}$$

The output MD is computed similarly:

$$\text{MD}_{\text{output}}(\boldsymbol{w}_{\text{test}}) := \text{MD}(\boldsymbol{w}_{\text{test}}; \boldsymbol{\mu}^{\boldsymbol{w}}, \boldsymbol{\Sigma}^{\boldsymbol{w}}) \qquad \textbf{(Output MD OOD score)}$$

Mahalanobis distance is equivalent to computing a negative log-likelihood of the Gaussian distribution (up to a constant and a scalar), i.e. $-\log p(\boldsymbol{z}) = \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) = \text{const.} + \frac{1}{2}\text{MD}(\boldsymbol{z})$. Ren et al. (2019) showed that normalizing the likelihood with the likelihood of a background model works better for OOD detection. In a similar vein, Ren et al. (2021) proposed an analogous **Relative Mahalanobis Distance (RMD)** for classification: using the relative distance between the class-conditional Gaussians and a single background Gaussian using data from all classes. That method cannot be directly applied for CLMs because outputs are not just class labels. Thus in this work, we extend the RMD idea to conditional language models,

$$\text{RMD}_{\text{input}}(\boldsymbol{z}_{\text{test}}) := \text{MD}_{\text{input}}(\boldsymbol{z}_{\text{test}}) - \text{MD}_0(\boldsymbol{z}_{\text{test}}), \qquad \textbf{(Input RMD OOD score)}$$

where $\text{MD}_0(\boldsymbol{z}_{\text{test}}) := \text{MD}(\boldsymbol{z}_{\text{test}}; \boldsymbol{\mu}_0^{\boldsymbol{z}}, \boldsymbol{\Sigma}_0^{\boldsymbol{z}})$ is the MD to a background Gaussian $\mathcal{N}(\boldsymbol{\mu}_0^{\boldsymbol{z}}, \boldsymbol{\Sigma}_0^{\boldsymbol{z}})$, fit using a large, broad dataset to approximately represent all domains. In practice, we use C4, a large Common Crawl-based English dataset (Raffel et al., 2020)[1] and ParaCrawl's English-French dataset

[1] https://www.tensorflow.org/datasets/catalog/c4

(Bañón et al., 2020)[2], as the data for fitting the background distributions for summarization and translation in our experiments, respectively.

While we use the ground-truth outputs to fit $\mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{w}}, \boldsymbol{\Sigma}^{\boldsymbol{w}})$, we decode outputs from the trained CLMs and use those output embeddings to fit the background output Gaussian, $\mathcal{N}(\boldsymbol{\mu}_\delta^{\boldsymbol{w}}, \boldsymbol{\Sigma}_\delta^{\boldsymbol{w}})$.

$$\text{RMD}_{\text{output}}(\boldsymbol{w}_{\text{test}}) := \text{MD}_{\text{output}}(\boldsymbol{w}_{\text{test}}) - \text{MD}_\delta(\boldsymbol{w}_{\text{test}}), \qquad \textbf{(Output RMD OOD score)}$$

where $\text{MD}_\delta(\boldsymbol{w}_{\text{test}}) := \text{MD}(\boldsymbol{w}_{\text{test}}; \boldsymbol{\mu}_\delta^{\boldsymbol{w}}, \boldsymbol{\Sigma}_\delta^{\boldsymbol{w}})$ is the MD to the decoded output background distribution $\mathcal{N}(\boldsymbol{\mu}_\delta^{\boldsymbol{w}}, \boldsymbol{\Sigma}_\delta^{\boldsymbol{w}})$. See Algorithm 1 and 2 for the detailed steps. Using decoded outputs serves two purposes: (1) We do not require supervised data (e.g. document-summary pairs) to fit the background Gaussian. (2) Decoded outputs may exhibit increased deficiencies that result from running the model on out-of-distribution data, which provides greater contrast with the in-domain ground-truth labels.

The RMD score can be regarded as a background contrastive score that indicates how close the test example is to the training domain compared to the background domains. A negative score suggests the example is relatively in-domain, while a positive score suggests the example is OOD. A higher score indicates greater OOD-ness.

Binary classifier for OOD detection Since we have explicitly defined two classes, in-domain and background/general domain, another option is to train a binary classifier to discriminate embeddings from the two classes. We train a logistic regression model and use the un-normalized logit for the background as an OOD score. The **Input Binary logits OOD score** uses the input embeddings as features, whereas the **Output Binary logits OOD score** uses the decoded output embeddings as features. A higher score suggests higher likelihood of OOD. The preferred use of the logits over probability was also recommended by previous OOD studies for classification problems (Hendrycks et al., 2019). Though RMD is a generative-model based approach and the binary classifier is a discriminative model, we show that RMD is a generalized version of binary logistic regression and can be reduced to a binary classification model under certain conditions (see Section A.5 for details).

3 Experiments: OOD detection

3.1 Experiment setup

We run our experiments using Transformer (Vaswani et al., 2017) encoder-decoder models trained for abstractive summarization and translation. Below we specify the dataset used for training/fine-tuning (i.e. in-domain) and the OOD datasets.

In the case of summarization, OOD datasets can be intuitively categorized as *near* or *far OOD* based on the nature of the documents. For example, news articles from different publishers may be considered as sourced from different distributions, but are closer than news articles are to dialogue transcripts. We also quantitatively showed that using n-gram overalp analysis in Table A.10. In contrast, the translation datasets we use consist of English-French sentence pairs with less variation between datasets due to the shorter length of sentences.

Summarization model We fine-tuned $\text{PEGASUS}_{\text{LARGE}}$ (Zhang et al., 2020) on the xsum (Narayan et al., 2018) dataset, consisting of BBC News articles with short, abstractive summaries.

Summarization datasets We use 10,000 examples from xsum and C4 training split to fit in-domain/foreground and background Gaussian distributions, respectively. For test datasets, we have cnn_dailymail (Hermann et al., 2015; See et al., 2017), news articles and summaries from CNN and DailyMail; newsroom (Grusky et al., 2018), article-summary pairs from 38 major news publications; reddit_tifu (Kim et al., 2018), informal stories from sub-reddit TIFU with author written summaries of very diverse styles; samsum (Gliwa et al., 2019) and forumsum (Khalman et al., 2021), high-quality summaries of casual dialogues.

Translation model We train a Transformer base model (Vaswani et al., 2017) with embedding size 512 on WMT15 English-French (Bojar et al., 2015). The model is trained with Adafactor optimizer (Shazeer & Stern, 2018) for 2M steps with 0.1 dropout and 1024 batch size. Decoding is done using beam search with 10 beam size and $\alpha = 0.6$ length normalization (Wu et al., 2016b). The best checkpoint scores 39.9 BLEU on newstest2014.

[2] https://www.tensorflow.org/datasets/catalog/para_crawl

Table 1: AUROCs for OOD detection. For summarization task (a), cnn_dailymail and newsroom are considered as near shift OOD since it shares news topics as xsum, and reddit_tifu, forumsum, and samsum are far shift OOD. For translation (b), WMT dataset contains various test WMT datasets collected from different years, OPUS contains five different domains (the degree of shift varies), and MTNT contains noisy data from Reddit.

(a) Summarization

	Near Shift OOD		Far Shift OOD		
Measure	cnn_dailymail	newsroom	reddit_tifu	forumsum	samsum
			INPUT OOD		
MD	0.651	0.799	0.974	0.977	0.995
RMD	0.828	0.930	0.998	0.997	**0.999**
Binary logits	**0.997**	0.959	**1.000**	**0.999**	0.998
			OUTPUT OOD		
Perplexity (baseline)	0.424	0.665	0.909	0.800	0.851
NLI score (baseline)	0.440	0.469	0.709	0.638	0.743
MD	0.944	0.933	0.985	0.973	0.985
RMD	0.958	0.962	0.998	0.993	0.998
Binary logits	0.989	**0.982**	**1.000**	0.998	0.997

(b) Translation

	WMT			OPUS					MTNT
Measure	nt2014	ndd2015	ndt2015	law	medical	Koran	IT	sub	
				INPUT OOD					
MD	0.534	0.671	0.670	0.511	0.704	0.737	0.828	0.900	0.668
RMD	0.798	0.866	0.863	0.389	0.840	0.957	**0.959**	**0.969**	0.943
Binary logits	**0.864**	**0.904**	**0.904**	0.485	0.813	**0.963**	0.928	0.950	**0.963**
				OUTPUT OOD					
Perplexity (baseline)	0.570	0.496	0.494	0.392	0.363	0.657	0.343	0.359	0.633
COMET (baseline)	0.484	0.514	0.525	0.435	0.543	0.632	0.619	0.518	0.724
Prism (baseline)	0.445	0.504	0.505	0.459	0.565	0.716	0.604	0.577	0.699
MD	0.609	0.733	0.739	0.482	0.784	0.838	0.900	0.935	0.794
RMD	0.786	0.858	0.861	0.355	**0.845**	0.939	0.951	0.959	0.922
Binary logits	0.822	0.860	0.865	0.507	0.783	0.942	0.890	0.910	0.931

Translation datasets We use 100,000 examples from WMT15 En-Fr and the same number of examples from ParaCrawl En-Fr to fit the foreground and background Gaussians, respectively. For test, we use newstest2014 (nt14), newsdiscussdev2015 (ndd15), and newsdiscusstest2015 (ndt15) from WMT15 (Bojar et al., 2015) and the law, Koran, medical, IT, and subtitles (sub) subsets from OPUS (Tiedemann, 2012; Aulamo & Tiedemann, 2019). We also use the English-French test set of MTNT (Michel & Neubig, 2018), consisting of noisy comments from Reddit.

Evaluation metric We use the area under the ROC curve (AUROC) 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.

Baseline methods We compare our proposed OOD scores with various baseline methods, including (1) the model perplexity score, (2) the embedding-based Mahalanobis distance. In addition, we also compare with (3) Natural Language Inference (NLI) score (Honovich et al., 2022) for summarization, and (4) COMET (Rei et al., 2020) and (5) Prism (Thompson & Post, 2020) for translation. NLI score measures the factual consistency by treating the input document as a premise and the generated summary as a hypothesis. Both COMET and Prism are quality estimation metrics designed to measure translation quality without access to a human reference. More specifically, COMET finetunes the large XLM-R model (Conneau et al., 2020) on human evaluation data, and Prism is the perplexity score from a multilingual NMT model trained on 99.8M sentence pairs in 39 languages.

3.2 RESULTS

RMD and Binary classifier are better at OOD detection than baselines Table 1 shows the AUROCs for OOD detection on the (a) summarization and (b) translation datasets. Overall, our proposed OOD scores RMD and Binary logits outperform the baselines with high AUROCs (above 0.8). The commonly used output metrics, perplexity, NLI, COMET and Prism, have generally low AUROC scores (many have values around 0.5-0.6), suggesting they are not suited for OOD detection. Interestingly, we noticed that the output OOD scores perform better for summarization, while the input OOD scores perform better for translation. One possible reason is that when summariza-

tion outputs are low-quality (e.g. producing repeated text or irrelevant summaries) they look very different than reference summaries, making OOD output score more sensitive to the contrast.

Though RMD and Binary logits OOD scores both perform well at OOD detection, **RMD OOD score is better at distinguishing near-OOD from far-OOD**. This can be seen in Figure 3 where near-OOD datasets have scores distributed in between in-domain and far-OOD. In the summarization task, near-OOD (news articles) datasets cnn_dailymail and newsroom have their RMD scores distributed in the middle of xsum and reddit_tifu, forumsum and samsum. In contrast, under the binary logits score, the near-OOD and far-OOD datasets have largely overlapping score distributions making it hard to distinguish between the two. In practice, RMD OOD score may be better suited for selective generation where domain shifts are expected. We explore this in more detail in Section 4.

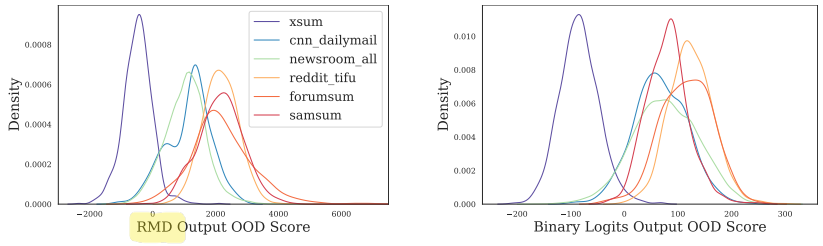


Figure 3: Density of RMD (left) and Binary logits (right) OOD scores evaluated on summarization datasets. RMD is better at distinguishing near-OOD from far-OOD.

For the translation task, we additionally note that all methods have small AUROC for law dataset, suggesting that none of the methods are detecting the dataset as OOD. To better understand the special characteristics of the law dataset, we conducted an n-gram overlap analysis between the various test sets including law and the in-domain training data. We observed that law has the highest unigram overlap rate (48.8%) and the second highest overall overlap with the in-domain data (Table A.9).[3] This shows that law is close to in-domain data in terms of surface features, which might contribute to the low AUROC scores for all tested methods.

We use ParaCrawl instead of C4 for translation because our translation model is trained on the sentence level, unlike the summarization model that takes the document as input. To further explore the effect of the background data on the performance, we split C4 documents into sentences and use that as the background data to compute the scores. The OOD detection performance using C4 sentences is very similar to that using ParaCrawl, as shown in Table A.3, suggesting that our method is not particularly sensitive to the choice of background data.

4 Using OOD scores for selective generation

The most conservative option for deployment of a conditional language model is to completely abstain from generating on inputs that are detected as out-of-distribution, for which we have shown in Section 3 our OOD scores are fairly accurate. However, it is often desirable to expand the use of models beyond strictly in-distribution examples, if the quality of outputs is sufficiently high. In classification, this has been framed as determining when to trust a classifier, or *selective prediction* (Geifman & El-Yaniv, 2017; Lakshminarayanan et al., 2017; Tran et al., 2022). In this section, we seek to predict the quality of generation given an example, which may be out-of-distribution and *abstain* if the predicted quality is low. We call this *selective generation*. In practice, abstaining may correspond to hiding the model's generated text, or turning off a summarization/translation feature.

4.1 Experiment setup

We use the same models and datasets described in Section 3.1 but instead of simply detecting out-of-distribution examples, our focus now is to *predict the quality of generation* for examples possibly outside the training distribution.

[3] We define overlap rate as the percentage of unique n-grams in the test set that are also present in the in-domain data. The overall overlap is defined as the geometric mean of all the n-gram overlap rates up to $n = 4$. All domains/splits including the in-domain data are subsampled to 1K for this analysis.

Measuring Translation quality We use BLEURT (Pu et al., 2021) as the main metric to measure translation quality. Previous work has demonstrated that neural metrics such as BLEURT are much better correlated with human evaluation, on both the system level and the sentence level (Freitag et al., 2021). BLEURT scores range from 0 to 1, with higher scores indicating better quality.

Measuring Summarization quality In general, it is unclear how to automatically measure the quality of summaries generated by a model on out-of-distribution examples (in this case, examples from different datasets). The reason is summarization datasets have dataset-specific summary styles that may be difficult to compare. For example, xsum summaries are typically single-sentence whereas cnn_dailymail summaries consist of multiple sentences. Thus we report ROUGE-1 score as an automatic measure but primarily use human evaluation to assess the quality. Amazon Mechanical Turk workers were asked to evaluate summaries generated by the xsum model on a scale of 1-5 (bad-good) using 100 examples from xsum, cnn_dailymail, reddit_tifu, and samsum. We collected 3 ratings per example and computed the median. See Section A.3 for more details.

4.2 Perplexity has diminishing capability in predicting quality on OOD data

Since the models are trained using negative log-likelihood as the loss, perplexity (which is monotonically related) is a good predictor of output quality for in-domain data. In fact, the Kendall rank correlation coefficient τ between perplexity and human judged quality score is 0.256 (See Table 2) for in-domain xsum for summarization. However, when including shifted datasets to test, we found that the perplexity score is worse at predicting quality on OOD data. For example the Kendall's τ decreases to 0.068 for OOD dataset samsum (see Table A.4). We observed similar trend in translation, although less severe, as data shifted from in-domain to OOD, the Kendall's τ between perplexity and BLEURT decreases (see Table A.5). Figure 4 further shows the correlation between perplexity and the quality score (ROUGE-1, human rating, and BLEURT, respectively) as a function of OOD score. It is clear to see the correlation decreasing as OOD score increases and the trend is consistent for both summarization and translation.

(a) Summarization, ROUGE-1 (b) Summarization, human rating (c) Translation, BLEURT

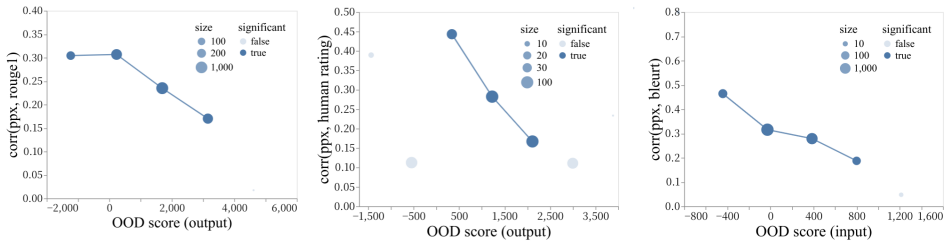


Figure 4: The Kendall's τ correlation between perplexity and (a) ROUGE-1, (b) human evaluation median rating, and (c) BLEURT decreases as OOD score increases respectively. Note that we use output RMD OOD score for summarization and input RMD OOD score for translation.

4.3 Combining OOD scores and perplexity

While model perplexity for quality estimation is worse for OOD examples, we observed that our OOD scores and perplexity are complementary in quality prediction. Figure A.1 shows a 2-D plot between the OOD score and perplexity regarding quality. We can see that neither perplexity nor OOD score can perfectly separate good and bad examples, and the combination of the two can work much better. Our observation echos work in uncertainty estimation in classification models (Mukhoti et al., 2021): perplexity based on softmax predictive distribution is regarded as an estimation for *aleatoric* uncertainty (caused by inherent noise or ambiguity in data), and the OOD distance based on representation estimates the *epistemic* uncertainty (caused by a lack of training data), and combining the two provides a comprehensive estimation of uncertainty.

We propose two simple methods to combine perplexity and OOD scores. (1) A simple linear regression, trained on a random 10% data split using ROUGE-1 or BLEURT as the quality score, and evaluated on the test split and human evaluation split. (2) the sum of the percentile ranks (PR) of the scores, i.e. $\text{PR}_{\text{sum}} = \text{PR}_{\text{perplexity}} + \text{PR}_{\text{OOD}}$. We sum PRs instead of their raw values because the two scores are in different ranges, $\text{PR}(x) = \frac{R(x)}{N} \times 100$, where $R(x)$ is x's rank in the list of size N.

Table 2 shows the Kendall's τ correlation coefficient between the various single and combined scores and the quality metric with only in-domain and all examples from all datasets. When all datasets

Table 2: Kendall's τ correlation (p-value < 0.05 are grayed out) between various measures and human evaluation for summarization and BLEURT for translation. The "All" column shows the correlation when both in-domain and OOD examples are merged. Note for negatively correlated scores (e.g. perplexity (ppx), RMD), we take the negative value of the score for easier comparison.

(a) Summarization

Measure	In-domain	All
Single Score		
Perplexity (baseline)	0.256	0.300
NLI score (baseline)	0.337	0.381
Input RMD	0.015	0.336
Output RMD	0.053	0.385
Combined Score		
PR_{sum}(ppx, input RMD)	0.186	0.358
PR_{sum}(ppx, output RMD)	0.250	0.415
Linear Reg. (ppx, input & output)	0.235	**0.422**

(b) Translation

Measure	In-domain	All
Single Score		
Perplexity (baseline)	0.309	0.286
COMET (baseline)	0.184	0.336
Prism (baseline)	0.184	0.301
Input RMD	0.147	0.195
Output RMD	0.086	0.170
Combined Score		
PR_{sum}(ppx, input RMD)	0.321	**0.361**
PR_{sum}(ppx, output RMD)	0.323	0.356
Linear Reg. (ppx, input & output)	0.318	0.352

are merged, **the combined scores significantly improve the correlation over perplexity by up to 12% (absolute) for summarization and 8% for translation, while the gains over the best external model-based (and much more expensive) baselines are 4% and 3%**. The two combination methods perform similarly. See Tables A.4 and A.5 for an expanded table of scores.

4.4 Selective Generation using the Combined Score

In selective generation, our goal is to generate when the model is more likely to produce high-quality output, and *abstain* otherwise, enabling safer deployment of generative language models. To evaluate that, we propose using the ***Quality vs Abstention Curve*** **(QA)**, analogous to accuracy versus rejection curve used for selective prediction in the classification (Chow, 1957; Bartlett & Wegkamp, 2008; Geifman & El-Yaniv, 2017). Similar concepts were proposed also in Malinin & Gales (2020); Xiao et al. (2020), but they only use automatic quality metrics for the analysis while we consider human evaluation to assess the quality as well. Specifically, at a given abstention rate α, the highest α-fraction scoring examples are removed and the average quality of remaining examples is computed. We want to maximize the quality of what is selectively generated and a better curve is one that tends to the upper-left which corresponds to removing bad examples earlier than good ones.

Figure 5 shows the QA curves for various methods on summarization and translation. Quality is measured by human evaluation for summarization (see Figure A.4 for similar ROUGE-1 plot), and BLEURT for translation. **The combined scores have the highest quality score at almost all abstention rates for both summarization and translation**, while linear regression and PR_{sum} perform similarly. For single scores, the OOD score performs better than perplexity and NLI scores at almost all abstention rates for summarization. For translation, the OOD score is better than perplexity when abstention rate $\alpha > 0.65$ and worse than perplexity when $\alpha < 0.65$. In other words, OOD score is better at abstaining slightly far-OOD while perplexity is better at abstaining near-OOD examples. Interestingly, our combined score is even marginally better than COMET that requires a separate neural network trained on human evaluation data. Prism is better than single scores, but much worse than our combined score. Area under the QA curves are shown in Tables A.6 and A.8 for reference.

Figures 5 (b, d) are the corresponding survival curves showing how many examples per dataset are selected for generation as a function of abstention rate, based on the PR_{sum} score. For summarization, the samples from far-OOD datasets reddit_tifu and samsum are eliminated first with their sample count decreasing rapidly. The near-OOD dataset cnn_dailymail and in-domain xsum are kept intact until $\alpha > 0.3$, and in-domain xsum examples survive the longest. Similarly for translation, the out-of-domain and worst-quality (as seen in Table A.1b) Koran, MTNT, and subtitles examples are eliminated first, and the best-performing law and in-domain datasets are abstained last. **The order in which datasets are eliminated corresponds to the aggregate quality by dataset**, which we report in Table A.1. Besides the quantitative results, we show a few real examples in Section A.14 to better demonstrate how our predicted quality score helps selective generation.

5 Related Work

OOD detection problem was first proposed and studied in vision classification problems (Hendrycks & Gimpel, 2016; Liang et al., 2017; Lakshminarayanan et al., 2017; Lee et al., 2018; Hendrycks et al., 2018; 2019), and later in text classification problems such as sentiment analysis (Hendrycks

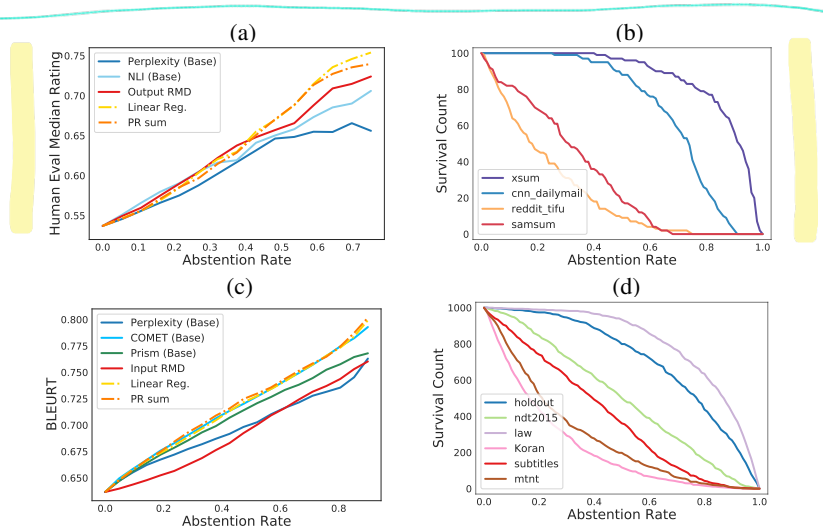


Figure 5: (a) The Quality (human eval) vs Abstention curve for summarization. Combined scores have the highest quality at almost all abstention rates. (b) Survival count of each dataset as a function of abstention rate, using PR_{sum} (we use output/input RMD for summarization/translation to pair with perplexity). OOD data is abstained earlier than in-domain. (c, d) The same as (a, b) for translation.

et al., 2020), natural language inference (Arora et al., 2021), intent prediction (Liu et al., 2020a; Tran et al., 2022), and topic prediction (Rawat et al., 2021). The widely used OOD methods can be characterized roughly into two categories (1) softmax probability or logits-based scores (Hendrycks & Gimpel, 2016; Liang et al., 2017; Hendrycks et al., 2019; Liu et al., 2020b), (2) embedding-based methods that measure the distance to the training distribution in the embedding space (Lee et al., 2018; Ren et al., 2021; Sun et al., 2022), (3) contrastive learning based methods which incorporate the contrastive loss into the classification cross-entropy loss to improve representation learning and consequently improve OOD detection (Winkens et al., 2020; Zhou et al., 2021). Though it is not straightforward to extend those classifier-based scores to CLMs especially for input OOD detection, we extend three of them based on our understanding as baselines for comparison with our methods. See Section A.6 for details. The results in Table A.2 show that those methods are in general not competitive with our proposed methods RMD and Binary logits, especially on near-OOD datasets.

OOD detection problem is less studied in CLMs. A few studies explored OOD detection in semantic parsing (Lukovnikov et al., 2021; Lin et al., 2022), speech recognition (Malinin & Gales, 2020), and machine translation (Malinin et al., 2021; Xiao et al., 2020), but many of them focus on ensemble-based methods like Monte Carlo dropout or deep ensemble which use the averaged perplexity after sampling multiple output sequences.The ensembling method costs N times of the inference time, which is not feasible in practice. In this work, we focus on developing scores that can be readily derived from the generative model itself, without much increase in computation. We include an ensemble-based baseline in Section A.6 and show that its performance is worse than our methods.

6 Conclusion and Future Work

We have proposed lightweight and accurate scores to detect out-of-distribution examples for conditional language generation tasks. For real-world deployment, we have also shown how our OOD scores can be combined with language model perplexity to selectively generate high-quality outputs while abstaining from low-quality ones in the setting of input distribution shift.

Although our experiments focus on summarization and translation, our methods do not make any assumptions about the task modality, and we believe our method is widely applicable to other tasks where the model output is a sequence, e.g. image captioning. While our analysis was restricted to conditional language modeling with encoder-decoder Transformers, we expect our method to also work with decoder-only (Liu et al., 2018) architectures, used by some large language models such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and LaMDA (Thoppilan et al., 2022).

Finally, analyzing why certain examples are OOD could lead to insights in how to make models more robust. Section A.13 presents one possible way to attribute OOD scores to sentences.

ACKNOWLEDGEMENTS

The authors would like to thank Jeremiah Zhe Liu, Sharat Chikkerur, and the anonymous reviewers for their helpful feedback on the manuscript. The authors would also like to thank Colin Cherry, George Foster, and Polina Zablotskaia for their feedback throughout the project.

REFERENCES

Udit Arora, William Huang, and He He. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*, 2021.

Mikko Aulamo and Jörg Tiedemann. The OPUS resource repository: An open package for creating parallel corpora and machine translation services. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 389–394, Turku, Finland, September–October 2019. Linköping University Electronic Press. URL https://aclanthology.org/W19-6146.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL https://aclanthology.org/2020.acl-main.417.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 1–46, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3001. URL https://aclanthology.org/W15-3001.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K Varshney, and Dawn Song. Anomalous example detection in deep learning: A survey. *IEEE Access*, 8:132330–132347, 2020.

Chi-Keung Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 733–774, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.73.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, pp. 166–175, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330955. URL https://doi.org/10.1145/3292500.3330955.

Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018. doi: 10.18653/v1/n18-1065. URL http://dx.doi.org/10.18653/v1/n18-1065.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pp. 1693–1701, 2015. URL http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3905–3920, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287.

Misha Khalman, Yao Zhao, and Mohammad Saleh. Forumsum: A multi-speaker conversation summarization dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4592–4599, 2021.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of reddit posts with multi-level memory networks, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS*, 2018.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 4160–4173, 2022.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020a.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hyg0vbWC-.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020b.

Denis Lukovnikov, Sina Daubener, and Asja Fischer. Detecting compositionally out-of-distribution examples in semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 591–598, 2021.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173.

Paul Michel and Graham Neubig. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 543–553, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1050. URL https://aclanthology.org/D18-1050.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A simple baseline. *arXiv e-prints*, pp. arXiv–2102, 2021.

Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL probml.ai.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206.

A Nguyen, J Yosinski, and J Clune. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. arxiv. *arXiv preprint arXiv:1412.1897*, 2014.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 751–762, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.58. URL https://aclanthology.org/2021.emnlp-main.58.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Mrinal Rawat, Ramya Hebbalaguppe, and Lovekesh Vig. Pnpood: Out-of-distribution detection for text classification via plug andplay data augmentation. *arXiv preprint arXiv:2111.00506*, 2021.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*, 2020.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *NeurIPS*, 2019.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://aclanthology.org/P17-1099.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022.

Brian Thompson and Matt Post. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. *arXiv preprint arXiv:2004.14564*, 2020.

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

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Dustin Tran, Jeremiah Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, et al. Plex: Towards reliability using pretrained large model extensions. *arXiv preprint arXiv:2207.07411*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016a. URL http://arxiv.org/abs/1609.08144.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016b.

Tim Z Xiao, Aidan N Gomez, and Yarin Gal. Wat zei je? detecting out-of-distribution translations with variational transformers. *arXiv preprint arXiv:2006.08344*, 2020.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. Contrastive out-of-distribution detection for pre-trained transformers. *arXiv preprint arXiv:2104.08812*, 2021.

A APPENDIX

A.1 THE OUTPUT QUALITY FOR SUMMARIZATION AND TRANSLATION DATASETS.

Table A.1: The output quality for summarization and translation datasets. (a) Summarization quality (higher is better) for xsum model. ROUGE-1 is based on all samples in the test split per dataset, while human evaluation is based on 100 samples. The raw human evaluation rating ranges from 1 to 5. We normalized the score by dividing 5.0, and toke the median of the ratings over 3 raters to reduce inter-rater noise. The standard deviation among 3 ratings are reported in brackets. (b) Translation quality for different datasets (higher is better). All datasets are sub-sampled to 1000 sentence pairs.

(a) Summarization

Dataset	ROUGE-1	Human evaluation
xsum	0.474	0.698 (0.182)
cnn_dailymail	0.226	0.624 (0.145)
reddit_tifu	0.140	0.450 (0.152)
samsum	0.210	0.376 (0.147)

(b) Translation

Dataset	BLEURT	BLEU
law	0.781	53.8
nt2014	0.731	39.8
holdout	0.674	41.8
ndt2015	0.671	37.9
ndd2015	0.664	30.9
medical	0.643	34.2
IT	0.588	28.3
MTNT	0.565	32.0
sub	0.552	22.8
Koran	0.491	12.9

A.2 OOD SCORE AND PERPLEXITY ARE COMPLEMENTARY FOR PREDICTING OUTPUT QUALITY.

(a) Summarization, ROUGE-1 (b) Summarization, human rating (c) Translation, BLEURT

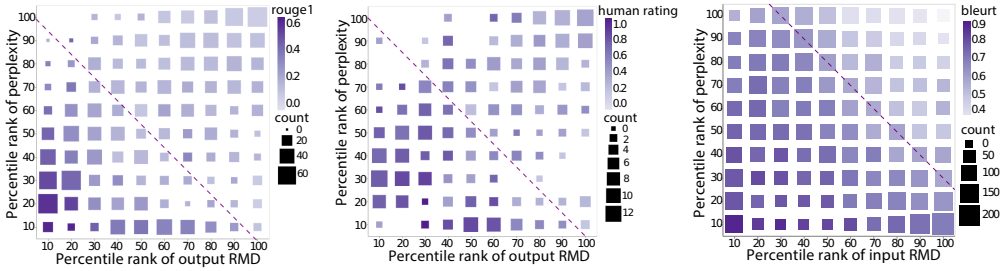


Figure A.1: 2D plot between OOD and perplexity. The two scores are self-normalized by its percentile rank respectively. Each square corresponds to a subset of samples whose OOD and perplexity scores are within the percentile bin. The size of the square represents the size of the bin where the color indicates the quality of the model's output. The OOD score and perplexity capture different properties of model outputs, and combining both scores can be beneficial for quality prediction.

A.3 AMAZON MECHANICAL TURK ASSESSMENT OF SUMMARY QUALITY

A PEGASUS$_{\text{LARGE}}$ model fine-tuned on xsum was run on a random sample of 100 examples from the test split of four datasets: xsum, cnn_dailymail, reddit_tifu, samsum. Each example was rated for general summarization quality on a rating of 1-5 by 3 AMT workers using the template shown in Figure A.2. Workers were required to be Masters located in the US with greater than 95% HIT Approval Rate, with at least 1000 HITs approved and were paid $0.80 per rating.

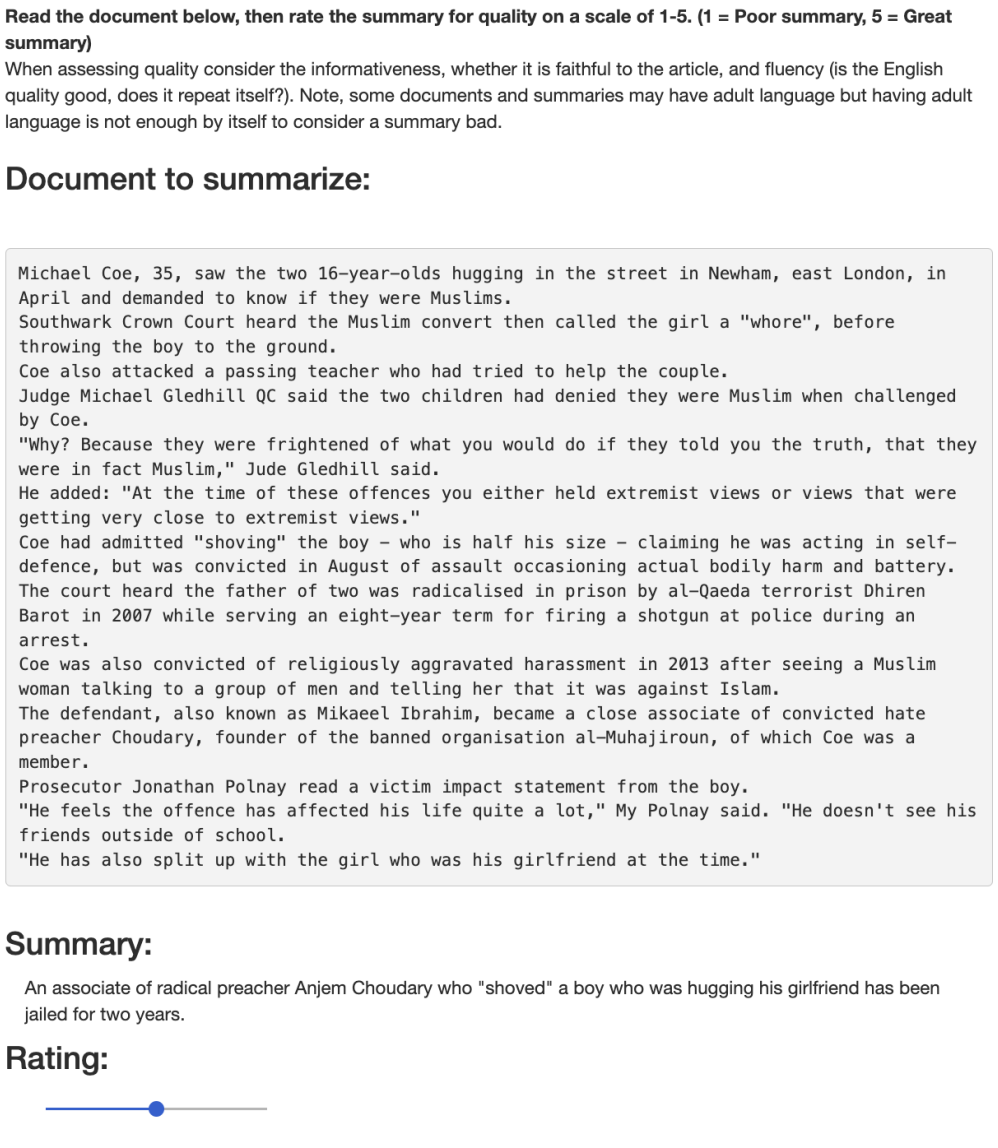


Figure A.2: AMT template for summarization human evaluation.

A.4 ALGORITHM FOR RMD OOD SCORES

Algorithm 1 Fitting Gaussians for input and output embeddings

1: **Input:** CLM M with encoder f_e and decoder f_d trained on in-domain train set $\mathcal{D}^{\text{in}}_{\text{train}} = \{(\boldsymbol{x}, \boldsymbol{y})\}$. A large and background dataset such as C4 or ParaCrawl $\mathcal{D}^{\text{bg}}_{\text{train}} = \{(\boldsymbol{x}, \hat{\boldsymbol{y}})\}$, where $\hat{\boldsymbol{y}} = M(\boldsymbol{x})$.
2: Generate the input embeddings $\mathcal{S}^{\text{in}}_{\text{train}} = \{\boldsymbol{z}|f_e(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{D}^{\text{in}}_{\text{train}}\}$ and $\mathcal{S}^{\text{bg}}_{\text{train}} = \{\boldsymbol{z}|f_e(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{D}^{\text{bg}}_{\text{train}}\}$.
3: Fit a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{z}}, \boldsymbol{\Sigma}^{\boldsymbol{z}})$ using $\mathcal{S}^{\text{in}}_{\text{train}}$, and a background Gaussian $\mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{z}}_0, \boldsymbol{\Sigma}^{\boldsymbol{z}}_0)$ using $\mathcal{S}^{\text{bg}}_{\text{train}}$.
4: Similarly, generate output embeddings $\mathcal{E}^{\text{in}}_{\text{train}} = \{\boldsymbol{w}|f_d(\boldsymbol{y}), \boldsymbol{y} \in \mathcal{D}^{\text{in}}_{\text{train}}\}$, and $\mathcal{E}^{\text{bg}}_{\text{train}} = \{\boldsymbol{w}|f_d(\hat{\boldsymbol{y}}), \hat{\boldsymbol{y}} \in \mathcal{D}^{\text{bg}}_{\text{train}}\}$.
5: Fit a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{w}}, \boldsymbol{\Sigma}^{\boldsymbol{w}})$ using $\mathcal{E}^{\text{in}}_{\text{train}}$ and a background Gaussian $\mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{w}}_\delta, \boldsymbol{\Sigma}^{\boldsymbol{w}}_\delta)$ using $\mathcal{E}^{\text{bg}}_{\text{train}}$.

Algorithm 2 OOD score inference

1: **Input:** In-domain test set $\mathcal{D}^{\text{in}}_{\text{test}} = \{(\boldsymbol{x}, \hat{\boldsymbol{y}})\}$. OOD test set $\mathcal{D}^{\text{ood}}_{\text{test}} = \{(\boldsymbol{x}, \hat{\boldsymbol{y}})\}$, where $\hat{\boldsymbol{y}} = M(\boldsymbol{x})$.
2: Generate input embeddings $\mathcal{S}^{\text{in}}_{\text{test}} = \{\boldsymbol{z}|f_e(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{D}^{\text{in}}_{\text{test}}\}$ and $\mathcal{S}^{\text{ood}}_{\text{test}} = \{\boldsymbol{z}|f_e(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{D}^{\text{ood}}_{\text{test}}\}$.
3: Compute input OOD score $\text{RMD}_{\text{input}}(\boldsymbol{z})$ for $\boldsymbol{z} \in \mathcal{S}^{\text{in}}_{\text{test}}$ and $\mathcal{S}^{\text{ood}}_{\text{test}}$, respectively. Compute AUROC based on the input OOD scores.
4: Similarly, generate output embeddings $\mathcal{E}^{\text{in}}_{\text{test}} = \{\boldsymbol{w}|f_d(\hat{\boldsymbol{y}}), \hat{\boldsymbol{y}} \in \mathcal{D}^{\text{in}}_{\text{test}}\}$ and $\mathcal{E}^{\text{ood}}_{\text{test}} = \{\boldsymbol{w}|f_d(\hat{\boldsymbol{y}}), \hat{\boldsymbol{y}} \in \mathcal{D}^{\text{ood}}_{\text{test}}\}$. Compute output OOD score $\text{RMD}_{\text{output}}(\boldsymbol{w})$ for $\boldsymbol{w} \in \mathcal{E}^{\text{in}}_{\text{test}}$ and $\mathcal{E}^{\text{ood}}_{\text{test}}$, respectively. Compute AUROC based on the output OOD scores.

A.5 THE CONNECTION BETWEEN RMD AND BINARY CLASSIFIER

RMD is a generative model based approach which assumes the distributions of the two classes are Gaussian, while the binary classifier is a discriminative model which learns the decision boundary between two classes. Though they have different settings, under certain condition, the Gaussian generative model can be reduced to a binary classifier. To see the connection, let us assume the label $y = 0$ if the sample is from in-domain, and $y = 1$ if the sample is from the general domain. Let us also assume the two classes have balanced sample size without loss of generality $p(y = 1) = p(y = 0)$. Since the log-probability of $\log p(y = 1|\boldsymbol{z})$ can be rewritten using the Bayes rule $\log p(y = 1|\boldsymbol{z}) = \log p(\boldsymbol{z}|y = 1) + \log p(y = 1) - \log p(\boldsymbol{z})$, the logit (log odds) can be written as,

$$\begin{aligned}\text{logit} = \log\left(\frac{p(y=1|\boldsymbol{z})}{p(y=0|\boldsymbol{z})}\right) &= \log p(y=1|\boldsymbol{z}) - \log p(y=0|\boldsymbol{z}) \\ &= \log p(\boldsymbol{z}|y=1) - \log p(\boldsymbol{z}|y=0) \\ &= -\frac{1}{2}\left(\text{MD}(\boldsymbol{z}; \boldsymbol{\mu}_{y=1}, \boldsymbol{\Sigma}_{y=1}) - \text{MD}(\boldsymbol{z}; \boldsymbol{\mu}_{y=0}, \boldsymbol{\Sigma}_{y=0})\right) + \text{const.}\end{aligned}$$

When $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{y=1} = \boldsymbol{\Sigma}_{y=0}$, the equation can be further simplified as

$$\begin{aligned}\text{logit} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=0})^T\boldsymbol{z} - \frac{1}{2}\left(\boldsymbol{\mu}_{y=1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{y=0}\right) + \text{const.} \\ &= \boldsymbol{\beta_1}\boldsymbol{z} + \beta_0.\end{aligned}$$

Therefore, when assuming the covariance matrices are identical for the two Gaussian distributions, the Gaussian generative model can be reduced to a binary classification model. However, our RMD does not assume the same covariance matrix in both distributions. We estimate the covariance matrix individually for each class. So our RMD is different from binary classifier, and it has higher model capacity than the binary classifier.

Table A.2: AUROCs for OOD detection for comparing our proposed method with more baseline

	Near Shift OOD		Far Shift OOD		
Measure	cnn_dailymail	newsroom	reddit_tifu	forumsum	samsum
		INPUT OOD			
KNN (α=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
Binary logits	0.997	0.959	1.000	0.999	0.998
		OUTPUT OOD			
NLI score	0.440	0.469	0.709	0.638	0.743
Perplexity	0.424	0.665	0.909	0.800	0.851
Mean(MSP)	0.343	0.616	0.877	0.715	0.826
Energy score	0.460	0.592	0.960	0.899	0.981
Ensemble using MC dropout (N=5)	0.496	0.768	0.970	0.937	0.944
Ensemble using MC dropout (N=10)	0.497	0.774	0.976	0.947	0.956
KNN (α=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
Binary logits	0.989	0.982	1.000	0.998	0.997

A.6 COMPARISON WITH MORE BASELINE METHODS

As we discussed in the related works, OOD detection problem was mainly studied in classification problems, and less studied in CLMs. Though it is not straight forward to extend classifier-based scores to CLMs especially for the input OOD detection, we would like to include as many possible methods as we can to present a comprehensive comparison for different methods.

For those methods which rely on classification head derived logits, MSP (Hendrycks & Gimpel, 2016), max-logit (Hendrycks et al., 2019), and energy score (Liu et al., 2020b), we simply consider the output decoding process as a sequence of classifications over tokens, and take the average of the corresponding score over the generated output tokens $y_1, \ldots, y_T$ as the output OOD scores. Therefore we added the following scores for CLMs,

- Mean(MSP) $-\frac{1}{T}\sum_{t=1}^{T} p(y_t|y_{<t}, \boldsymbol{x})$.
- Energy score $\frac{1}{T}\sum_{t=1}^{T} E(\boldsymbol{x}, f_t)$, where $E(\boldsymbol{x}, f_t) = -\tau \log \sum_{v \in V} e^{f(y_t=v|y_{<t}, \boldsymbol{x})/\tau}$, $f(y_t = v|y_{<t}, \boldsymbol{x})$ is the logit corresponding to the v-th token at the t-th decoding step, V is the token-vocabulary, and τ is the temperature parameter. We set $\tau = 1$ since the original paper (Liu et al., 2020b) suggested the energy score can be used parameter-free by simply setting $\tau = 1$.
- Ensemble estimation of the output perplexity from multiple Monte-Carlo dropout samples. Malinin & Gales (2020); Xiao et al. (2020) propose to turn on the MC dropout layer at the inference time and sample multiple times (N) using different random seeds as a way to approximate the Bayesian neural networks. We follow their idea and generate multiple output sequences and use the averaged perplexity as the uncertainty score. Note that the inference time for ensemble based method is N times of that for the single model based score.
- KNN-based OOD score. Sun et al. (2022) propose to use the distance to the k-th nearest neighbour in the training set in the embedding space as an OOD score. 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.

Mean(MSP), energy score, and ensembled perplexity score, are all derived from the logits of the tokens in output sequences, so they are output OOD scores. The KNN-based method can be applied for both input sequence embeddings and output sequence embeddings.

Table A.2 shows the AUROCs for OOD detection for the above newly added baselines, as a comparison to our methods. First, the logits based output OOD scores, perplexity, mean(MSP), energy score, even the ensembled perplexity score which costs N times of the inference time, are in general not competitive with our proposed method RMD and Binary logits. Though the energy score

Table A.3: Comparison of the OOD detection performance using two different background data, ParaCrawl and C4 sentence.

	WMT			OPUS					MTNT
Measure	nt2014	ndd2015	ndt2015	law	medical	Koran	IT	sub	
				INPUT OOD					
RMD (ParaCrawl)	0.798	0.866	0.863	0.389	0.840	0.957	**0.959**	0.969	0.943
RMD (C4 sent)	0.833	0.916	0.911	0.269	0.811	0.954	0.924	0.985	0.953
Binary logits (ParaCrawl)	**0.864**	0.904	0.904	0.485	0.813	0.963	0.928	0.950	0.963
Binary logits (C4 sent)	0.848	0.916	0.916	0.285	0.808	0.944	0.918	**0.987**	**0.976**
				OUTPUT OOD					
RMD (ParaCrawl)	0.786	0.858	0.861	0.355	**0.845**	0.939	0.951	0.959	0.922
RMD (C4 sent)	0.818	0.901	0.898	0.259	**0.845**	0.953	0.947	0.979	0.947
Binary logits (ParaCrawl)	0.822	0.860	0.865	0.507	0.783	0.942	0.890	0.910	0.931
Binary logits (C4 sent)	0.853	**0.925**	**0.919**	0.294	0.809	**0.964**	0.901	0.981	0.975
				OTHER BASELINES					
Input MD	0.534	0.671	0.670	0.511	0.704	0.737	0.828	0.900	0.668
Output MD	0.609	0.733	0.739	0.482	0.784	0.838	0.900	0.935	0.794
Perplexity	0.570	0.496	0.494	0.392	0.363	0.657	0.343	0.359	0.633
COMET	0.484	0.514	0.525	0.435	0.543	0.632	0.619	0.518	0.724
Prism	0.445	0.504	0.505	0.459	0.565	0.716	0.604	0.577	0.699

is a bit better than perplexity and mean(MSP), and ensembled score is better than energy score, the performance gap between those methods and our proposed method is still big, especially for the near-OOD datasets. Second, KNN-based methods are not as good as MD and RMD either. Though it is possible that the optimal hyper-paramaters suggested by the paper may not be the optimal ones for our problem, searching for the optimal hyper-parameters requires a separate validation set. In contrast, our proposed methods have no hyperparameters.

A.7 EFFECT OF THE CHOICE OF THE BACKGROUND DATASET

Our principle for choosing the background data is to make it as general as possible. For summarization we use the C4 dataset, which contains a large amount of web crawl documents, to represent a broad range of topics. Similarly for translation, we use ParaCrawl dataset, which is also a large web crawl of sentences, because our translation model is a sentence to sentence model, unlike the summarization model that takes the document as the input. To further explore the effect of the background data on the performance, we split C4 documents into sentences and use that as the background data to compute the scores, and compare that with the version using ParaCrawl dataset. The OOD detection performance using C4 sentences is very similar to that using ParaCrawl, as shown in Table A.3. For example, ParaCrawl-based input OOD score has slightly better performance on medial, Koran, IT datasets, while C4 based input score is slightly better at the other datasets. Both are significantly better than the baseline methods, and both give the same ranking of datasets on their OOD-ness, so our conclusion remains. Those results verify that our method is robust to the choice of background data.

A.8 ROC PLOTS FOR THE CORRESPONDING AUROC SCORES FOR OOD DETECTION

To better visualize the OOD detection performance, we present Figure A.3 to show the ROC plots for the corresponding AUROC scores for OOD detection in Table 1. Each of the OOD measures is used for separating the in-domain test data as negative and the OOD test data as positive sets. The AUROC is defined as the area under the ROC curves. The closer an ROC curve is to the upper left corner, the larger the AUROC value is. 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.

A.9 CORRELATION BETWEEN DIFFERENT SCORES AND THE QUALITY METRICS

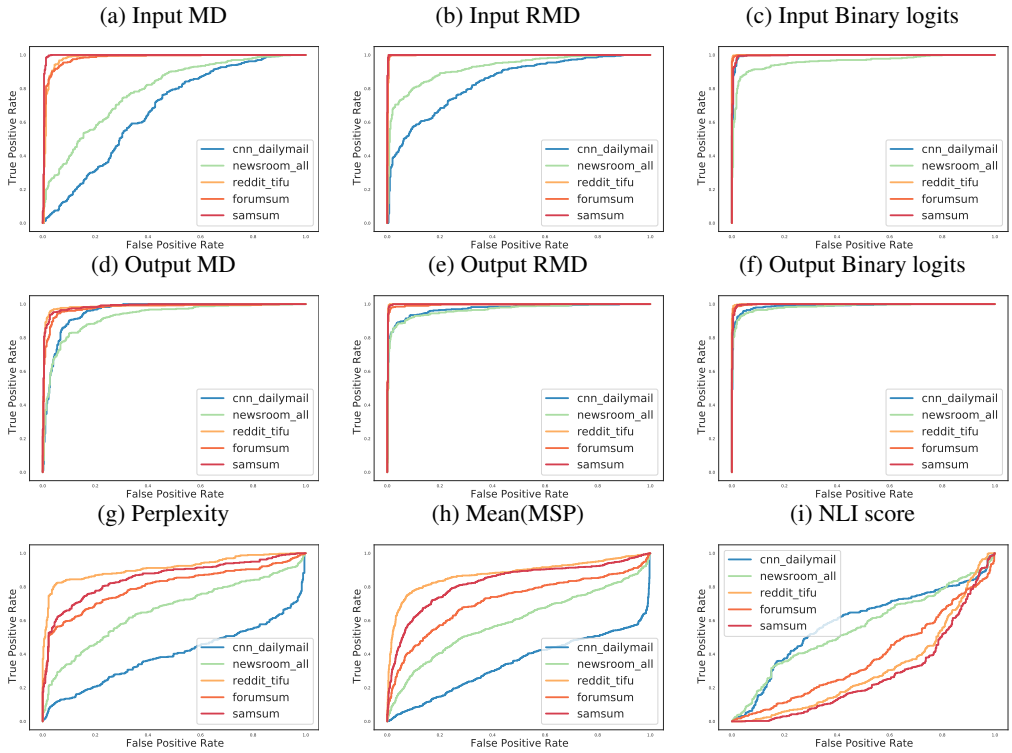


Figure A.3: ROC plots for the corresponding AUROC scores in Table 1 for OOD detection in summarization

Table A.4: Kendall's τ correlation (p-value < 0.05 are greyed out) between various measures with human-judged quality of a PEGASUS xsum model decoded on summarization datasets. The "All" column shows the correlation when examples from all datasets are included. Note for negatively correlated scores (e.g. perplexity, OOD score), we take the negative value of the score for easier comparison. A few intra-dataset correlations have p-value < 0.05 due to the small sample size (only 100 examples per dataset were sent for human evaluation).

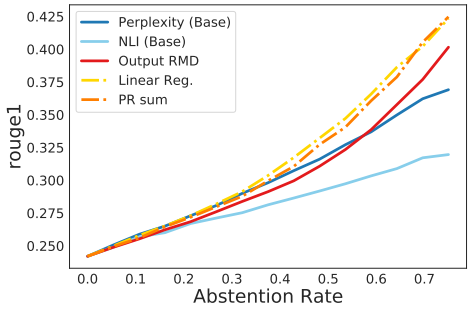
Measure	In-domain xsum	Near Shift OOD cnn_dailymail	Far Shift OOD reddit_tifu	Far Shift OOD samsum	All
Single Score					
INPUT OOD					
MD	0.044	-0.018	-0.017	0.133	0.328
RMD	0.015	-0.033	0.017	0.133	0.336
Binary Logits	-0.022	-0.061	0.028	0.106	0.233
OUTPUT OOD					
Perplexity (baseline)	0.256	0.186	0.081	0.068	0.300
NLI score (baseline)	0.337	0.308	0.226	0.132	0.381
MD	0.106	-0.055	0.202	0.352	0.384
RMD	0.053	0.177	0.214	0.314	0.385
Binary logits	0.199	-0.100	0.091	0.026	0.213
Combined Score					
PR sum (perplexity, input RMD)	0.186	0.134	0.082	0.109	0.358
PR sum (perplexity, output RMD)	0.250	0.350	0.168	0.237	0.415
PR sum (perplexity, input & output RMD)	0.171	0.242	0.158	0.250	0.401
PR sum (perplexity, input binary logits)	0.214	0.079	0.126	0.090	0.322
PR sum (perplexity, output binary logits)	0.347	0.086	0.114	0.052	0.330
PR sum (perplexity, input & output binary logits)	0.277	0.003	0.127	0.096	0.307
Lineare regression (perplexity, input & output)	0.235	0.402	0.170	0.250	**0.422**

Table A.5: Kendall τ correlation (p-value < 0.05 are grayed out) between various measures and quality measured by BLEURT on translation datasets. For easier comparison, we negate the signs of the coefficients for measures that are expected to have negative correlation with BLEURT (e.g., OOD score). Within the same dataset, perplexity shows good correlation, but it deteriorates (with the exception of MTNT) as we move to more OOD datasets such as Koran.

Measure	WMT				OPUS					MTNT	All
	holdout	nt2014	ndd2015	ndt2015	law	medical	Koran	IT	sub		
Single Score											
INPUT OOD											
MD	-0.081	-0.131	-0.129	-0.117	-0.171	0.041	-0.147	-0.093	0.012	-0.117	0.007
RMD	0.147	0.091	0.049	0.115	0.197	0.013	-0.071	-0.060	0.098	0.083	0.195
Binary logits	0.144	0.116	0.141	0.162	0.124	-0.003	0.025	-0.071	0.104	0.161	0.202
OUTPUT OOD											
Perplexity (baseline)	0.309	0.337	0.352	0.375	0.389	0.224	0.222	0.225	0.227	0.341	0.286
COMET (baseline)	0.184	0.397	0.402	0.443	0.324	0.253	0.359	0.174	0.297	0.414	0.336
Prism (baseline)	0.184	0.329	0.337	0.342	0.179	0.188	0.192	0.151	0.286	0.370	0.301
MD	-0.029	-0.066	-0.064	-0.048	-0.096	0.032	-0.105	-0.057	0.041	-0.020	0.083
RMD	0.086	0.049	0.044	0.095	0.135	-0.026	-0.077	-0.056	0.061	0.077	0.170
Binary logits	0.106	0.058	0.075	0.114	0.094	-0.036	-0.013	-0.059	-0.012	0.075	0.151
Combined Score											
RR sum (perplexity, input RMD)	0.321	0.361	0.351	0.410	0.382	0.230	0.161	0.154	0.261	0.354	**0.361**
PR sum(perplexity, output RMD)	0.323	0.357	0.359	0.414	0.371	0.200	0.152	0.164	0.240	0.350	0.356
PR sum(perplexity, input & output RMD)	0.291	0.284	0.264	0.329	0.346	0.119	0.082	0.084	0.231	0.290	0.311
PR sum(perplexity, input binary logits)	0.323	0.352	0.372	0.384	0.391	0.195	0.211	0.111	0.234	0.359	0.335
PR sum(perplexity, output binary logits)	0.318	0.302	0.314	0.350	0.356	0.168	0.162	0.127	0.156	0.293	0.299
PR sum(perplexity, input & output binary logits)	0.300	0.262	0.288	0.309	0.340	0.125	0.145	0.053	0.163	0.287	0.288
Linear regression (perplexity, input & output)	0.318	0.370	0.355	0.414	0.383	0.243	0.180	0.119	0.268	0.367	0.352

A.10 SELECTIVE GENERATION AND OUTPUT QUALITY PREDICTION

(a) ROUGE-1 vs Abstention Rate



(b) Survival count per dataset

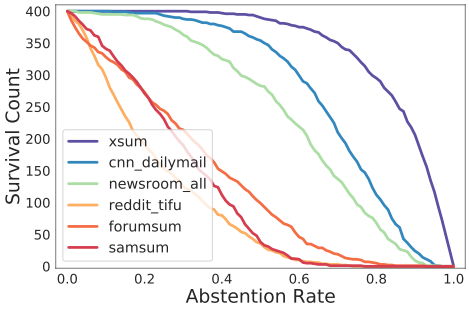


Figure A.4: (a) The summarization quality ROUGE-1 vs abstention curve for single scores, including input and output RMD OOD scores, output perplexity score, and NLI score, and combined scores, including linear regression machine learning model, percentile sum of RMD OOD scores and perplexity score. The corresponding area under the curve is in Table A.7. (b) The survival count of each dataset as the joint dataset is abstained. Each dataset is sub-sampled to 400 examples for this analysis.

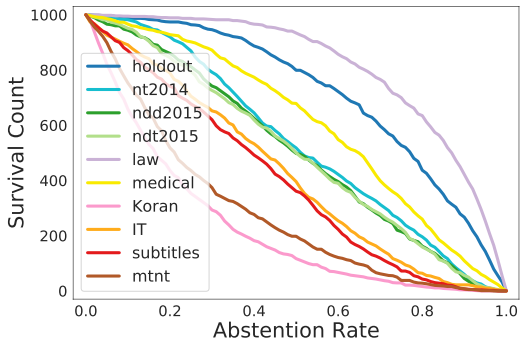


Figure A.5: The translation survival count of each dataset as the joint dataset is abstained. Complete results for Figure 5 (d).

Table A.6: Area under the quality (human eval) vs abstention curve for summarization for various single scores and the proposed combined scores.

Measure	Area under the quality (human eval) vs abstention curve
Single Score	
Input OOD	
MD	0.464
RMD	0.466
Binary logits	0.445
Output OOD	
Perplexity (baseline)	0.458
NLI score (baseline)	0.469
MD	0.469
RMD	0.474
Binary logits	0.441
Combined Score	
PR_{sum}(perplexity, input RMD)	0.468
PR_{sum}(perplexity, output RMD)	<u>0.478</u>
PR_{sum}(perplexity, input & output RMD)	0.476
PR_{sum}(perplexity, input binary logits)	0.461
PR_{sum}(perplexity, output binary logits)	0.461
PR_{sum}(perplexity, input & output binary logits)	0.456
Linear regression (perplexity, input & output RMD)	**0.481**

Table A.7: Area under the quality (ROUGE-1) vs abstention curve for summarization for various single scores and the proposed combined scores.

Measure	Area under the quality (rouge1) vs abstention curve
Single Score	
Input OOD	
MD	0.208
RMD	0.214
Binary logits	0.217
Output OOD	
Perplexity (baseline)	0.221
NLI score (baseline)	0.207
MD	0.219
RMD	0.221
Binary logits	0.207
Combined Score	
PR_{sum}(perplexity, input RMD)	0.222
PR_{sum}(perplexity, output RMD)	<u>0.228</u>
PR_{sum}(perplexity, input & output RMD)	0.224
PR_{sum}(perplexity, input binary logits)	0.225
PR_{sum}(perplexity, output binary logits)	0.221
PR_{sum}(perplexity, input & output binary logits)	0.220
Linear regression (perplexity, input & output RMD)	**0.229**

Table A.8: Area under the quality (BLEURT) vs abstention curve for translation using various single scores and the proposed combined scores.

Names	Area under the quality vs abstention curve
Single Score	
Input OOD	
MD	0.583
RMD	0.623
Binary logits	0.621
Output OOD	
Perplexity (baseline)	0.627
Comet (baseline)	0.644
Prism (baseline)	0.638
MD	0.601
RMD	0.618
Binary logits	0.608
Combined Score	
PR_{sum}(perplexity, input RMD)	**0.647**
PR_{sum}(perplexity, output RMD)	<u>0.646</u>
PR_{sum}(perplexity, input & output RMD)	0.641
PR_{sum}(perplexity, input binary logits)	0.639
PR_{sum}(perplexity, output binary logits)	0.632
PR_{sum}(perplexity, input & output binary logits)	0.633
Linear regression (ppx, input & output)	0.645

A.11 INVESTIGATION OF THE N-GRAM OVERLAP BETWEEN LAW DATASET AND IN-DOMAIN DATASETS

domain/split	overall average	n-gram overlap			
		$n=1$	$n=2$	$n=3$	$n=4$
holdout	**8.3**	45.4	**16.8**	**4.8**	**1.3**
nt2014	4.9	39.0	12.3	2.7	0.5
ndd2015	5.1	40.7	12.9	2.7	0.5
ndt2015	4.6	39.0	12.8	2.6	0.3
law	7.7	**48.8**	16.1	4.2	1.1
medical	4.3	33.5	10.7	2.4	0.4
Koran	2.8	32.6	8.7	1.4	0.2
IT	4.0	35.9	10.6	2.2	0.3
sub	2.8	38.6	10.9	1.4	0.1
MTNT	2.5	31.4	8.4	1.2	0.1

Table A.9: n-gram overlap analysis between the various test sets including law and the in-domain training data, we observe that law has the highest unigram overlap rate (48.8%) and the second highest overall overlap (defined as the geometric mean) with the in-domain data.

A.12 QUANTITATIVE ANALYSIS USING N-GRAM OVERLAP TO DETERMINE NEAR- AND FAR-OOD DATASETS IN SUMMARIZATION

To support our claim that the news related test datasets, cnn_dailymail and newsroom are closer to the in-domain xsum than the other dialogue datasets reddit_tifu, samsum, and forumsum, we compute the n-gram overlap between each of the test datasets and the in-domain dataset. We use Jaccard similarity score, $J(\mathcal{A},\mathcal{B}) = \frac{|\mathcal{A}\cap\mathcal{B}|}{|\mathcal{A}\cup\mathcal{B}|}$, where $\mathcal{A}$ and $\mathcal{B}$ are the set of n-gram in dataset A and dataset B, to measure the similarity between two datasets. Table A.10 shows the similarity scores based on $1-4$ grams. It is clear to see that cnn_dailymail and newsroom have significantly higher similarity with the in-domain xsum data than other three datasets. Therefore, we call the news-related datasets *near*-OOD and the other dialogue based datasets *far*- OOD.

domain/split	overall average	n-gram overlap			
		$n=1$	$n=2$	$n=3$	$n=4$
xsum	7.3	32.4	13.3	4.6	1.4
cnn_dailymail	6.2	31.1	12.7	4.0	0.9
newsroom	5.3	28.8	11.1	3.3	0.7
reddit_tifu	2.8	17.2	6.9	1.8	0.3
forumsum	2.7	18.0	6.5	1.6	0.3
samsum	1.2	10.4	3.1	0.7	0.1

Table A.10: Jaccard similarity based on n-gram overlap between the various test sets and the in-domain xsum training data. We observe that the news-related datasets cnn_dailymail and newsroom have significantly higher similarity scores with the in-domain xsum data than the other three OOD datasets reddit_tifu, forumsum, and samsum.

A.13 VISUALIZATION OF OOD SCORE ON SHIFTED DATASET

We explore how individual parts of an input text contribute to the OOD score, which can help us visualize which parts of the text are OOD. We define the OOD score of each sentence in the text using a leave-one-out strategy: For any given sentence, we compute the OOD score of the article with and without that sentence in it. The negative of the change in the OOD score after removing the sentence denotes the OOD score of that sentence. Intuitively, if removing the sentence decreases the overall OOD score, that sentence is assigned a positive OOD score and vice-versa. Figure A.6 illustrates an example where an article contains noise in the form of tweets with emojis, and the OOD scoring mechanism described above assigns positive OOD scores to those tweets and negative scores to the main text.

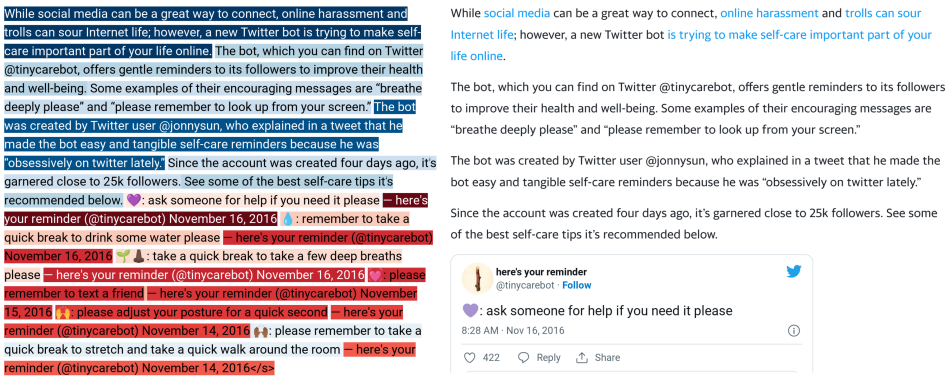


Figure A.6: OOD score can be attributed to individual sentences to highlight the out-of-domain noisy parts of text (red denotes out-of-domain and blue denotes in-domain text), e.g. tweets present in articles scraped from internet. Example taken from Newsroom dataset.

A.14 SUMMARIZATION EXAMPLES WITH LOW/ HIGH PREDICTED QUALITY SCORES

Besides the quantitative results, here we show a few real examples to better demonstrate how well our predicted quality score helps for selective generation on out-of-distribution examples. The model here was fine-tuned on xsum but inference was run on examples from cnn_dailymail.

Figure A.7, A.8, and A.9 show 3 examples in cnn_dailymail that have the highest PR_{sum}(perplexity, output RMD) scores that predict for low quality summaries.

Figure A.10, A.11, and A.12 show 3 examples in cnn_dailymail that have the lowest PR_{sum}(perplexity, output RMD) scores that predict for high quality summaries.

Document: A man trying to elude police jumped into a Missouri creek overnight wearing only his underwear – but his daring gambit did not pay off. Responding officers and firefighters followed the fugitive into the murky waters of Brush Creek in Kansas City and fished him out early Friday morning. The 38-year-old suspect has been taken to an area hospital to be treated for injuries to his arm and leg. He may face charges in connection to a hit-and-run crash. Escape by water: A 38-year-old man stripped down to his skivvies and jumped into Brush Creek in Kansas City, Missouri, after being stopped by police. Up Brush Creek without a paddle: The suspect reached the middle of the creek and spent 10-15 minutes swimming back and forth. According to a Kansas City Police Department's arrest report, officers were called to a gas station in the 4600 block of Prospect at around 2am after receiving complaints from neighbors about a car blasting loud music. The report states that when police approached the car, a grey 2007 Infinity, and asked to see the driver's license, the man smiled, said, 'I'm out!' and took off from the scene. The Infinity promptly smashed into the north side of the Brush Creek bridge, after which the driver got out of the mangled car and jumped into the water. Police say the 38-year-old suspect stripped down to his underwear and spent 10-15 minutes swimming in chest-deep water, with officers waiting for him on north and south sides of the creek. Surrounded: When firefighters tried to pull him out, he threatened them with a log. Fish out of water: Police officers armed with a BB gun went after the nighttime bather and apprehended him. The bather was complaining of a broken leg, according to Fox4KC, so the Kansas City Fire Department's water rescue crew were sent in to fish him out. But the half-naked man in the water was not going to go quietly. 'The suspect picked up a large log and started swinging it at the firemen so they backed off as to not escalate the situation,' the arrest report states. That is when uniformed police officers armed with a BB gun followed the man into the creek, got him in a choke hold and pulled him out of the creek. Police suspect the man may have been under the influence of drugs or alcohol. Prelude: Before he jumped in the water, the 38-year-old driver fled from police and smashed his 2007 Infinity into a bridge. Police suspect the man may have been under the influence of drugs or alcohol at the time. As of Friday morning, the 38-year-old has not been formally charged with any crime.
Reference Summary: The 38-year-old suspect was questioned by Kansas City police after neighbors complained he was blasting music in his 2007 Infinity. Instead of handing over his ID, driver smiled, said 'I'm out!' and took off. After crashing into bridge, the man stripped down to his underwear and jumped into Brush Creek. It took cops armed with a BB gun 15 minutes to fish out the fugitive.
Model Summary: All images are copyrighted.
Human rating score (↑ means high quality): 0.2
PR_{sum}(perplexity, output RMD) (↓ means high quality): 0.67

Figure A.7: Examples in cnn_dailymail that have the highest PR_{sum}(perplexity, output RMD) scores that predict for low quality summaries.

Document: A crisp fan who gets through 42 bags in a week has discovered a skull-shaped deep-fried potato snack in one of his packets. Barry Selby, 54, who lives with his dog in Poole, Dorset, was eating a bag of cheese and onion crisps when he made the bizarre discovery, which appears to be a profile of a human skull. The floor-fitter has decided to keep the two inches tall by two-and-a-half inches wide snack as he believes it is far more impressive than other oddly-shaped examples he has seen on the internet. Scroll down for video. Spooky find: Barry Selby was eating a bag of Tesco cheese and onion crisps when he found the 'skull' snack. Mr Selby said: 'I was shocked when I found it. I was just eating a bag of cheese and onion crisps from Tesco and when I pulled it out it did take me back a bit. 'I thought it was worth keeping as I don't think I will ever find one like it again. It must have been a very weird-shaped potato. 'It's about two inches tall and two-and-a-half inches wide and it's in perfect detail, it even has an eye socket. 'I sometimes give my dog, Max, crisps in a bowl, so it's lucky he didn't have this packet or I wouldn't have found it. Weird snack: Mr Selby has decided to keep the unusual find, which appears to show a jaw, nose and eye. Comparison: The 54-year-old said he was 'shocked' to make the discovery, although it is not his first. In the 1990s he came across a 3D heart-shaped crisp, which he kept until it broke. And it's not the first odd-shaped snack he has come across - in the 1990s he found a crisp shaped like a 3D heart, which he kept for several years until it broke. But he says this find was different: 'This one was a big one. I just thought "wow" and wanted to share it. 'I've been keeping it on top of my computer in the front room, but it should be in a protective box really. 'I'm going to keep it forever, it's just so spooky. I looked on the internet for other funny-shaped crisps but this is a one-off.'
Reference Summary: Barry Selby from Dorset was eating bag of Tesco cheese and onion crisps. The 54-year-old discovered a snack shaped like profile of the human skull. He said he was 'shocked' with the find and has decided to 'keep it forever' It's not his first weird food find - he once discovered a heart-shaped crisp.
Model Summary: All images are copyrighted.
Human rating score (↑ means high quality): 0.2
PR_{sum}(perplexity, output RMD) (↓ means high quality): 0.66

Figure A.8: Examples in $\mathrm{cnn\_dailymail}$ that have the highest PR_{sum}(perplexity, output RMD) scores that predict for low quality summaries.

Document: Last week she was barely showing – but Demelza Poldark is now the proud mother to the show's latest addition. Within ten minutes of tomorrow night's episode, fans will see Aidan Turner's dashing Ross Poldark gaze lovingly at his new baby daughter. As Sunday night's latest heartthrob, women across the country have voiced their longing to settle down with the brooding Cornish gentleman – but unfortunately it seems as if his heart is well and truly off the market. Scroll down for video. Last week she was barely showing – but Demelza Poldark is now the proud mother to the show's latest addition. He may have married his red-headed kitchen maid out of duty, but as he tells her that she makes him a better man, audiences can have little doubt about his feelings. What is rather less convincing, however, is the timeline of the pregnancy. With the climax of the previous episode being the announcement of the pregnancy, it is quite a jump to the start of tomorrow's instalment where Demelza, played by Eleanor Tomlinson, talks about being eight months pregnant. Just minutes after – once again without any nod to the passing of time – she is giving birth, with the last month of her pregnancy passing in less than the blink of an eye. With the climax of the previous episode being the announcement of the pregnancy, it is quite a jump to the start of tomorrow's instalment where Demelza, played by Eleanor Tomlinson, talks about being eight months pregnant. As Sunday night's latest heartthrob, women across the country have voiced their longing to settle down with Poldark – but unfortunately it seems as if his heart is well and truly off the market. Their fast relationship didn't go unnoticed by fans. One posted on Twitter: 'If you are pregnant in Poldark times expect to have it in the next 10 minutes' It is reminiscent of the show's previous pregnancy that saw Elizabeth, another contender for Ross's affection, go to full term in the gap between two episodes. This didn't go unnoticed by fans, who posted on Twitter: 'Poldark is rather good, would watch the next one now. Though if you are pregnant in Poldark times expect to have it in the next 10 minutes.'
Reference Summary: SPOILER ALERT: Maid gives birth to baby on Sunday's episode. Only announced she was pregnant with Poldark's baby last week.
Model Summary: It's all change in the world of Poldark.
Human rating score (↑ means high quality): 0.4
PR_{sum}(perplexity, output RMD) (↓ means high quality): 0.62

Figure A.9: Examples in $\mathrm{cnn\_dailymail}$ that have the highest PR_{sum}(perplexity, output RMD) scores that predict for low quality summaries.

Document: Rangers boss Stuart McCall says he is already working on a dossier of signing targets for next season - even though he may not be around to parade them. The interim Ibrox manager still does not know if he will be in charge beyond the current campaign after being lured back to his old club to kick-start their faltering promotion bid. So far, everything is going to plan with Gers second in the Scottish Championship table and destined for a semi-final play-off slot. Stuart McCall says he is already looking at transfer targets for next season, though he may not be at Rangers. But with 12 players out of contract, McCall knows the Light Blues will need to strengthen if they have any chance of keeping pace with rivals Celtic next season - if they go up - and is already piecing together a wish list of potential new arrivals. He said: 'I've been speaking to a lot of agents and putting things in place for if and when... Even if I'm not here, if I'm getting players put to me who would like to come to Rangers regardless of the manager, then we build a little portfolio of positions that we will be needing next year. 'It's not a case of us standing still and then thinking come June 1, 'Oh we need to get into action'. 'No, there are a lot of agents who come to us and we build a little dossier of players that as a staff, we think will be good for next season, regardless of what league we are in. 'It would be slightly naive [if we were not doing that]. If I'm in charge or not, I still want the club to do well and I will put my view across to the board on who I think should be coming into the club and who should be here.' McCall is compiling a dossier on targets as he looks to put the club in the best possible position. Rangers have operated a haphazard transfer policy since re-emerging from the embers of liquidation. The club's team of scouts were jettisoned under the disastrous Craig Whyte regime and former boss Ally McCoist was largely forced to turn to a list of former Ibrox servants he had personal knowledge of when trying to bolster his squad. But McCall revealed the club's new board are now starting the process of re-establishing their spying network - albeit on a smaller level than before. 'I think there has been discussions behind the scenes with different people,' said the former Motherwell boss. 'I don't think we are at the stage where we were 10 or 15 years ago where we were aiming to get into the Champions League and bringing players in for three and four million yet. 'I don't think Rangers will be at the stage yet next year where we need international scouts everywhere. Rangers have expanded their scouting network after a haphazard system over the past few years. 'But certainly a scouting network needs to be put in place. 'Having said that, I spoke to Craig Levein at Hearts and they do a lot of their scouting with [online service] Wyscout. When I brought Henrik Ojamaa in at Motherwell, that was after I'd seen a clip of him on YouTube. I sold him for £350,000 after signing him for nothing. That was great. 'So you can still do your own background work. Personally I would always like to see the player myself. I've only ever signed one player without watching him first and slightly regretted it. 'So yeah we need a scouting network but at this moment where Rangers are, not to the extent where we have scouts all over Europe.' McCall admitted he still does not know if he will rejoin Gordon Strachan's Scotland staff for the June 13 Euro 2016 qualifier with Ireland in Dublin. And he also confessed to uncertainties ahead of Saturday's match with Falkirk. McCall's side are still in line for promotion, sitting in the play-off positions in the Scottish Championship. Peter Houston's Bairns - five points behind fourth-placed Queen of the South with two games to play - need an unlikely series of results to make the play-offs but McCall says that raises more questions than answers. He said: 'Housty is a wily old fox who has done terrifically well in his career so I don't know what to expect. 'It will take a difficult set of results for them to get into the play-offs so I don't know if they will come here and think the pressure is off and play care free. 'They don't lose many goals so we may have to be patient through the 90 minutes. We have had a couple of decent results against them but they have capable players and we will need to be at our best.'
Reference Summary: Rangers are currently second in the Scottish Championship. Stuart McCall's side are in pole position to go up via the play-offs. But McCall is still not certain of his future at the club next season. Rangers boss says he is still trying to build the squad for next year. Rangers have begun to expand their scouting after several poor years.
Model Summary: Stuart McCall says he is already looking at transfer targets for next season, though he may not be at Rangers.
Human rating score (↑ means high quality): 0.8
PR_{sum}(perplexity, output RMD) (↓ means high quality): 0.10

Figure A.10: Examples in cnn_dailymail that have the lowest PR_{sum}(perplexity, output RMD) scores that predict for high quality summary.

Document: An Alberta student who'd accidentally left his headlights on all day was greeted by what may have been the world's friendliest note from a stranger when he returned to his car. But Derek Murray, a University of Alberta law student, found more than just the note that cold November day in Edmonton–he also found an extension cord and battery charger left by the stranger to bring his dead Acura back to life. Now that Murray's life-affirming tale has now gone viral, he says 'It just shows you how such a pure act of kindness from one person can just spread through everyone and help make everyone's day a little brighter.' Good Samaritan: A friendly stranger left this unbelievably friendly letter to Alberta law student Derek Murray in order to help him get his car started after he left the headlights on all day. At first, though, he assumed the letter was from an angry fellow motorist, he told the National Post. 'When I first saw the note, I was expecting it to be an angry letter from someone telling me not to park there. Instead, I got someone just totally brightening my day. My day could have been ruined but, because of this guy, it was the highlight of my day.' The note reads, in part:. I noticed you left your lights on. The battery will probably not have enough charge to start your vehicle. I left a blue extension cord on the fence and a battery charger beside the fence in the cardboard box. If you know how to hook it up, use it to start your car. What followed was a detailed explanation of how to use the equipment. 'Sure enough,' Derek recalled to the National Post, 'I looked over at the house my car was parked beside, and there was a blue extension cord plugged into an outlet behind the guy's house with a battery charger right there beside it.' Derek was able to get his car started, but when he rang the good Samaritan's doorbell, there was no answer. So, Derek left his own note as a thank you for the kind gesture. He later snapped a photo of the stranger's friendly note to post to Facebook, where it has now gone viral. The note has been viewed millions of times and even Edmonton Mayor Don Iveson retweeted the photo. Derek snapped a photo of the note for Facebook and it has since gone viral. e 'It just shows you how such a pure act of kindness from one person can just spread through everyone and help make everyone's day a little brighter,' Derek said.
Reference Summary: Derek Murray, a University of Alberta law student, could have had his day ruined by the mistake by a stranger's kindness brightened it up. Murray posted his story and the note online and the random act of kindness has now gone viral.
Model Summary: A Canadian student who accidentally left his headlights on all day was greeted by what may have been the world's friendliest note from a stranger when he returned to his car.
Human rating score (↑ means high quality): 0.8
PR_{sum}(perplexity, output RMD) (↓ means high quality): 0.11

Figure A.11: Examples in cnn_dailymail that have the lowest PR_{sum}(perplexity, output RMD) scores that predict for high quality summary.

Document: Bayern Munich had to make do without FOUR important first-team stars as Pep Guardiola's side attempted to overturn a 3-1 deficit against Porto on Tuesday night. Injured quartet Franck Ribery, Mehdi Benatia, David Alaba and Arjen Robben were forced to watch on from the sidelines as the German giants bid to reach the Champions League semi-finals. However, the absence of Robben and Co appeared to make no difference as Bayern raced into a 5-0 lead at half-time before claiming a 6-1 victory to win the tie 7-4 on aggregate. Injured trio Franck Ribery, Mehdi Benatia and David Alaba chat ahead of Bayern's clash with Porto. Injured Ribery acknowledges a steward before taking a seat at the Allianz Arena on Tuesday night. Ribery looks on as former Roma defender Benatia chats with the France international in the dugout. While Ribery, Benatia and Alaba chatted in the home dugout ahead of kick-off, Holland international Arjen Robben was in front of the mic doing some punditry alongside Bayern goalkeeping legend Oliver Kahn. Ribery missed the game after failing to recover from a recent ankle injury while former Roma defender Benatia faces another two weeks out with a groin problem. Robben was unavailable for the encounter with an abdominal injury. David Alaba, meanwhile, is set for a month on the sidelines having partially ruptured knee ligaments playing for Austria at the start of April. Bayern had just 14 fit players to choose from against Porto in the first leg but tore the Portuguese giants apart at the Allianz Arena to progress. Holland international Arjen Robben was pictured doing punditry alongside Bayern legend Oliver Kahn (right) Bayern Munich wideman Robben was unavailable for the Champions League clash with an abdominal injury.
Reference Summary: Bayern Munich beat Porto 6-1 at the Allianz Arena on Tuesday night. German giants were without Franck Ribery, David Alaba and Mehdi Benatia. Arjen Robben was also sidelined and did some punditry for the tie.
Model Summary: Arjen Robben, Mehdi Benatia, Franck Ribery and David Alaba all missed Bayern Munich's Champions League quarter-final second leg against Porto. Holland international Arjen Robben was pictured doing punditry alongside Bayern legend Oliver Kahn (right) Bayern Munich wideman Robben was unavailable for the Champions League clash with an abdominal injury.
Human rating score (↑ means high quality): 0.8
PR_{sum}(perplexity, output RMD) (↓ means high quality): 0.11

Figure A.12: Examples in cnn_dailymail that have the lowest PR_{sum}(perplexity, output RMD) scores that predict for high quality summary.