Out-of-Distribution Detection and Selective Generation for Conditional Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

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 out-of-distribution (OOD) inputs as the prediction is done auto-regressively over many steps. 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* 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., 2016; 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, user provided inputs might 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). This has led to active research on OOD detection for a variety of domains, including vision and text but focused primarily on classification. Conditional language models are typically trained given input sequence to auto-regressively generate the next token in a sequence. Consequently, the perils of out-of-distribution are arguably more severe as errors propagate and magnify through auto-regression. Common errors from text generation models include disfluencies (Holtzman et al., 2020) and factual inaccuracies (Goodrich et al., 2019; Maynez et al., 2020).

In this work, we propose OOD detection methods for CLMs using abstractive summarization and translation as case studies. We show that CLMs have untrustworthy likelihood estimation on OOD examples, making perplexity a poor choice for OOD detection. 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 near-domain data. Thus the ability to selectively generate where the model is more likely to produce higher-quality outputs, enables safer deployment

Submitted to Workshop on Robustness in Sequence Modeling at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

of conditional language models. We call this procedure *selective generation*, analogous to the commonly used term *selective prediction* in classification (Geifman & El-Yaniv, 2017). We show that while model perplexity is a reasonable choice for in-domain examples, combining with our OOD score works much better when the input distribution is shifted.

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). 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. However, we found that the perplexity distributions overlap significantly with each other even though the input documents are significantly different, suggesting that perplexity is not well suited for OOD detection (see Figure A.1 for details).

Detecting OOD using CLM's embeddings We use Transformer encoder-decoder models and obtain the **input embedding** $\boldsymbol{z}$ by averaging the encoder's final-layer hidden state vectors $\boldsymbol{h}_i$ 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 $\boldsymbol{g}_i$ corresponding to the output token y_i. See Figure A.2.

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, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, to the training embeddings and use the *Mahalanobis distance (MD)*:

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

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 CLMs 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 and output MD are computed as

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

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

Ren et al. (2019) and Ren et al. (2021) showed that MD can be confounded by background effect for OOD in classification. In this work, we extend the idea to CLMs and propose,

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

where $\mathrm{MD}_0(\boldsymbol{z}_{\text{test}}) := \mathrm{MD}(\boldsymbol{z}_{\text{test}}; \boldsymbol{\mu}_0^{\boldsymbol{z}}, \boldsymbol{\Sigma}_0^{\boldsymbol{z}})$ 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. Similarly we define,

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

where $\mathrm{MD}_{\delta}(\boldsymbol{w}_{\text{test}}) := \mathrm{MD}(\boldsymbol{w}_{\text{test}}; \boldsymbol{\mu}_{\delta}^{\boldsymbol{w}}, \boldsymbol{\Sigma}_{\delta}^{\boldsymbol{w}})$ 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 details. The RMD score can be regarded as a background contrastive score that indicates how close the test example is to the in-domain compared to the background domain. We also consider a **binary classifier** between in-domain and background domain and use the logit as the OOD score, but we found RMD is better at distinguishing near-OOD from far-OOD than binary logits (See Section A.1.2).

3 Experiments

Experiment setup For summarization, we use a $\text{PEGASUS}_{\text{LARGE}}$ model (Zhang et al., 2020) fine-tuned on the `xsum` (Narayan et al., 2018), consisting of BBC News articles with short, abstractive summaries. We use 10,000 examples from `xsum` and `C4` (Raffel et al., 2020) 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

Table 1: AUROCs for OOD detection for summarization (upper) and translation (lower) tasks.

	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

	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
Output MD	0.609	0.733	0.739	0.482	0.784	0.838	0.900	0.935	0.794
Output 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

DailyMail; newsroom (Grusky et al., 2018), from 38 major news publications; reddit_tifu (Kim et al., 2018), informal stories from sub-reddit TIFU with author written summaries; samsum (Gliwa et al., 2019) and forumsum (Khalman et al., 2021), high-quality summaries of casual dialogues.

For translation, we train a Transformer base model (Vaswani et al., 2017) on WMT15 En-Fr (Bojar et al., 2015). We use 100K examples from WMT15 En-Fr and 100K from ParaCrawl En-Fr (Bañón et al., 2020) to fit the foreground and background Gaussian distributions, respectively. For test, we use newstest2014 (nt14), newsdiscussdev2015 (ndd15), and newsdiscusstest2015 (ndt15) from WMT15 (Bojar et al., 2015), the law, Koran, medical, IT, and subtitles (sub) subsets from OPUS100 (Tiedemann, 2012; Aulamo & Tiedemann, 2019), and MTNT (Michel & Neubig, 2018) consisting of noisy comments from Reddit.

3.1 OOD detection

Baseline methods We compare our proposed OOD scores with various baseline methods, including (1) perplexity score, (2) embedding-based Mahalanobis distance, (3) Natural Language Inference (NLI) score (Honovich et al., 2022) for summarization task, and (4) COMET (Rei et al., 2020) and (5) Prism (Thompson & Post, 2020) for translation task.

RMD and Binary classifier are better at OOD detection than baselines Table 1 shows the AUROCs for OOD detection for summarization and translation. Overall, our proposed OOD score outperforms the baselines with high AUROCs. The commonly used output metrics, perplexity, NLI, COMET and Prism, have generally low AUROCs, 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. Note that all methods have small AUROC for law dataset, suggesting that none of the methods can detect it as OOD. However we found law has the highest unigram overlap (48.8%) with the in-domain data (Table A.7). This confirms that law is actually not OOD data and explains why no method can detect it.

3.2 Using OOD scores for selective generation

Though completely abstaining from OOD inputs is a conservative option, it is more desirable to expand the use of models beyond strictly in-distribution examples, if output quality 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). Here we

Table 2: Kendall's τ (p-value < 0.05 are grayed out) between various measures and quality score. The "All" column shows the correlation when both in-domain and OOD examples are merged.

(a) Summarization

Measure	In-domain	All
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
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

seek to predict the generation quality given a potentially OOD example, and *abstain* if the quality is low. We call this *selective generation*. To measure output quality, we use BLEURT (Pu et al., 2021) for translation, and human evaluation for summarization (details in Section A.2.1).

We first observe that **perplexity has diminishing capability in predicting quality on OOD data**. Since the models are trained using negative log-likelihood as the loss, perplexity is a good predictor of output quality for in-domain. However, we found that perplexity is worse at predicting quality on shifted datasets. Figure A.5 plots the correlation between perplexity and quality as a function of OOD score. As OOD score increases, a decreasing correlation is observed for both summarization and translation. We further observed that **our OOD score and perplexity are complementary in quality prediction**. As illustrated in Figure A.6, neither perplexity nor OOD score can perfectly separate good and bad examples, and the combination of the two can work much better.

We propose two simple methods to combine perplexity and OOD scores. (1) A simple linear regression, (2) the sum of the percentile ranks (PR) of the scores, i.e. $PR_{sum} = PR_{perplexity} + PR_{OOD}$. Table 2 shows the 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 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.2 and A.3 for an expanded table of scores.

To evaluate selective generation, we propose the ***Quality vs Abstention Curve*** **(QA)**: 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. As shown in Figure 1 (a,c), **the combined scores have the highest quality score at almost all abstention rates for both summarization and translation**, and the linear regression and the PR_{sum} perform similarly. Interestingly, we see our simple combined score is even marginally better than COMET, which is a separate neural network model trained on annotated human evaluation data while our combined score does not use any human annotation for supervision. Area under the QA curves are shown in Tables A.4 and A.6 for reference. Figures 1 (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. OOD and worst-quality data are eliminated first and the in-domain data are absteained last. **The order in which datasets are eliminated corresponds to the aggregate quality by dataset**, which we report in Table A.1.

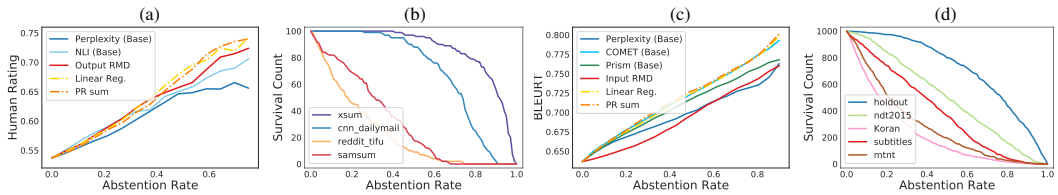

Figure 1: (a) The Quality (`human eval`) vs Abstention curve for summarization task. (b) The survival count of each dataset as a function of population-wise abstention rate of all examples, when abstaining using the PR_{sum}combined score with perplexity (for summarization we use output RMD and for translation we use input RMD to pair with perplexity). (c,d) Same as (a, b) but for translation.

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.

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.

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.

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.

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.

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.

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.

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.

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, 2016. URL http://arxiv.org/abs/1609.08144.

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.

A Appendix

A.1 OOD Detection in Conditinoal Language Models

A.1.1 Perplexity is ill-suited for OOD detection

In Figure A.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.

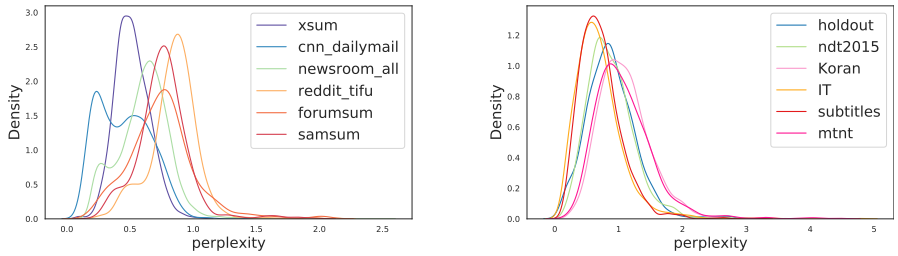

Figure A.1: Perplexity scores density of a CLM trained on `xsum` for summarization (left), and `WMT` for translation (right), evaluated on other datasets/domains. Perplexity is not well suited for OOD detection due to significant overlap between in-domain and OOD scores.

A.1.2 Our proposed OOD detector based on input and output embeddings

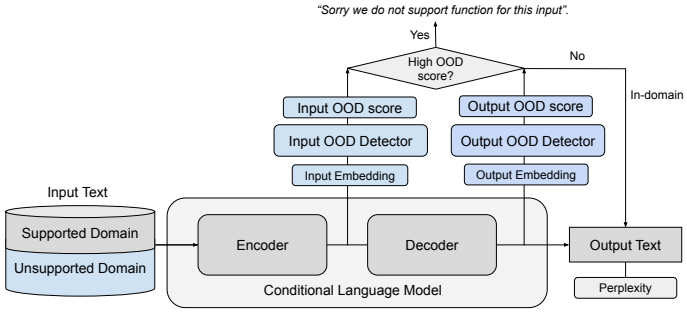

Figure A.2: OOD detector based on input and output embeddings.

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.

Binary classifier for OOD detection Since we have implicitly 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

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 A.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 3.2.

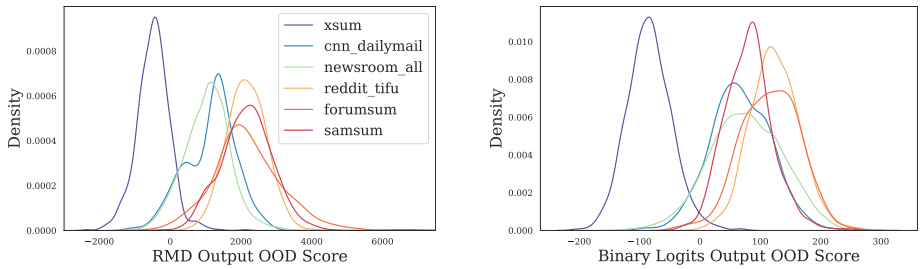

Figure A.3: Density of RMD (left) and Binary logits (right) OOD scores evaluated on summarization datasets. RMD is better at distinguishing near-OOD from far-OOD.

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,

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

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

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

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.

A.2 Using OOD scores for selective generation

A.2.1 Quality metrics used for translation and summarization

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 translation 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 took the median to reduce inter-rater noise. Specifically, a $\text{PEGASUS}_{\text{LARGE}}$ 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.4. 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. The median rating was used to reduce noise.

Read the document below, then rate the summary for quality on a scale of 1-5. (1 = Poor summary, 5 = Great summary)

When assessing quality consider the informativeness, whether it is faithful to the article, and fluency (is the English quality good, does it repeat itself?). Note, some documents and summaries may have adult language but having adult language is not enough by itself to consider a summary bad.

Document to summarize:

```
Michael Coe, 35, saw the two 16-year-olds hugging in the street in Newham, east London, in
April and demanded to know if they were Muslims.
Southwark Crown Court heard the Muslim convert then called the girl a "whore", before
throwing the boy to the ground.
Coe also attacked a passing teacher who had tried to help the couple.
Judge Michael Gledhill QC said the two children had denied they were Muslim when challenged
by Coe.
"Why? Because they were frightened of what you would do if they told you the truth, that they
were in fact Muslim," Jude Gledhill said.
He added: "At the time of these offences you either held extremist views or views that were
getting very close to extremist views."
Coe had admitted "shoving" the boy - who is half his size - claiming he was acting in self-
defence, but was convicted in August of assault occasioning actual bodily harm and battery.
The court heard the father of two was radicalised in prison by al-Qaeda terrorist Dhiren
Barot in 2007 while serving an eight-year term for firing a shotgun at police during an
arrest.
Coe was also convicted of religiously aggravated harassment in 2013 after seeing a Muslim
woman talking to a group of men and telling her that it was against Islam.
The defendant, also known as Mikaeel Ibrahim, became a close associate of convicted hate
preacher Choudary, founder of the banned organisation al-Muhajiroun, of which Coe was a
member.
Prosecutor Jonathan Polnay read a victim impact statement from the boy.
"He feels the offence has affected his life quite a lot," My Polnay said. "He doesn't see his
friends outside of school.
"He has also split up with the girl who was his girlfriend at the time."
```

Summary:

An associate of radical preacher Anjem Choudary who "shoved" a boy who was hugging his girlfriend has been jailed for two years.

Rating:

Figure A.4: AMT template for summarization human evaluation.

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. (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
`cnn_dailymail`	0.226	0.624
`reddit_tifu`	0.140	0.450
`samsum`	0.210	0.376

(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.2.2 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.2). 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.3). Figure A.5 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.

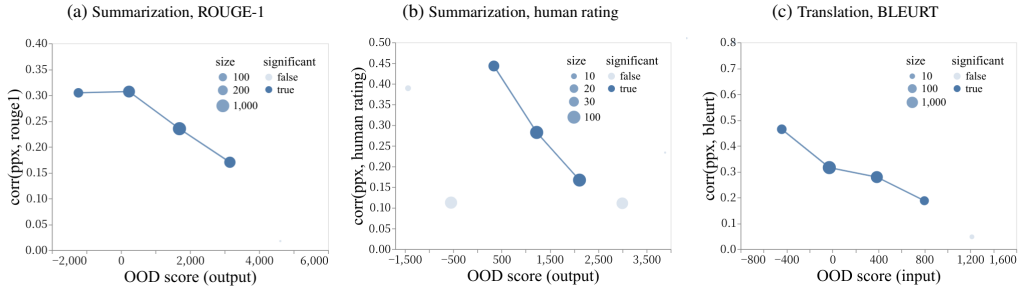

Figure A.5: The Kendall rank correlation coefficient between perplexity and (a) ROUGE-1, (b) human evaluation median rating, and (c) BLEURT decreases as OOD score increases respectively. Note that we use output RMD OOD score for summarization and input RMD OOD score for translation.

A.2.3 OOD score and perplexity are complementary for predicting output quality.

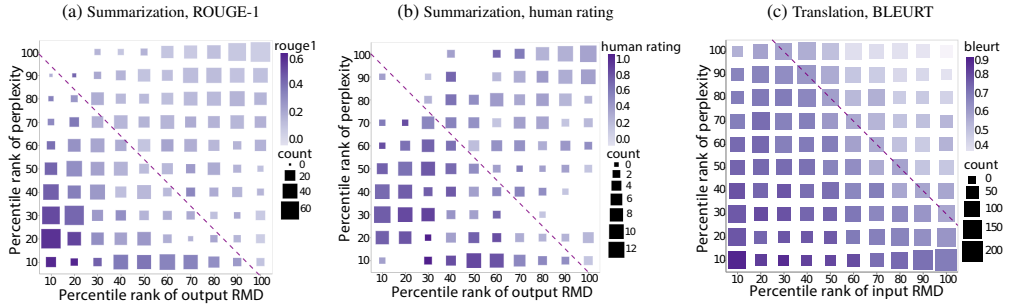

Figure A.6: 2D plot between OOD and perplexity. The two scores are self-normalized by its percentile rank respectively. Each square corresponds to a subset of samples whose OOD and perplexity scores are within the percentile bin. The size of the square represents the size of the bin where the color indicates 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.2.4 Correlation between different scores and the quality metrics

Table A.2: 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	Near Shift OOD	Far Shift OOD		All
	`xsum`	`cnn_dailymail`	`reddit_tifu`	`samsum`	
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.3: 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.2.5 Selective generation and output quality prediction

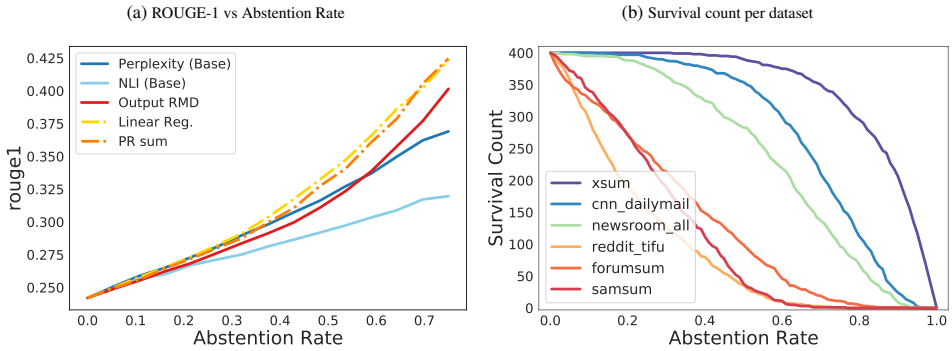

Figure A.7: Similar QA plot as Figure 1 (a, b) but using the automatic quality metric ROUGE-1. The corresponding area under the curve is in Table A.5.

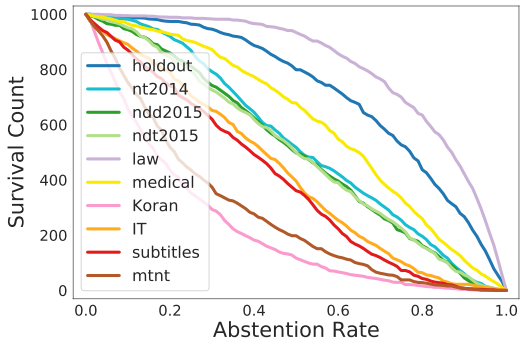

Figure A.8: The translation survival count of each dataset as the joint dataset is abstained. Complete results for Figure 1 (d).

Table A.4: 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.5: 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.6: 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.3 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.7: 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.4 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.9 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.

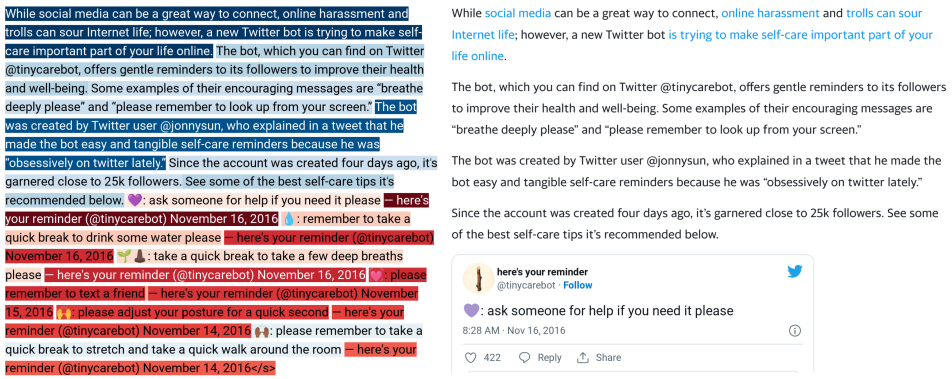

Figure A.9: OOD score can be attributed to individual sentences to highlight the out-of-domain noisy parts of text (red denotes out-of-domain and blue denotes in-domain text), e.g. tweets present in articles scraped from internet. Example taken from Newsroom dataset.

A.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 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).

OOD detection problem is relatively less studied in CLMs. A few studies explored the 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 sample multiple output sequences auto-regressively and use the averaged perplexity as the uncertainty score. 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.