RAINPROOF: AN UMBRELLA TO SHIED TEXT GENERATOR FROM OUT-OF-DISTRIBUTION DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

As more and more conversational and translation systems are deployed in production, it is essential to implement and develop effective control mechanisms to ensure their proper functioning and security. An essential component to ensure the safe behavior of the system is out-of-distribution (OOD) detection, which aims to detect whether an input sample is statistically far from the training distribution. While OOD detection is a widely covered topic in classification tasks, it has received much less attention in text generation. This paper addresses the problem of OOD detection for machine translation and dialog generation from an operational perspective. Our contribution includes (i) `RAINPROOF` a Relative informAItioN Projection Out OF distribution detection framework and (ii) a more operational evaluation setting for OOD detection. Surprisingly, we find that OOD detection is not necessarily aligned with task-specific measures. The OOD detector may filter out samples that are well processed by the model and keep samples that are not, leading to weaker performance. Our results show that `RAINPROOF` breaks this curse and achieve good results in OOD detection while increasing system performance.

1 INTRODUCTION

Significant progress has been made in Natural Language Generation in recent years with the development of powerful generic (*e.g.,* GPT Radford et al. (2018; 2019); Brown et al. (2020)) and task-specific (*e.g,* Grover Zellers et al. (2019), Pegasus Zhang et al. (2020) and DialogGPT Zhang et al. (2019)) text generators. Text generators power machine translation systems or chat bots that are by definition exposed to the public and whose reliability is therefore a prerequisite for adoption. Text generators are trained in the context of a so-called closed world Antonucci et al. (2021); Fei & Liu (2016), where training and test data are assumed to be drawn i.i.d. from a single distribution, known as the in-distribution. However, when deployed, these models operate in an open world Parmar et al. (2021); Zhou (2022) where the i.i.d. assumption is often violated. This change in data distribution is detrimental and creates a drop in performance as illustrated in **??** and Tab. 2. Thus, to ensure thrutworthyness and adoption, it is necessary to develop tools to protect them from harmful distribution shifts. For example, a trained translation model is not expected to be reliable when the input sample is far from the training distribution, as may occur in the case of language drift (*e.g.,* a Spanish model exposed to Catalan, or a Dutch model exposed to Afrikaans) or domain drift ((*e.g.,* a colloquial translation model exposed to rare technical terms from the medical field).

The problem of distributional is inherent to languages that can evolve over time. However, most existing research work focuses on the classification problem, and despite their importance, (conditional) text generation has received much less attention even though it is one of the most exposed applications. Existing solutions fall into two categories. The first one called training-aware methods Zhu et al. (2022); Vernekar et al. (2019a;b) modifies the classifier training by often exposing the DNN to OOD samples during training. The second one called plug-in methods aim to distinguish regular samples in the distribution (ID) from out-of-distribution (OOD) samples based on the behavior of the model on a new input. Plug-in methods include maximum softmax prediction (MSP) Hendrycks & Gimpel (2016) or energy Lee et al. (2018a) or feature-based anomaly scores that compute an anomaly score per class Ming et al. (2022); Ryu et al. (2017); Huang et al. (2020); Ren et al. (2021a). Although plug-in methods seem attractive, their adaptation to text generation may not be straightforward, as the

number of words in a vocabulary is an order of magnitude higher than the number of classes involved in the classical classification framework.

In this work, we aim to develop new tools to build more reliable text generators that can be used in practical systems. First, ($\mathcal{A}_1$) we work in the *unsupervised* detection setting where we do not assume that we have access to OOD samples as in many text applications they are not available. Second, ($\mathcal{A}_2$) we work in the black-box scenario which is the most common in the Software as a Service framework, where the detection methods only has access to the output of the DNN architecture. Finally, ($\mathcal{A}_3$) we want an easy-to-use and effective method to ensure adoptability.

Our contributions This work presents a benchmark and new unsupervised and untrained methods for OOD sample detection in a conditional text generation context, including neural machine translation and dialogue generation. Our main contributions can be summarized as follows:

1. *A more operational benchmark for text generation OOD detection.* OOD detection in text generation remains largely neglected, and often focuses on the GLUE benchmark Arora et al. (2021), moreover they focus on OOD detection performance metrics and not operational metrics relevant for the evaluated task (Such as BLEU Papineni et al. (2002) or BERT-S Unanue et al. (2021) for NMT). Instead, we propose a realistic operational benchmark focused on machine translation and dialog generation; two applications whose systems are widely deployed and in close contact with the public. For translation, we focus on two types of mismatch: linguistic mismatch with language pairs known to be difficult to distinguish (*e.g.*, Spanish and Catalan or Dutch and Afrikaans[1]) and domain change (*e.g.*, medical technical data or news). For dialogues, we focus on dialogues from different sources (*e.g.*, the Multi WOZ dataset and daily dialogues, among others).

2. *Novel information theoretic-based detectors*. We present a new framework `RAINPROOF` for *unsupervised* OOD detection. `RAINPROOF` is flexible and can adapt to the scenario ($\mathbb{s}_0$) where no reference samples (ID) are available by computing the neural network prediction negentropy. We also provide a natural extension using Information Projection in the scenario ($\mathbb{s}_1$) where reference samples (ID) are available. While yielding strong results, our framework remains cheap in terms of computation time. It maintains the same runtime as soft probability based methods while outperforming them as well as Mahalanobis distance which requires precomputations on the reference set.

3. *New insights on the operational value of OOD detectors* We showed that even if common OOD detectors perform well in terms of OOD detection, they do not always filter out samples wisely with respect to task specific performance metrics. More specifically, we show that in machine translation usual OOD detectors tend to hurt the system in average by filtering out samples that would be well handled while keeping more samples that are not. We argue that OOD detectors should help increase the average performances of these systems and that they should not only be evaluated based on their OOD detection performance but also on task-specific performance metrics. We show that `RAINPROOF` outperform baselines from both standpoints.

4. *Code and reproductibility* After acceptance, we will publish the open-source code on `github.com` and the data to facilitate future research, ensure reproducibility and reduce computational costs.

2 Problem Statement & Related Works

2.1 Notations & Conditional Text Generation

Let us denote Ω a vocabulary of size $|\Omega|$ and Ω^* its Kleene closure Fletcher et al. (1990)[2]. Denote $\mathcal{P}(\Omega) = \left\{\mathbf{p} \in [0,1]^{|\Omega|} : \sum_{i=1}^{|\Omega|} \mathbf{p}_i = 1\right\}$ the set of probability distributions defined over Ω. Let $\mathcal{D}_{train}$ be the training set, composed of $N \geqslant 1$ i.i.d. samples $\{(\mathbf{x}^i,\ \mathbf{y}^i)\}_{i=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$ with probability law p_{XY}. We denote p_X and p_Y the associated marginal laws of p_{XY}. Each $\{\mathbf{x}^i\}_{i=1}^N$ is a sequence of token and we denote $x_j^i \in \Omega$ their tokens and $\mathbf{x}_{\leqslant t}^i = \{x_1^i, \cdots, x_t^i\} \in \Omega^*$ the prefix of length t. The same notations holds for $\mathbf{y}$.

[1] Afrikaans comes directly from the Dutch leading to very close languages: Dutch: Appelen zijn gewoonlijk groen, geel of rood. Afr: "Appels is gewoonlik groen, geel of rooi."

[2] The Kleene closure corresponds to sequences of arbitrary size written with words in Ω. Formally: $\Omega^* = \bigcup_{i=0}^{\infty} \Omega^i$.

Conditional textual generation. In conditional textual generation, the goal is to model a probability distribution $p_\star(\mathbf{x}, \mathbf{y})$ over variable-length text sequences $(\mathbf{x}, \mathbf{y})$ by finding p_θ which approximates $p_\star(\mathbf{x}, \mathbf{y})$, *i.e.,* $p_\star(\mathbf{x}, \mathbf{y}) \approx p_\theta(\mathbf{x}, \mathbf{y})$ for any $(\mathbf{x}, \mathbf{y})$. In this work, we assume to have access to a pretrained conditional language model $f_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}^{|\Omega|}$ where the output is the (unormalized) logits scores. f_θ parameterized p_θ, *i.e.,* for any $(\mathbf{x}, \mathbf{y})$, $p_\theta(\mathbf{x}, \mathbf{y}) = \mathrm{softmax}(f_\theta(\mathbf{x}, \hat{\mathbf{y}})/T)$ where $T \in \mathbf{R}$ denotes the temperature. Given an input sequence $\mathbf{x}$, the pretrained language f_θ can recursively generate an output sequence $\hat{\mathbf{y}}$ by sampling $y_{t+1} \sim p_\theta^T(\cdot|\mathbf{x}, \hat{\mathbf{y}}_{\leqslant t})$, for $t \in [1, |\mathbf{y}|]$. Note that $\hat{y}_0$ is the start of sentence ($< SOS >$ token). We denote by $\mathcal{S}(\mathbf{x})$, the set of model normalized logits scores generated by the model when the initial input is $\mathbf{x}$ *i.e.,* $\mathcal{S}(\mathbf{x}) = \{\mathrm{softmax}(f_\theta(\mathbf{x}, \hat{\mathbf{y}}_{\leqslant t}))\}_{t=1}^{|\hat{\mathbf{y}}|}$. Note that any element of $\mathcal{S}(\mathbf{x})$ can be interpreted as a discrete probability distribution on Ω.

2.2 Problem Statement

In OOD detection the goal is to find an anomaly score $a : \mathcal{X} \to \mathbf{R}_+$ that quantifies how a much a sample is far from the IN distribution. $\mathbf{x}$ is classified as in-distribution or OOD according to the score $a(\mathbf{x})$. Following previous work Hendrycks & Gimpel (2016), one fixes a threshold γ and classifies the test sample IN if $a(\mathbf{x}) \leqslant \gamma$ or OOD if $a(\mathbf{x}) > \gamma$. Formally, let us denote $g(\cdot, \gamma)$ the decision function, we take: $g(\mathbf{x}, \gamma) = \begin{cases} 1 & \text{if } a(\mathbf{x}) > \gamma \\ 0 & \text{if } a(\mathbf{x}) \leqslant \gamma \end{cases}$

Remark 1. In practice OOD examples might not be available. In our experiments we take γ such that at least 80% of the train set is classified as IN data. This assumption is reasonable since in practice even a well tailored dataset might contains significant shares of outliers Mishra et al. (2020).

2.3 Review of OOD detectors

OOD detection for classification. Most work on OOD detection has focused on detectors for classifiers and relies either on internal representations (*features-based detector*) or on the final soft probabilities produced by the classifier (*softmax based detectors*).

Features-based detectors leverage latent representations to derive anomaly scores Kirichenko et al. (2020); Zisselman & Tamar (2020). The most well known is the Mahanalobis distance Lee et al. (2018b); Ren et al. (2021b) but there are other methods employing Grams matrices Sastry & Oore (2020), Fisher Rao distance Gomes et al. (2022) or other statistical tests Haroush et al. (2021). Other methods rely on the gradient space Huang et al. (2021) or the moment of the features Quintanilha et al. (2019); Sun et al. (2021). Most of the time, these methods require access to the latent representation of the models, which does not fit our black box scenario. Moreover, they often rely on a per-class decision, which is fine for classifiers but becomes intractable in the case of language models due to the size of Ω.

Softmax-based detectors rely on the soft probabilities produced by the model. The maximum softmax probability Hendrycks & Gimpel (2017); Hein et al. (2019); Liang et al. (2018); Hsu et al. (2020) uses the probability of the mode while others take into account the entire distribution, such as the energy-based OOD detection scores. scores Liu et al. (2020). Due to the large vocabulary size, it is unclear how these methods generalize to sequence generation tasks.

OOD detection for text generation. Little work has been done on OOD detection for text generation. Therefore, we will follow Arora et al. (2021) and will rely on their baseline but also generalize common OOD scores such as MSP or energy to the context of text generation (see **??**).

Remark 2. Note that *features-based* detectors assume a white-box framework where the internal representations of an input are accessible, whereas *softmax based* detectors rely only on the final probability output. Following Arora et al. (2021), we work in a black-box framework Chen et al. (2020), developing scores based solely on the output probability distributions. We always compare our results to the Mahalanobis distance Lee et al. (2018b), as it is known to be a strong baseline.

3 Information Theoretic OOD detectors

In this section, we first present the information measures we will use to construct the detectors for the no-reference scenario ($\mathbb{s}_0$) and the reference scenario ($\mathbb{s}_1$). Finally, we generalize existing OOD detection methods to the sequence generation task. ~~and highlight the connection to our detectors.~~

3.1 INFORMATION THEORETICAL BACKGROUND

An information measures $\mathcal{I} : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \rightarrow \mathbf{R}$ measures the similarity between any pair of discrete distributions $\mathbf{p}, \mathbf{q} \in \mathcal{P}(\Omega)$. Since Ω is a finite set, we will adopt the following notations $\mathbf{p} = [p_1, \cdots, p_{|\Omega|}]$ and $\mathbf{q} = [q_1, \cdots, q_{|\Omega|}]$. The development of new information measures for specific applications has received much attention over the years Fujisawa & Eguchi (2008); Cichocki et al. (2011) (we refer the reader to Basseville (2013) for a complete review). Existing information measures can be grouped into two broad categories: *divergences* and *distances*. Divergence is a measure of dissimilarity that is always positive or equal to zero if (and only if) the two distributions under consideration are strictly identical. The distance is a function that is symmetric, positive, respects the triangular inequality and is equal to zero if (and only if) the two distributions considered are strictly identical. In what follows, we present and motivate the information measures we will use in this work.

First we rely on the *Renyi divergence* Csiszár (1967). The Renyi divergences are a family of f-divergences parametrized by a parameter $\alpha \in \mathbf{R}_+ - \{1\}$ and are flexible because they include well-known divergences such as the Kullback-Leiber (KL) (when $\alpha \rightarrow 1$) distance or the Hellinger distance Hellinger (1909) (when $\alpha = 0.5$). The Rényi divergence between $\mathbf{p}$ and $\mathbf{q}$ can be defined as follows: $D_\alpha(\mathbf{p}\|\mathbf{q}) = \frac{1}{\alpha-1} \log \left(\sum_{i=1}^{|\Omega|} \frac{p_i^\alpha}{q_i^{\alpha-1}} \right)$. The Renyi divergence is useful and widely used in machine learning Peters et al. (2019) because α allows weighting the influence of $\frac{p_i}{q_i}$.

Second, we investigate the *Fisher Rao distance* (FR). FR is a Riemannian metric and is defined on the space of Fisher information-based parametric distributions. It is based on the Fisher information matrix which provides a natural Riemannian structure Amari (2012) on the probability space. It computes the geodesic distance between two discrete distributions Rao (1992); Pinele et al. (2020) between two discrete distributions and is defined as follows: $\mathbf{p}, \mathbf{q}$: $\mathrm{FR}(\mathbf{p}\|\mathbf{q}) = \frac{2}{\pi} \arccos \sum_{i=1}^{|\Omega|} \sqrt{p_i \times q_i}$. It has found many applications in recent machine learning Picot et al. (2022); Colombo et al. (2022b;a) and is known to be more accurate than popular divergence measures Costa et al. (2015).

3.2 RAINPROOF FOR THE NO-REFERENCE SCENARIO

At inference time~~At the time of testing~~, the no-reference scenario ($\mathbb{s}_0$) does not assume the existence of a reference set of IN samples~~clean reference samples~~ to decide whether a new input sample is OOD. Softmax based detectors such as MSP Hendrycks & Gimpel (2016), Energy Liu et al. (2020) or the sequence likelihood Arora et al. (2021) are examples of OOD scores operating under these assumptions.

Under these assumptions our OOD detector RAINPROOF is composed of three steps. For a given input $\mathbf{x}$ with generated sentence $\hat{\mathbf{y}}$:

1. We first use f_θ to extract the step-by-step sequence of soft distributions $\mathcal{S}(\mathbf{x})$.

2. We then compute an anomaly score by averaging the score output by an information measure $\mathcal{I}$ between a reference distribution $\mathbf{u} \in \mathcal{P}(\Omega)$ and each element of $\mathcal{S}(\mathbf{x})$. Formally, $a(\mathbf{x})$ becomes:

$$a_{\mathcal{I}}(\mathbf{x}) = \frac{1}{|\mathcal{S}(\mathbf{x})|} \sum_{\mathbf{p} \in \mathcal{S}(\mathbf{x})} \mathcal{I}\left(\mathbf{p}\|\mathbf{u}\right), \tag{1}$$

where $|\mathcal{S}(\mathbf{x})|$ denotes the cardinal of $\mathcal{S}(\mathbf{x})$.

3. The last step consists in thresholding the previous anomaly score $a_{\mathcal{I}}(\mathbf{x})$. If $a_{\mathcal{I}}(\mathbf{x})$ is over a given threshold γ, we classify $\mathbf{x}$ as an OOD example

Interpretation of Eq. 1. It measures the average distance to normality between the probability distribution of the next token and the reference distribution $\mathbf{u}$~~It measures the average similarity between the probability distributions generated by f_θ at each time step and the reference probability $\mathbf{u}$~~. It corresponds to the token average uncertainty of the model f_θ to generate $\hat{\mathbf{y}}$ when the input is $\mathbf{x}$. The intuition behind Eq. 1 is that the produced distribution of f_θ, when exposed to an OOD sample, should be far from normality and thus should have a higher score.

Choice of $\mathbf{u}$ and $\mathcal{I}$. The uncertainty definition of Eq. 1 depends of the choice of both the reference distribution $\mathbf{u}$ and the information measure $\mathcal{I}$. A natural choice for $\mathbf{u}$ is the uniform distribution *i.e.,*

$\mathbf{u} = [\frac{1}{|\Omega|}, \cdots, \frac{1}{|\Omega|}]$ which we will use in this work. It is worth pointing out that $\mathcal{I}(\cdot||\mathbf{u})$ yields the negentropy of a distribution. Other possible choice includes one hot or tf-idf distribution Colombo et al. (2022b). For $\mathcal{I}$, we rely on Renyi divergence to obtain $a_{\mathcal{D}_\alpha}$ and the Fisher Rao distance to obtain a_{FR}.

3.3 RAINPROOF FOR THE REFERENCE SCENARIO

In the reference-based scenario ($\mathbb{s}_1$) we assume that one has access to a set of **IN samples** $\mathcal{R} = \{x_i \in \mathcal{D}_{train}\}_{i=1}^{|\mathcal{R}|}$ where $|\mathcal{R}|$ is the size of the reference set. For example, the Mahalanobis distance works under this assumption. This reference set can be exploited to obtain a notion of normality. For example, Eq. 1 imposes an ad-hoc choice when using $\mathbf{u}$ (a normal distribution). In what follows we describe how we use $\mathcal{R}$. To leverage $\mathcal{R}$, we build on the intuition that if f_θ is exposed to an OOD sample, the induced distribution should be atypical. In this case, typicality will obtained using $\mathcal{R}$.

Under $\mathbb{s}_1$, RAINPROOF computes the OOD score following four steps:

1. (*Offline*) For each $\mathbf{x} \in \mathcal{R}$, we generate $\hat{\mathbf{y}}(\mathbf{x})$ and the associated sequence of probability distributions. Overall we thus generate $\sum_{\mathbf{x}\in\mathcal{R}} |\hat{\mathbf{y}}(\mathbf{x})|$ which will explodes for long sequences. Thus, we will instead rely on the bag of distributions of each sequence Colombo et al. (2022b) and for the set of these bags of distributions $\bar{\mathcal{S}}^* = \{\frac{1}{|\mathcal{S}(\mathbf{x_r})|} \sum_{\mathbf{p}\in\mathcal{S}(\mathbf{x_r})} \mathbf{p} \; \forall \; \mathbf{x_r} \in \mathcal{R}\}$.

2. (*Online*) For a given input $\mathbf{x}$ with generated sentence $\hat{\mathbf{y}}$, we compute its bag of distributions representation $\bar{\mathbf{p}}(\mathbf{x}) = \frac{1}{|\mathcal{S}(\mathbf{x})|} \sum_{\mathbf{p}\in\mathcal{S}(\mathbf{x})} \mathbf{p}$.

3. (*Online*) We then compute an anomaly score $a^\star_\mathcal{I}$ by projecting $\bar{\mathbf{p}}(\mathbf{x})$ on the set $\bar{\mathcal{S}}^*$. Formally $a^\star_\mathcal{I}(\mathbf{x})$ is defined as:

$$a^\star_\mathcal{I}(\mathbf{x}) = \min_{\mathbf{p}\in\mathcal{D}^\star} \mathcal{I}(\mathbf{p}||\bar{\mathbf{p}}(\mathbf{x})). \tag{2}$$

4. The last step consists in thresholding the previous anomaly score $a_\mathcal{I}(\mathbf{x})$. If $a_\mathcal{I}(\mathbf{x})$ is over a given threshold γ, we classify $\mathbf{x}$ as an OOD example

Interpretation of Eq. 2. It defines a Generalized Information Projection Kullback (1954); Csiszár (1975; 1984)[3] which measures the similarity between the bag of distribution representation of $\mathbf{x}$ and the set of reference distributions $\bar{\mathcal{S}}^*$. Note that the closest element of $\bar{\mathcal{S}}^*$ in the sens of $\mathcal{I}$ can give insights on the decision and allow to interpret the decision of the detector.

Choice of $\mathcal{I}$. Similarly to **??**, we will rely on the Renyi divergence to define $a^\star_{\mathcal{R}_\alpha}(\mathbf{x})$ and the Fisher Rao distance $a^\star_{FR}(\mathbf{x})$.

3.4 GENERALIZATION OF EXISTING OOD DETECTOR TO SEQUENCE GENERATION

In this section, we extend classical OOD detection score to the conditional text generation settting.

Likelihood Score The most naive approach to build a OOD score is to rely solely on the log-likelyhood of the sequence. For a conditioning $\mathbf{x}$ we define the log-likelyhood score by $a_L(\mathbf{x}) = -\sum_{t=0}^{|\hat{\mathbf{y}}|-1} \log \mathbf{p}_\theta(\hat{\mathbf{y}}_{t+1}|\mathbf{x}, \hat{\mathbf{y}}_{\leqslant t})$

Average Maximum Softmax Probability score The maximum softmax probability Hendrycks & Gimpel (2017) takes the probability of the mode of the categorical distribution as score of OOD. We extend thise definition in the case of sequence of probability distribution by averaging this score along the sequence. For a given conditioning $\mathbf{x}$, we define the average MSP score $a_{\text{MSP}}(\mathbf{x}) = \frac{1}{|\hat{\mathbf{y}}|} \sum_{t=1}^{|\hat{\mathbf{y}}|} \max_{i\in[|0,K|]} \mathbf{p}^T_\theta(i|\mathbf{x}, \hat{\mathbf{y}}_{\leqslant t}))$. While it is closely linked to uncertainty measures it discards most of the information contained in the probability distribution. It discards the whole probability distribution. We claim that much more information can be retrieve by studying the whole distribution.

[3]The minimization problem of Eq. 2 finds numerous connections in the theory of large deviation Sanov (1958) or in statistical physics Jaynes (1957).

Average Energy score We extend the definition of the energy score described in Liu et al. (2020) to a sequence of probability distributions by averaging the score along the sequence. For a given conditioning $\mathbf{x}$ and a temperature T we define the average energy of the sequence:$a_E(\mathbf{x}) \triangleq -\frac{T}{|\hat{\mathbf{y}}|}\sum_{t=1}^{|\hat{\mathbf{y}}|}\log\sum_{i}^{|\Omega|} e^{f_\theta(\mathbf{x},\hat{\mathbf{y}}_{\leqslant t})_i/T}$. It corresponds to the normalization term of the $\mathrm{softmax}$ function applied on the logits. While it takes into account the whole distribution, it only takes into account the amount of unormalized mass before normalization without attention to how this mass is distributed along the features.

Mahalanobis distance Following Lee et al. (2018c) compute the Mahalanobis matrice based on the samples of a given reference set $\mathcal{R}$. In our case we are using encoder-decoder models we use the output of the last hidden layer of the encoder as embedding. Let's denote $\phi(\mathbf{x})$ this embedding for a conditionning $\mathbf{x}$. Let's μ and Σ be respectively the mean and the covariance of these embedding on the reference set. We define $a_{\mathrm{M}}(\mathbf{x}) = \left(1 + (\phi(\mathbf{x}) - \mu)^\top \Sigma^{-1}(\phi(\mathbf{x}) - \mu)\right)^{-1}$.

Ours. We investigate ways to aggregate OOD scores along a sequence to compute a sequence-wise OOD score because usual scores only address single distributions situations. Moreoever the uncertainty metrics we propose take into account the whole probability distribution and not only its mode or amount of energy. In addition, the information projection metrics extend our proposal to compare probability distributions to any reference set and not only to the uniform distribution.

4 Experimental Setting

We present present our Language shift Out Of diStribution pErformaNce bEnchmaRk (`LOOSENER`), a benchmark for evaluating perfomance of text OOD detectors. It consists of datasets covering neural machine translation and dialog models and featuring small shifts in input distributions (such as domain shifts or linguistically close languages). It aims at filling a gap in text OOD detection beyond classification tasks and English only tasks.

4.1 NMT Benchmarks

Shifts typology We consider two main types of shifts: Language shifts and Domain shifts. Language shifts can be huge, for example when an English model faces Chinese however we put these situations aside to consider smaller shifts that can arise in some situations. Some languages can be very close linguistically speaking to a point that they can't be trivially distinguished. While they can be extremely similar (such as Afrikaans and Dutch[4]), they can still induce important errors (See Table 4 in Appendix). Thus we focus our attention on closely related language pairs (See 3). The same holds for domain shifts, when a model faces a text on a very technical topic that it has not been trained on. In Tables **??** and **??** we show performance drops in translation on OOD samples.

Language Shifts Benchmarks We use the Tatoeba MT data set Tiedemann (2020) and the news commentary data set Tiedemann (2012a) as our main sources of sample from common language.

Domain Shift Benchmark As technical data we used the European parallel translation corpus from the European Medicines Agency (EMEA) which consists of regulatory texts featuring all pairs of languages of the EU countries and the EuroParl dataset constituted of regulatory texts of the EU Tiedemann (2012b).

Building OOD datasets For each pair of IN distribution data and Out of distribution data we constitute a balanced dataset of IN and OOD samples as presented in Table 3.

Measuring translation quality In addition to common OOD detection metrics outlined in Section 4.3 we measure the quality of the translations generated on the different datasets (See Figure **??**,Tables 2 with and without filtering the samples with our method. To do so we relied on common automatic text evaluation metrics: `BLEU` Papineni et al. (2002) and `BERT-S` Unanue et al. (2021). Moreover we assess the gain in translation performance in a realistic setting, we chose the detection threshold using only the IN dataset as OOD samples would not be available in a real scenario. Thus the gains we present are achievable in practice.

[4]Afrikaans comes directly from the Dutch leading to very close languages: Dutch: Appelen zijn gewoonlijk groen, geel of rood. Afr: "Appels is gewoonlik groen, geel of rooi."

Models We benchmark our OOD method on translation models provided by Helsinky NLP Tiedemann & Thottingal (2020) on several pairs of languages with large and small shifts. We extended the experiment to detect domain shifts.

4.2 DIALOG BENCHMARK

Chat bot systems are also widely publicly used systems that are likely to encounter OOD inputs Perez et al. (2022). In order to avoid providing answers outside their domain of expertise it is crucial to be able filter out OOD samples.

Model We used a dialogGPT Zhang et al. (2019) model fine-tuned on the `Multi WOZ` dataset as chat bot model.

Shift Typology We focused on domain shifts in dialog systems. Indeed dialogues datasets come in all shapes and forms. There are day to day type of dialogues, customers services type of dialogues, open or constrained dialogues. These tasks cover a wide range of domain and often require domain specifics knowledge to perform well Shalaby et al. (2020).

Building domain shifts for dialog We build dataset consisting of pair of dialog data. We use the `Multi WOZ` as IN distribution data since the dialog model we use has been finetuned on it. Then we pair it with datasets from the `SILICONE` benchmark. Chapuis et al. (2020) which consists of a compilation of NLP tasks covering a wide range of domains from daily life to joint tasks, telivision dialogues or phone calls.

Presentation of `SILICONE` The `SILICONE` dataset consists of a collection of widely recognized datasets covering a wide range of domains, tasks and topics.

- *Switchboard Dialog ACT Corpus (SWDA)* contains two sided telephone spee

4.3 METRICS

4.3.1 OUT-OF-DISTRIBUTION DETECTION PERFORMANCE

OOD Detection is usually an unbalanced binary classification problem where the class of interest is OUT. Therefore we can assess the performance of our OOD detection tool focusing on **False alarm rate** and on the **True detection rate**. The **False alarm rate** or False positive rate (FPR) which is the proportion of samples missclassified as OUT. For a score threshold γ, we have FPR $= \Pr\left(s(\mathbf{X}) \leqslant \gamma \,|\, Z = 0\right)$. The **True detection rate** or True positive rate (TPR) is the proportion of OOD sample that are detected by the method. It is theoretically given by For a score threshold γ, we have TPR $= \Pr\left(s(\mathbf{X}) \leqslant \gamma \,|\, Z = 1\right)$.

In order to evaluate the performance of our methods we will focus on two commonly used metrics.

Area Under the Receiver Operating Characteristic curve (`AUROC` ↑) Bradley (1997). The Receiver Operating Characteristic curve is curve obtained by plotting the True positive rate against the False positive rate. The area under this curve is the probability that an in-distribution example $\mathbf{X}_{in}$ has a anomaly score higher than an OOD sample $\mathbf{X}_{out}$: `AUROC` $\uparrow = \Pr(s(\mathbf{X}_{in}) > s(\mathbf{X}_{out}))$. It is given by $\gamma \mapsto (\Pr\left(s(\mathbf{X}) > \gamma \,|\, Z = 0\right), \Pr\left(s(\mathbf{X}) \leqslant \gamma \,|\, Z = 1\right))$.

False Positive Rate at 95% True Positive Rate (`FPR` ↓). We accept to allow only a given false positive rate r corresponding to a defined level of safety and we want to know what share of positive samples we actually catch under this constraint. It leads to select a threshold γ_r such that the corresponding TPR equals r. At this threshold, one then computes: $\Pr(a(\mathbf{X}) \geqslant \gamma_r \,|\, z = 0)$ with γ_r s.t. TPR$(\gamma_r) = r$ r is chosen depending of the difficulty of task at hand and the required level of safety.

4.3.2 MACHINE TRANSLATION PERFORMANCE

The goal of OOD filtering is to remove samples that could hurt the model. However, what matters in practice is the performances of the model on the samples it get to see. For this part we focus on machine translation tasks to leverage automatic translation metrics to evaluate the quality of the filtering applied by our detectors. We measure the average BLEU Papineni et al. (2002) and

BERT-S Unanue et al. (2021) scores of the samples that pass the filter on datasets containing only IN samples, on datasets containing only OOD samples and datasets containing a balanced combination of IN and OOD samples.

5 Experimental results

We evaluated RAINPROOF on our benchmark LOOSENER. We aimed to distinguish OOD samples from IN samples in non-trivial situations and to use these filters to improve model performances. In Table 1 we present our average results on each task (Language shifts, domain shifts and dialog shifts). We showed that RAINPROOF consistently outperformed previous baselines while remaining tractable and plug-and-play. We first discuss our results on Neural Machine Translation tasks in Section 5.1.1 and then our results on dialog tasks in Section 5.1.2.

5.1 Performance in OOD detection

Scenario		Score	Language shifts AUROC ↑	Language shifts FPR ↓	Domain shifts AUROC ↑	Domain shifts FPR ↓	Dialog shifts AUROC ↑	Dialog shifts FPR ↓
$\mathcal{S}_0$	Ours	a_{D_α}	**0.95**	**0.25**	**0.85**	**0.62**	0.79	0.64
		a_{FR}	0.93	0.28	0.81	0.67	0.76	0.68
	Baselines	S_E	0.89	0.44	0.79	0.78	0.65	0.76
		a_{MSP}	0.87	0.44	0.79	0.77	0.66	0.72
		a_L	0.78	0.79	0.73	0.88	0.65	0.95
$\mathcal{S}_1$	Ours	$a_{D^*_\alpha}$	0.88	0.34	**0.86**	**0.50**	**0.86**	**0.52**
		a_{FR^*}	0.88	0.35	0.81	0.69	0.76	0.75
	Baselines	a_M	0.92	0.26	0.78	0.59	0.84	0.55

Table 1: Summary of the performance of our detectors compared to commonly used strong baselines.

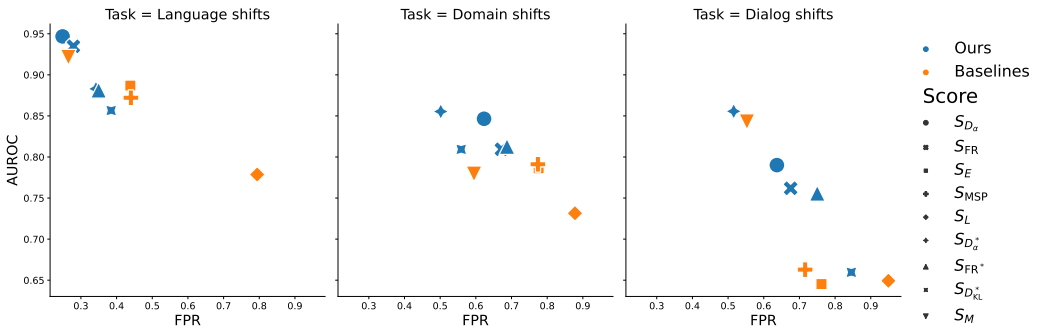

Figure 1: Trade-offs between AUROC ↑and FPR ↓ for each tasks and metrics

5.1.1 Neural Machine Translation

As stated in Section 4.1 we evaluate the performances of our detectors on dataset balanced datasets constituted of different IN and OUT data pairs.

Results on language shifts We show that our detectors outperform every baseline for any language pair in terms of OOD detection (see Table **??** in the appendix). Furthermore, we show that our metrics correlate with BLEU and BERT scores in close OOD data (Table **??**). This correlation translates well in performance gains in practice when filtering out samples that are deemed OOD by our measures (See Figure **??**,Tables 2). It is worth noting that RAINPROOF is not only able to improve the performance on OOD samples but also on *a priori* IN samples.

Results on domain shifts Our detectors also outperform all the baselines on domain shifts. However, it appears that, for these smaller shifts, a reference set is needed to distinguish OOD samples efficiently.

			spa-cat	spa-por	nld-afr	spa:tat-parl	de:news-parl	spa:tat-EMEA	de:news-EMEA
BERT-S	ALL	✓	0.85 +0.19	0.86 +0.29	0.85 +0.10	0.82 +0.02	0.72 +0.01	0.88 +0.01	0.75 +0.02
		✗	0.66	0.57	0.75	0.80	0.71	0.83	0.73
	IN	✓	0.89 +0.01	0.89 +0.01	0.90 +0.01	0.89 +0.02	0.75 +0.01	0.89 +0.01	0.75 +0.01
		✗	0.88	0.88	0.89	0.88	0.74	0.88	0.74
	OOD	✓	0.51 +0.25	0.61 +0.35	0.62 +0.16	0.73 +0.02	0.69 +0.01	0.81 +0.02	0.74 +0.02
		✗	0.26	0.26	0.46	0.71	0.68	0.79	0.72
BLEU	ALL	✓	58 +14	59 +22	58 +9	49 +3	32 +1	62 +3	41
		✗	44	37	49	46	31	59	43
	IN	✓	62 +2	62 +3	63.91 +2	62 +3	35 +1	62 +3	35 +1
		✗	59	59	61	59	34	59	34
	OOD	✓	27 +12	36 +21	34 +11	34 +1	29 +1	62 +3	53 +1
		✗	15	15	23	33	28	59	52

Table 2: Performance gains in translation when filtering OOD samples.

In addition, while yielding smaller gains in translation performance, they still improve the average quality of the translations on OOD data and on IN data.

5.1.2 Application to dialog

Setting

Results Consistently with the results obtained in neural machine translation, uncertainty based detectors fail to discriminate small shifts such as topic changes whereas the information projection is able to discriminate IN and OUT topics. We also show that while our uncertainty based metrics can not compete with reference based ones, they still clearly outperform other common detectors (Fig. 1).

5.1.3 Takeaways

Performance estimation. While our main concern was to filter our OOD samples, we showed that they were also able to exclude *a priori* IN samples that would not be well translated by the model. We discuss this more in depth later on.

Importance of distribution tails. Our results show that when it comes to domain shift (domain shifts in translation or dialog shifts) reference based detectors are required to obtain good results. They also show that the more these detectors take into account the tail of the distributions the better they are as shown in Section B.1 and B.2. We show indeed that low values (close to 0) of α yield better results with the Renyi Information projection S_{D_α}. It is clue that the tail of the distributions used during text generation carry context information and insights on the processed text. Such results are consistent with recent works in automatic evaluation of texts Colombo et al. (2022b).

Comparison to the Mahalanobis distance Our reference based metrics work with a small reference set. In our experiments with used reference set of size 10 to 2000. The mahalanobis distance requires to approximate the covariance matrix of the reference set embeddings. In our case, these embeddings are of dimension 512. Which makes unpractical to correctly approximate the covariance matrix and thus makes the method unreliable in practice while our projection continue to be robust even with really small reference set.

Runtimes Computing the Mahalanobis distance for a sample is fast ($3.10^{-3}s$ in average) provided that we have learnt the covariance matrix but it is still slower than our uncertainty based metrics ($2.10^{-3}s$). This does not take into account the overhead induced by the computation of the covariance matrix, which can take up to several minutes. The information projection does not require precomputation but is, however, significantly slower ($9.10^{-2}s$).

5.2 Link with performance in translation

OOD detection is important because we cannot test the performance of the model outside its training (IN) distribution: we remove the samples for which the model is not tested. Therefore, we prefer not to take any risks since the translation might be inaccurate. However, a more interesting problem is to remove samples that have a higher risk of bad translation altogether and keep those which are likely to be well translated. Hence, we analyse our OOD detectors through the lens of translation performances. An OOD detector should remove samples whose translation would turn out badly and keep those that would be well-translated.

We showed that if common OOD detectors achieve great performances, they do not filter OOD samples as wisely as ours, since they often reduce the average translation performances of the model.

We analysed the effect of filtering OOD samples in 3 settings and present our results in the form of gain (or loss) of BLEU and BERT-S.

Operational setting In real life, IN and OOD samples come at the same time, thus we are interested in the overall performance of the translation model on a dateset composed of IN and OOD samples.

Dataset cleaning setting Even well-tailored datasets contains OOD samples or harder to process samples on which the model can fail even though they are *a priori* IN distribution. We show that we can use OOD detectors to clean out those samples from IN data.

Worst case setting In the worst case scenario a model is confronted only to OOD samples and it is important to know how it would behave in such situation and its performances on the OOD samples that are not filtered out by the OOD detector. We show that the detectors keep a significant part of the OOD dataset but that the translation model still achieve good performance on it.

REFERENCES

Shun-ichi Amari. *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media, 2012.

Alessandro Antonucci, Alessandro Facchini, and Lilith Mattei. Structural learning of probabilistic sentential decision diagrams under partial closed-world assumption. 2021. doi: 10.48550/ARXIV. 2107.12130. URL https://arxiv.org/abs/2107.12130.

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

Michéle Basseville. Divergence measures for statistical data processing—an annotated bibliography. *Signal Processing*, 93(4):621–633, 2013.

Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2636–2648, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.239. URL https://www.aclweb.org/anthology/2020.findings-emnlp.239.

Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pp. 1277–1294. IEEE, 2020.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011.

Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. Learning disentangled textual representations via statistical measures of similarity. *arXiv preprint arXiv:2205.03589*, 2022a.

Pierre Jean A Colombo, Chloé Clavel, and Pablo Piantanida. Infolm: A new metric to evaluate summarization & data2text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10554–10562, 2022b.

Sueli IR Costa, Sandra A Santos, and Joao E Strapasson. Fisher information distance: A geometrical reading. *Discrete Applied Mathematics*, 197:59–69, 2015.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

Imre Csiszár. Sanov property, generalized i-projection and a conditional limit theorem. *The Annals of Probability*, pp. 768–793, 1984.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Geli Fei and Bing Liu. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 506–514, 2016.

Peter Fletcher, Hughes Hoyle, and C Wayne Patty. *Foundations of discrete mathematics*. Brooks/Cole, Florence, KY, November 1990.

Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

Eduardo Dadalto Camara Gomes, Florence Alberge, Pierre Duhamel, and Pablo Piantanida. Igeood: An information geometry approach to out-of-distribution detection. *arXiv preprint arXiv:2203.07798*, 2022.

Matan Haroush, Tzviel Frostig, Ruth Heller, and Daniel Soudry. A statistical framework for efficient out of distribution detection in deep neural networks, 2021.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41–50, 2019.

Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, 1909.

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 and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 2020.

Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Bin Dong, and Xinyu Zhou. Feature space singularity for out-of-distribution detection. *arXiv preprint arXiv:2011.14654*, 2020.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *ArXiv*, abs/2110.00218, 2021.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20578–20589. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ecb9fe2fbb99c31f567e9823e884dbec-Paper.pdf.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1954.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018a.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 7167–7177. Curran Associates, Inc., 2018b. URL http://papers.nips.cc/paper/7947-a-simple-unified-framework-for-detecting-out-of-distribution-samples-and-adve pdf.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018c. URL https://arxiv.org/abs/1807.03888.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

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

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. Cider: Exploiting hyperspherical embeddings for out-of-distribution detection. *arXiv preprint arXiv:2203.04450*, 2022.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. Dqi: Measuring data quality in nlp, 2020. URL https://arxiv.org/abs/2005.00816.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Jitendra Parmar, Satyendra Singh Chouhan, Vaskar Raychoudhury, and Santosh Singh Rathore. Open-world machine learning: Applications, challenges, and opportunities, 2021. URL https://arxiv.org/abs/2105.13448.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. *arXiv preprint arXiv:1905.05702*, 2019.

Marine Picot, Francisco Messina, Malik Boudiaf, Fabrice Labeau, Ismail Ben Ayed, and Pablo Piantanida. Adversarial robustness via fisher-rao regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Julianna Pinele, João E. Strapasson, and Sueli I. R. Costa. The fisher–rao distance between multivariate normal distributions: Special cases, bounds and applications. *Entropy*, 22(4), 2020. ISSN 1099-4300. doi: 10.3390/e22040404. URL https://www.mdpi.com/1099-4300/22/4/404.

Igor M. Quintanilha, Roberto de M. E. Filho, José Lezama, Mauricio Delbracio, and Leonardo O. Nunes. Detecting out-of-distribution samples using low-order deep features statistics, 2019. URL https://openreview.net/forum?id=rkgpCoRctm.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*, pp. 235–247. Springer, 1992.

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*, 2021a.

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, 2021b.

Seonghan Ryu, Seokhwan Kim, Junhwi Choi, Hwanjo Yu, and Gary Geunbae Lee. Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems. *Pattern Recognition Letters*, 88:26–32, 2017.

Ivan N Sanov. *On the probability of large deviations of random variables*. United States Air Force, Office of Scientific Research, 1958.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8491–8501. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sastry20a.html.

Walid Shalaby, Adriano Arantes, Teresa GonzalezDiaz, and Chetan Gupta. Building chatbots from large scale domain-specific knowledge bases: challenges and opportunities, 2020. URL https://arxiv.org/abs/2001.00100.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *ArXiv*, abs/2111.12797, 2021.

Jörg Tiedemann. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 1174–1182, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.139.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012a. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Jrg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012b. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Inigo Jauregi Unanue, Jacob Parnell, and Massimo Piccardi. Berttune: Fine-tuning neural machine translation with bertscore, 2021. URL https://arxiv.org/abs/2106.02208.

Sachin Vernekar, Ashish Gaurav, Vahdat Abdelzad, Taylor Denouden, Rick Salay, and Krzysztof Czarnecki. Out-of-distribution detection in classifiers via generation, 2019a. URL https://arxiv.org/abs/1910.04241.

Sachin Vernekar, Ashish Gaurav, Taylor Denouden, Buu Phan, Vahdat Abdelzad, Rick Salay, and Krzysztof Czarnecki. Analysis of confident-classifiers for out-of-distribution detection. *arXiv preprint arXiv:1904.12220*, 2019b.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pp. 11328–11339. PMLR, 2020.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 9(8):nwac123, 2022.

Qiuyu Zhu, Guohui Zheng, and Yingying Yan. Effective out-of-distribution detection in classifier based on pedcc-loss, 2022. URL https://arxiv.org/abs/2204.04665.

Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

6 APPENDIX

A Experimental setting **16**
A.1 Language pairs 16
A.2 Samples 16
A.3 Additional metrics 16

B Parameters tuning **16**
B.1 Impact of temperatures 16
B.2 Impact of α 16

C Full results **17**
C.1 Detailed results of OOD detection performances 17
C.2 ROC AUC curves 19

D NTM performance gains **20**

E Interpretability **21**

F Negative results **21**
F.1 Negentropy of bag of distributions 21

A EXPERIMENTAL SETTING

A.1 LANGUAGE PAIRS

Model	IN data	OUT data
Language shift		
DE-EN	Tatoeba DE	News FR
DE-EN	Tatoeba DE	Tatoeba NLD
ES-EN	Tatoeba ES	News FR
ES-EN	Tatoeba ES	Tatoeba CAT
ES-EN	Tatoeba ES	Tatoeba POR
NLD-EN	Tatoeba ES	AFR
Domain shift		
DE-EN	Tatoeba DE	EMEA DE
DE-EN	Tatoeba DE	Eurparl DE
ES-EN	Tatoeba ES	EMEA DE
ES-EN	Tatoeba ES	Eurparl DE

Table 3: Summary of models and studied shifts.

A.2 SAMPLES

Source sentence	Expected translation	Translation	BLEU
A en Tom li agrada la tecnologia.	Tom likes technology.	Tom li likes technology.	42.73
Ací està la teua bossa.	Here is your bag.	Ací está la teua bossa.	8.12
Això et posarà en perill.	That'll put you in danger.	Això et posarà en perill.	8.12
A Londres hi han molts parcs bonics.	There are many beautiful parks in London.	To London hi han molts parcs bonics.	6.57
Aquest pa és molt deliciós.	This bread is very delicious.	Aquest pa és molt deliciós.	8.12
A tots els meus amics els agraden els videojocs.	All my friends like playing videogames.	A tots els meus amics els agrade els videojocs.	4.20
Açò és un peix.	This is a fish.	Aaaaaaaaaaaaaaaaaaa ... aaaaaaaaaaaaaaaaaaaaaaaa	0.00
Moltes felicitats!	Congratulations!	Moltes congrats!	27.52
Bon any nou!	Happy New Year!	Bon any nou!	15.97
Aquell que menteix, robarà.	He that will lie, will steal.	The one who's mindless, he'll steal.	12.22
Jo sóc qui té la clau.	I'm the one who has the key.	Jo soc qui te la clau.	5.69
En Tom surt a treballar cada matí a dos quarts de set.	Tom leaves for work at 6:30 every morning.	In Tom surt to pull each matí to two quarts of set.	3.67
Ell m'ha dit que la seva casa era embruixada.	He told me that his house was haunted.	Ell m'ha dit that the seva house was haunted.	27.78
Aquest és el lloc on va nèixer el meu pare.	This is the place where my father was born.	Aquest is the lloc on va nèixer el meu pare.	8.30

Table 4: Example of behavior of a language model trained to handle Spanish inputs on Catalan inputs.

A.3 ADDITIONAL METRICS

For the sake of brevity we present only AUROC ↑and FPR ↓ metrics in our aggregated results but we also used Detection error and Area Under the Precision-Recall curve metrics and those are presented in our full results section (Section C).

Detection error. It is simply the probability of miss-classification for a given True positive rate.

Area Under the Precision-Recall curve (AUPR-IN/AUPR-OUT) Davis & Goadrich (2006). The Precision-Recall curve plots the recall (true detection rate) against the precision (actual proportion of OOD amongst the predicted OOD). The area under this curve $\gamma \mapsto (\Pr\big(Z = 1 \,|\, s(\mathbf{X}) \leqslant \gamma\big), \Pr\big(s(\mathbf{X}) \leqslant \gamma \,|\, Z = 1\big))$ captures the trade-off between precision and recall made by the model. A high value represents a high precision and a high recall *i.e.* the detector captures most of the positive samples while having few False positive.

B PARAMETERS TUNING

B.1 IMPACT OF TEMPERATURES

B.2 IMPACT OF α

Indeed, in Figure 3 we present the impact of the size of the reference set and of the paramet α on Renyi information projection to distinguish dialog shifts, as expected, the larger the reference set, the better. However, we see that smaller values of α yield better results.

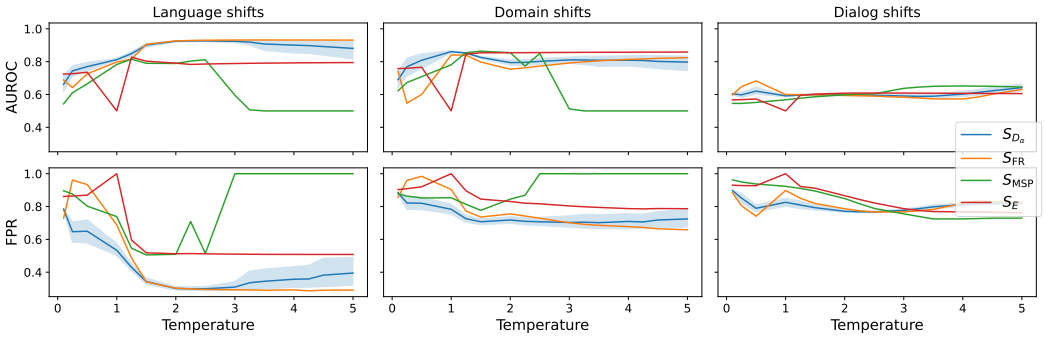

Figure 2: Impact of the temperature on the separation performance of our metrics.

We recall that the Renyi divergence is defined as $D_\alpha(\mathbf{p}\|\mathbf{q}) = \frac{1}{\alpha-1}\log\left(\sum_{i=1}^{|\Omega|}\frac{p_i^\alpha}{q_i^{\alpha-1}}\right)$, where $\alpha \in \mathbb{R}_+ - \{1\}$. Smaller values of α distribute the weight of each feature more equally in the final divergence, more specifically they tend to give an equal weight to the very likely outcome as well as to the less likely ones, therefore giving more weight to the tail of the distribution. When α tends to 0 the Renyi divergence actually counts the number of nonzero common probabilities. That makes sens in terms of topic detection, it counts the common tokens considered during text generation.

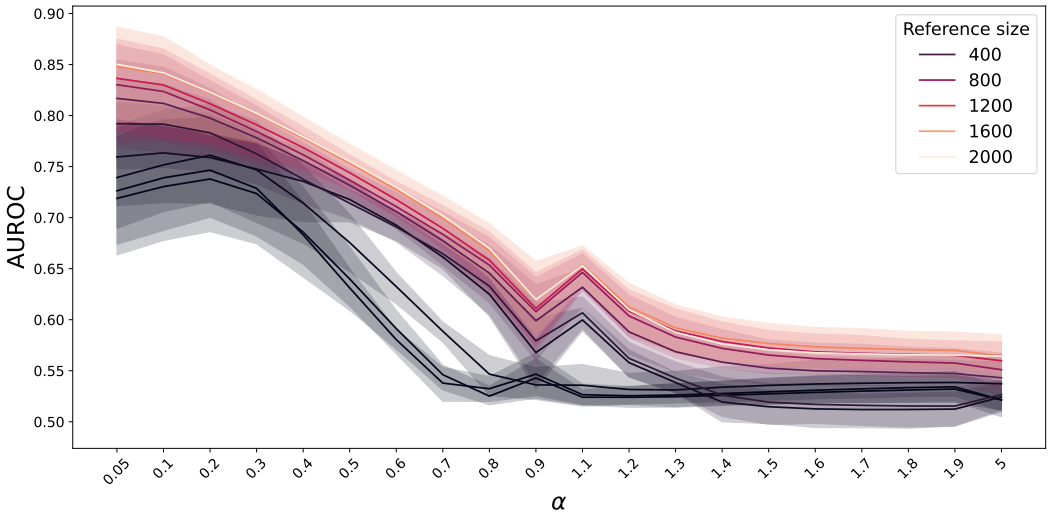

Figure 3: Impact of α on the performance of the Renyi information projection for dialog shift detection. A smaller α increases the weight of the tail of the distribution. An α of 0 would consist in counting the number of the common non zero elements.

C Full results

C.1 Detailed results of OOD detection performances

In this section we present the performances of our OOD detectors on each detailed tasks, *i.e.* for each pair of IN and OOD data with all the considered metrics. We show that our metrics outperform other OOD detectors baselines in almost all scenarios.

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**1.00**	**0.99**	**1.00**	**0.03**	**0.01**
		a_{FR}	0.99	**0.99**	0.99	0.05	0.04
	Baselines	a_E	0.98	0.95	0.98	0.05	0.05
		a_L	0.81	0.75	0.80	0.41	0.78
		a_{MSP}	0.99	**0.99**	0.99	0.04	0.03
s_1	Ours	$a_{D^*_\alpha}$	**1.00**	**1.00**	**1.00**	**0.03**	**0.01**
		$a_{D^*_{\text{KL}}}$	0.98	0.99	0.99	**0.03**	0.02
		a_{FR^*}	**1.00**	**1.00**	**1.00**	**0.03**	**0.01**
	Baselines	a_M	**1.00**	**1.00**	**1.00**	**0.03**	**0.01**

(a) deu-fr

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.76**	**0.91**	**0.85**	**0.51**	**0.72**
		a_{FR}	0.71	0.88	0.80	0.53	0.76
	Baselines	a_E	0.74	0.88	0.82	0.56	0.79
		a_L	0.73	0.88	0.82	0.58	0.82
		a_{MSP}	0.70	0.88	0.81	0.56	0.80
s_1	Ours	$a_{D^*_\alpha}$	**0.71**	**0.94**	**0.87**	**0.23**	**0.31**
		$a_{D^*_{\text{KL}}}$	0.61	**0.94**	0.84	0.25	0.35
		a_{FR^*}	**0.71**	**0.94**	**0.87**	**0.23**	**0.31**
	Baselines	a_M	0.62	0.93	0.85	0.29	0.39

(b) nld-afr

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.86**	**0.93**	**0.91**	**0.30**	**0.44**
		a_{FR}	0.85	0.92	0.89	0.35	0.51
	Baselines	a_E	0.71	0.82	0.77	0.58	0.87
		a_L	0.67	0.79	0.75	0.59	0.88
		a_{MSP}	0.68	0.83	0.73	0.58	0.87
s_1	Ours	$a_{D^*_\alpha}$	0.56	0.77	0.68	0.59	0.88
		$a_{D^*_{\text{KL}}}$	0.49	0.74	0.64	0.60	0.91
		a_{FR^*}	0.54	0.76	0.67	0.60	0.90
	Baselines	a_M	**0.71**	**0.88**	**0.81**	**0.41**	**0.60**

(c) spa-cat

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**1.00**	**1.00**	**1.00**	**0.02**	**0.00**
		a_{FR}	**1.00**	**1.00**	**1.00**	0.03	0.01
	Baselines	a_E	0.99	0.97	0.98	0.04	0.04
		a_L	0.70	0.75	0.75	0.38	0.72
		a_{MSP}	0.98	0.99	0.99	0.05	0.04
s_1	Ours	$a_{D^*_\alpha}$	**0.99**	**0.99**	**0.99**	**0.03**	**0.02**
		$a_{D^*_{\text{KL}}}$	0.90	0.96	0.94	0.09	0.13
		a_{FR^*}	**0.99**	**0.99**	**0.99**	**0.03**	**0.02**
	Baselines	a_M	0.93	0.94	0.94	0.18	0.31

(d) spa-fr

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.94**	**0.93**	**0.94**	**0.18**	**0.32**
		a_{FR}	**0.94**	**0.93**	**0.94**	0.20	0.34
	Baselines	a_E	0.82	0.75	0.79	0.44	0.83
		a_L	0.79	0.70	0.77	0.46	0.88
		a_{MSP}	0.77	0.77	0.74	0.45	0.84
s_1	Ours	$a_{D^*_\alpha}$	0.79	0.70	0.76	0.44	0.84
		$a_{D^*_{\text{KL}}}$	0.75	0.70	0.74	0.46	0.87
		a_{FR^*}	0.78	0.71	0.76	0.45	0.86
	Baselines	a_M	**0.92**	**0.95**	**0.94**	**0.15**	**0.24**

(e) spa-por

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**1.00**	**0.99**	**1.00**	**0.03**	**0.00**
		a_{FR}	0.99	**0.99**	0.99	0.04	0.02
	Baselines	a_E	0.98	0.95	0.98	0.05	0.05
		a_L	0.77	0.77	0.80	0.37	0.69
		a_{MSP}	0.98	0.98	0.98	0.05	0.05
s_1	Ours	$a_{D^*_\alpha}$	**1.00**	**1.00**	**1.00**	**0.03**	**0.00**
		$a_{D^*_{\text{KL}}}$	0.97	0.99	0.99	0.04	0.03
		a_{FR^*}	**1.00**	**1.00**	**1.00**	**0.03**	**0.00**
	Baselines	a_M	0.99	**1.00**	0.99	0.04	0.03

(f) deu-nl

Table 5: Detailed results of the performances of our OOD detectors on different language shifts. The first language of the pair is the reference language of the model and the second one is the studied shift.

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.90**	0.75	**0.86**	0.44	0.83
		a_{FR}	0.86	0.71	0.81	0.46	0.87
	Baselines	a_E	0.87	0.75	0.83	0.47	0.89
		a_L	0.86	0.73	0.82	0.48	0.90
		a_{MSP}	0.89	**0.76**	0.85	**0.43**	**0.81**
s_1	Ours	$a_{D^*_\alpha}$	**0.89**	**0.87**	**0.89**	**0.27**	**0.49**
		$a_{D^*_{\text{KL}}}$	**0.89**	0.85	**0.89**	0.29	0.52
		a_{FR^*}	0.83	0.66	0.77	0.51	0.97
	Baselines	a_M	0.80	0.79	0.79	0.35	0.66

(a) de:news-EMEA

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.95**	**0.91**	**0.94**	**0.19**	**0.33**
		a_{FR}	0.91	0.89	0.91	0.24	0.42
	Baselines	a_E	0.91	0.83	0.90	0.32	0.59
		a_L	0.78	0.69	0.75	0.45	0.86
		a_{MSP}	0.90	0.88	0.89	0.36	0.68
s_1	Ours	$a_{D^*_\alpha}$	**0.97**	0.94	**0.96**	0.13	0.20
		$a_{D^*_{\text{KL}}}$	0.95	**0.96**	**0.96**	**0.12**	**0.18**
		a_{FR^*}	0.95	0.89	0.94	0.25	0.44
	Baselines	a_M	0.92	0.92	0.92	0.20	0.35

(b) spa:tat-EMEA

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.75**	**0.75**	**0.76**	**0.35**	**0.66**
		a_{FR}	0.68	0.72	0.71	0.39	0.72
	Baselines	a_E	**0.75**	**0.75**	0.75	0.41	0.77
		a_L	0.65	0.69	0.71	0.42	0.79
		a_{MSP}	**0.75**	**0.75**	**0.76**	0.39	0.73
s_1	Ours	$a_{D^*_\alpha}$	**0.89**	**0.88**	**0.90**	0.27	**0.48**
		$a_{D^*_{\text{KL}}}$	0.65	0.77	0.72	0.34	0.63
		a_{FR^*}	0.88	**0.88**	**0.90**	**0.26**	**0.48**
	Baselines	a_M	0.77	0.82	0.78	0.28	0.51

(c) spa:tat-parl

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.75**	**0.75**	**0.70**	**0.43**	**0.81**
		a_{FR}	0.60	0.64	0.63	0.45	0.85
	Baselines	a_E	**0.75**	**0.75**	0.67	0.45	0.85
		a_L	0.63	0.58	0.64	0.51	0.96
		a_{MSP}	**0.75**	**0.75**	0.67	0.46	0.87
s_1	Ours	$a_{D^*_\alpha}$	**0.68**	**0.64**	**0.68**	**0.44**	**0.83**
		$a_{D^*_{\text{KL}}}$	0.65	0.62	0.67	0.48	0.90
		a_{FR^*}	0.62	0.63	0.65	0.45	0.86
	Baselines	a_M	0.64	0.62	0.63	0.46	0.87

(d) de:news-parl

Table 6: Detailed results of the performances of our OOD detectors on different domain shifts. For Spanish (spa) and German (de), we present two domains shifts: Technical medical (EMEA) data and legal parlementary texts (parl) against common language emboddied by the Tatoeba dataset (tat).

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	0.71	0.74	**0.72**	**0.36**	0.67
		a_{FR}	0.68	0.73	**0.72**	**0.36**	0.67
	Baselines	a_E	**0.75**	**0.75**	0.66	0.39	0.74
		a_L	0.69	0.48	0.57	0.50	1.00
		a_{MSP}	0.61	0.74	0.71	**0.36**	**0.66**
s_1	Ours	$a_{D^*_\alpha}$	0.88	0.89	0.88	0.24	0.44
		$a_{D^*_{\text{KL}}}$	0.68	0.70	0.70	0.42	0.80
		a_{FR^*}	0.79	0.80	0.80	0.36	0.68
	Baselines	a_M	0.87	0.95	0.93	0.10	0.21

(a) swda

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	0.52	**0.87**	**0.72**	**0.52**	**0.69**
		a_{FR}	0.41	0.86	0.69	0.55	0.73
	Baselines	a_E	**0.63**	**0.87**	0.63	0.56	0.74
		a_L	0.40	0.66	0.47	0.73	1.00
		a_{MSP}	0.36	0.86	0.67	**0.52**	**0.69**
s_1	Ours	$a_{D^*_\alpha}$	0.69	0.93	0.85	0.40	0.53
		$a_{D^*_{\text{KL}}}$	0.42	0.84	0.67	0.63	0.85
		a_{FR^*}	0.53	0.88	0.75	0.58	0.77
	Baselines	a_M	0.53	0.87	0.74	0.73	0.99

(b) melde

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	0.79	**0.80**	0.81	**0.36**	**0.68**
		a_{FR}	0.68	0.73	0.70	0.39	0.74
	Baselines	a_E	0.75	0.75	0.64	0.45	0.84
		a_L	**0.91**	0.73	**0.85**	0.49	0.94
		a_{MSP}	0.70	0.65	0.64	0.43	0.80
s_1	Ours	$a_{D^*_\alpha}$	0.92	0.91	0.91	0.23	0.42
		$a_{D^*_{\text{KL}}}$	0.60	0.60	0.61	0.47	0.89
		a_{FR^*}	0.75	0.74	0.75	0.41	0.77
	Baselines	a_M	0.85	0.93	0.91	0.10	0.21

(c) dydae

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	0.79	**0.80**	0.81	**0.36**	**0.68**
		a_{FR}	0.68	0.73	0.70	0.39	0.74
	Baselines	a_E	0.75	0.75	0.64	0.45	0.84
		a_L	**0.91**	0.73	**0.85**	0.49	0.94
		a_{MSP}	0.70	0.65	0.64	0.43	0.80
s_1	Ours	$a_{D^*_\alpha}$	0.92	0.91	0.91	0.23	0.42
		$a_{D^*_{\text{KL}}}$	0.60	0.60	0.61	0.47	0.89
		a_{FR^*}	0.75	0.74	0.75	0.41	0.77
	Baselines	a_M	0.85	0.93	0.91	0.10	0.21

(d) dydada

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	0.52	**0.87**	**0.72**	**0.52**	**0.69**
		a_{FR}	0.41	0.86	0.69	0.55	0.73
	Baselines	a_E	**0.63**	**0.87**	0.63	0.56	0.74
		a_L	0.40	0.66	0.47	0.73	1.00
		a_{MSP}	0.36	0.86	0.67	**0.52**	**0.69**
s_1	Ours	$a_{D^*_\alpha}$	0.69	0.93	0.85	0.40	0.53
		$a_{D^*_{\text{KL}}}$	0.42	0.84	0.67	0.63	0.85
		a_{FR^*}	0.53	0.88	0.75	0.58	0.77
	Baselines	a_M	0.53	0.87	0.74	0.73	0.99

(e) melds

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	**0.87**	**0.87**	**0.87**	**0.31**	**0.56**
		a_{FR}	0.73	0.81	0.79	0.34	0.63
	Baselines	a_E	0.75	0.75	0.62	0.40	0.75
		a_L	0.81	0.79	0.82	0.39	0.72
		a_{MSP}	0.53	0.64	0.57	0.42	0.78
s_1	Ours	$a_{D^*_\alpha}$	0.69	0.69	0.72	0.43	0.81
		$a_{D^*_{\text{KL}}}$	0.66	0.68	0.70	0.44	0.82
		a_{FR^*}	0.69	0.69	0.72	0.43	0.82
	Baselines	a_M	0.87	0.84	0.87	0.33	0.61

(f) dailydialog-default

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	0.42	**0.91**	**0.72**	0.57	**0.70**
		a_{FR}	0.41	0.90	**0.72**	0.57	0.71
	Baselines	a_E	**0.61**	0.89	0.63	0.61	0.76
		a_L	0.53	0.76	0.58	0.80	1.00
		a_{MSP}	0.35	0.90	0.68	**0.56**	**0.70**
s_1	Ours	$a_{D^*_\alpha}$	0.67	0.95	0.86	0.44	0.55
		$a_{D^*_{\text{KL}}}$	0.29	0.85	0.63	0.72	0.90
		a_{FR^*}	0.41	0.90	0.73	0.61	0.77
	Baselines	a_M	0.42	0.89	0.70	0.78	0.99

(g) iemocap

Scenario		Score	AUPR-IN	AUPR-OUT	AUROC	ERR	FPR
s_0	Ours	a_{D_α}	0.71	**0.77**	**0.75**	0.33	0.60
		a_{FR}	0.69	0.74	0.74	0.36	0.67
	Baselines	a_E	**0.75**	0.75	0.70	0.37	0.68
		a_L	0.72	0.48	0.57	0.50	1.00
		a_{MSP}	0.62	0.76	0.73	**0.32**	**0.59**
s_1	Ours	$a_{D^*_\alpha}$	0.86	0.88	0.87	0.23	0.42
		$a_{D^*_{\text{KL}}}$	0.66	0.71	0.70	0.41	0.77
		a_{FR^*}	0.77	0.80	0.79	0.36	0.66
	Baselines	a_M	0.88	0.96	0.94	0.10	0.21

(h) mrda

Table 7: Detailed performance results of our OOD detectors on dialog shift against the Multi WOZ dataset as reference set.

C.2 ROC AUC CURVES

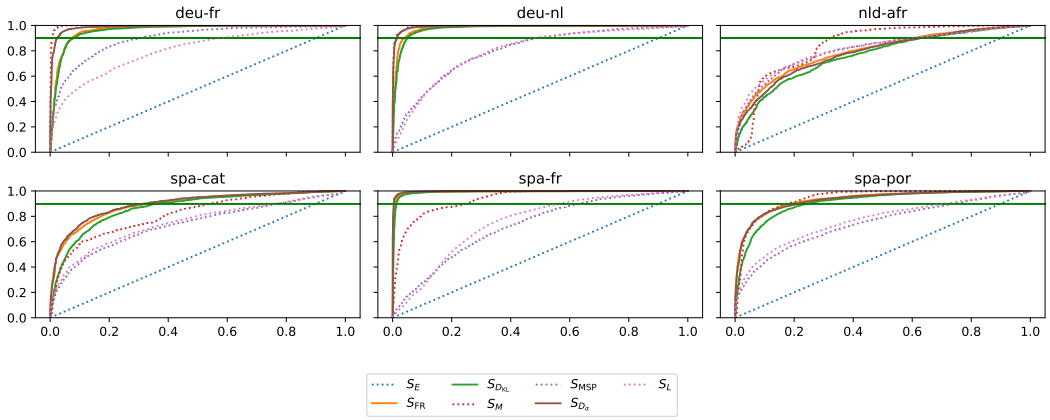

Figure 4: ROCAUC curves for our uncertainty based metrics compared to common baselines on language shifts. Baselines are represented in dashed lines.

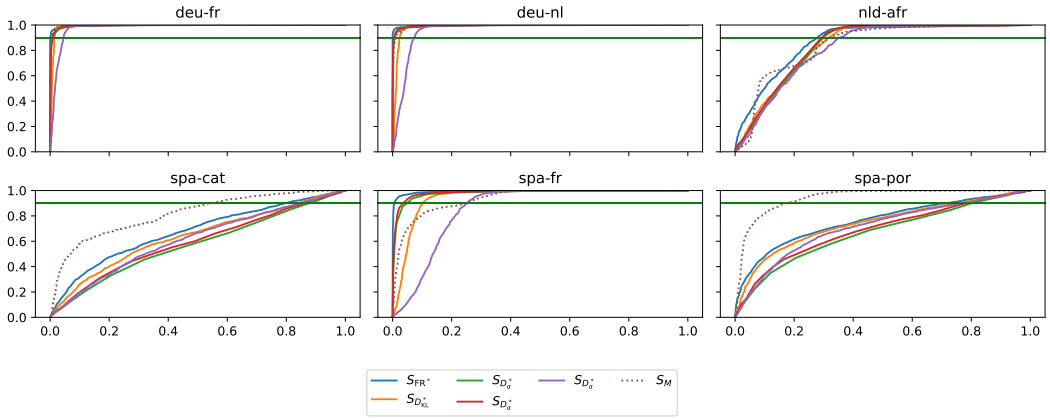

Figure 5: ROCAUC curves for our reference based metrics compared to common baselines. Baselines are represented in dashed lines.

D NTM PERFORMANCE GAINS

Surprisingly we show that common OOD detectors tend to exclude samples that are well handled by the model and keep some that are not leading to decreasing overall performance in terms of translation metrics. Moreoever it seems this phenomenon is more dominant in reference based detectors. We show that our uncertainty based detectors mostly avoir that downfall and provide good OOD detection and improved translation performances.

Scenario		Score	spa-cat	spa-por	nld-afr	spa:tat-parl	de:news-parl	spa:tat-EMEA	de:news-EMEA
$\mathbb{s}_0$	Ours	a_{D_α}	+2.38	+2.38	+2.19	+2.38	+1.10	+2.38	+1.10
		a_{FR}	+2.69	+2.69	+2.51	+2.69	+1.53	+2.69	+1.53
	Baselines	a_E	-1.87	-1.87	-2.28	-1.87	-1.76	-1.87	-1.76
		a_L	-4.53	-4.53	-3.78	-4.53	-3.08	-4.53	-3.08
		a_{MSP}	-4.88	-4.88	-4.28	-4.88	-3.21	-4.88	-3.21
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	+0.79	+0.79	+0.97	+0.79	-0.29	+0.79	-0.29
		a_{FR^*}	+0.81	+0.81	+0.97	+0.81	-0.29	+0.81	-0.29
	Baselines	a_M	+0.06	+0.06	+1.27	+0.06	+0.19	+0.06	+0.19

(a) IN

Scenario		Score	spa-cat	spa-por	nld-afr	spa:tat-parl	de:news-parl	spa:tat-EMEA	de:news-EMEA
$\mathbb{s}_0$	Ours	a_{D_α}	+12.06	+20.64	+11.15	+1.28	+0.67	+3.24	+1.46
		a_{FR}	+12.05	+19.71	+11.11	+3.13	+1.17	+5.64	+2.68
	Baselines	a_E	+1.07	+2.88	-0.34	-2.68	-2.31	-19.14	-19.91
		a_L	-0.32	+0.10	-0.87	-1.72	-3.25	-14.56	-23.93
		a_{MSP}	-0.23	+0.18	-1.24	-1.76	-3.48	-14.91	-25.12
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	+4.25	+7.81	+9.54	+0.27	-0.67	-9.75	-7.25
		a_{FR^*}	+4.22	+7.72	+9.63	+0.23	-0.71	-9.84	-7.28
	Baselines	a_M	+0.01	+8.14	+4.07	-2.00	-1.83	-10.23	-7.78

(b) OOD

Scenario		Score	spa-cat	spa-por	nld-afr	spa:tat-parl	de:news-parl	spa:tat-EMEA	de:news-EMEA
$\mathbb{s}_0$	Ours	a_{D_α}	+14.50	+21.50	+9.11	+2.92	+0.69	+2.65	-1.98
		a_{FR}	+14.67	+22.02	+9.63	+4.74	+1.22	+3.46	+0.29
	Baselines	a_E	+0.25	+2.79	-2.83	-2.91	-2.23	-4.57	-10.79
		a_L	-4.13	-3.54	-4.07	-4.03	-3.28	-8.10	-12.79
		a_{MSP}	-4.38	-3.76	-4.67	-4.34	-3.46	-8.46	-13.16
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	+5.65	+10.91	+9.04	+10.26	-0.71	-0.05	-5.55
		a_{FR^*}	+5.66	+10.85	+9.10	+10.25	-0.71	-0.04	-5.52
	Baselines	a_M	+7.56	+18.78	+7.73	+1.98	-0.50	-1.95	-5.47

(c) ALL

Table 8: Effect different OOD detectors on the overall translation system performance in terms of translation metrics on different dataset settings: an operational setting with ALL data (IN and OOD), a worst case scenario setting with only OOD samples (OOD) and a cleaning data scenario with *a priori* IN data only.

Scenario		Dataset	spa-cat		spa-por		nld-afr		spa:tat-parl		de:news-parl		spa:tat-EMEA		de:news-EMEA	
		Score	ALL	OOD	ALL	OOD	ALL	OOD	ALL	OOD	ALL	OOD	ALL	OOD	ALL	OOD
$\mathbb{s}_0$	Ours	a_{D_α}	42%	83%	55%	89%	34%	63%	26%	32%	15%	9%	53%	85%	41%	62%
		a_{FR}	42%	83%	55%	90%	34%	66%	30%	40%	16%	12%	48%	77%	33%	46%
	Baselines	a_E	23%	29%	28%	36%	16%	8%	16%	12%	15%	9%	52%	84%	48%	76%
		a_L	14%	8%	12%	6%	13%	4%	11%	4%	16%	12%	36%	53%	46%	73%
		a_{MSP}	16%	9%	14%	7%	15%	5%	12%	4%	17%	13%	37%	54%	47%	74%
$\mathbb{s}_1$	Ours	$a_{D^*_\alpha}$	30%	47%	41%	62%	37%	73%	53%	87%	13%	7%	56%	92%	39%	59%
		a_{FR^*}	29%	47%	40%	61%	37%	74%	53%	87%	12%	6%	56%	92%	38%	58%
	Baselines	a_M	37%	67%	56%	91%	35%	68%	34%	47%	26%	32%	50%	79%	40%	60%

Table 9: Share of the datasets removed when taking γ so that we keep 80% of the IN distribution.

E INTERPRETABILITY

F NEGATIVE RESULTS

F.1 NEGENTROPY OF BAG OF DISTRIBUTIONS

We introduced in Section **??** the bag of distributions as a way to aggregate a sequence of probability distribution and compare it to a set of reference using information projections **??**. A natural idea would be to apply the Negentropy methods (Section 3.2) to these aggregated distributions.

Source sentence	Expected translation	Translation	Projection score	Closest	BLEU	BERT-S
Aguanta.	Stick it out.	Hang in there.	0.34	Véngate.	15.97	0.65
Aquesta és una carta molt estranya.	This is a very strange letter.	This is a molt estranya card.	0.74	Este carro es chiquito.	26.27	0.34
Amb qui estàs?	Who are you with?	Who are you?	0.02	¿Estás bien?	49.76	0.65

(a) Catalan samples projected on a Spanish reference set.

Source sentence	Expected translation	Translation	Projection score	Closest	BLEU	BERT-S
Wofür wird Abilify angewendet?	What is Abilify used for?	What is Abilify used for?	0.91	Für was bin ich hier?	100.00	1.00
Es können auch depressive Episoden auftreten.	They may also have episodes of depression.	Depressive episodes may also occur.	1.26	Kann sein, dass du gebraucht wirst.	13.83	0.73

(b) German samples from EMEA projected on a Tatoeba reference set.

Table 10: Intrepretable information projection and performance in translation.

More formally given a sequence of probability distribution $\mathcal{S}_\theta(\mathbf{x}) = \{p_\theta^T(, \hat{\mathbf{y}}_{\leqslant t})\}_{t=1}^n$ we would compute its bag of distributions:

$$\bar{p}_\theta(x) \triangleq \frac{1}{|y|} \sum_{t=1}^{|y|} p_\theta(x, y_{\leqslant t}) \tag{3}$$

And then compute as novelty score:

$$J_D(\mathbf{p}) = D(\mathbf{p} \| \mathcal{U}) \tag{4}$$

Further experiments have shown that this process was unable to discriminate OOD samples or improve performance translation. We suspect that the uncertainty at each step is key to capture the behavior of the language model and that this uncertainty information is lost when averaging probability distribution along the sequence.