Intrinsic Uncertainty-Aware Calibration Metric

Anonymous ACL submission

Abstract

Deep learning models have made great strides in recent years. Subsequently, model calibration and measurements of the quantity have gained much attention, with the degree being an indication of reliability of a model. In this study, we explore the limitations of the existing calibration metrics, and propose a simple calibration metric that caters to natural language generation (NLG) tasks. Unlike existing calibration metrics, our metric is not confined to/not sorely based on a single prediction; it considers a distribution mapped by a model. In this regard, the proposed metric takes intrinsic uncertainty present in a natural language into account when quantifying the calibration degree. The metric has been tested on machine translation datasets, a popular NLG task with intrinsic uncertainty. A thorough analysis illustrates that the proposed metric possesses the ability to handle intrinsic uncertainty and hence is more suitable measure under NLG tasks.

1 Introduction

A predictive score of a well-calibrated model reflects the true likelihood of correctness (Guo et al., 2017; Jiang et al., 2021). Therefore, model calibration demonstrates the reliability of a model (Nixon et al., 2019), answering the question of *"how much we can trust a decision made by a model"*. Deep learning models are being applied to various sectors of society. For this reason, not only the performance but calibration is of significance (Tomani and Buettner, 2021).

The growing importance has introduced calibration measures with the aim of accurate quantification of the quantity (Naeini et al., 2015; Nixon et al., 2019; Guo et al., 2017; Ding et al., 2021; Jagannatha and Yu, 2020). The calibration metrics have been utilized to test the trustworthiness of a model not just in safety-critical domain (Mehrtash et al., 2020), but also in image classification (Krishnan and Tickoo, 2020), text classification (Jung et al., 2020), and *text generation* (Müller et al., 2019; Wang et al., 2020).

Of the domains mentioned above, accurate evaluation of a language model (LM)'s calibration is of most significance *in regards to model output*. Model calibration does not change model prediction of a classifier; on the contrary, model calibration affects model output of an LM (Müller et al., 2019). Common generation schemes of an autoregressive LM, such as top-p (Holtzman et al., 2020), top-k sampling (Fan et al., 2018) and beam search, are grounded on an assumption that a predictive score represents the likelihood of the word in a given context (Holtzman et al., 2020). In this regard, when a probability distribution is not calibrated, the assumption fails to hold, leading to the degradation in quality of an output (Müller et al., 2019). Therefore, an accurate measure of model calibration of an LM is in need.

In this paper, we discuss several limitations of existing calibration metrics, especially from the perspective of NLG tasks; the measures do not take the ***intrinsic uncertainty*** of a natural language into consideration. A semantic equivalence can be achieved with a variable size of utterances; this aspect of a natural language is referred to as intrinsic uncertainty (Ott et al., 2018). Previous calibration metrics overlook this aspect in evaluating an LM, consequently generating inaccurate approximations.

To this end, we propose e-ECE, a calibration metric designed to evaluate model calibration in - but not limited to - NLG tasks. The metric intakes **a distribution**, reflecting intrinsic uncertainty in the course of evaluation. We empirically find that the measure lowers the level of mis-calibration error brought by the uncertainty that has otherwise remained as error in previous metrics.

The contributions of our work are as follows:

- Our work discusses the limitations of the existing calibration metrics under NLG environment.

- We present e-ECE, a calibration metric that is designed for NLG environment by considering the intrinsic uncertainty of a natural language in computing model calibration.
- e-ECE is stable, evaluates broader pool of generation schemes, and, with high accuracy, quantifies model calibration, its level superior to that of the existing metrics.

2 Preliminaries & Related Work

2.1 Calibration

A model calibration is a measure of how predictive scores truly reflect the accuracy of predictions (Guo et al., 2017). In this paper, perfect calibration is defined as follows[1]:

$$P(\hat{Y} = Y | \hat{P} = p) = p, \quad \forall p \in [0, 1] \tag{1}$$

where $\hat{Y}$ and $\hat{P}$ indicate model predictions and corresponding confidence scores (predictive scores). In plain English, the predicted probability should match the accuracy when given a calibrated model; model predictions with 0.5 predictive scores are expected to achieve 50 percent accuracy. Therefore, the quantity is an indication of trustworthiness of a model prediction (Tomani and Buettner, 2021).

Naeini et al. (2015) approximate Equation 1 with a binning approach. The test predictions are binned to M bins based on their predictive scores. The accuracy and confidence of each bin are computed as follows:

$$\begin{aligned} \mathrm{acc}(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} a^{(i)}, \quad a^{(i)} \in \{0, 1\} \\ \mathrm{conf}(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} p^{(i)} \end{aligned} \tag{2}$$

where B_m refers to the m-th bin and $a^{(i)}$ is computed with an indicator function $\mathbb{1}(y^{(i)} = \hat{y}^{(i)})$. Reliability diagram (DeGroot and Fienberg, 1983) visualizes the gap between the accuracy and confidence of each bin. ECE is a weighted sum of the differences, where the weights are proportional to the size of bins.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{\sum_{j=1}^{M} |B_j|} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)| \tag{3}$$

[1]The notations are borrowed from (Guo et al., 2017)

Dataset	Greedy	Pure	Top-k	Top-p
WMT14 EN→DE	1.52	1.31	1.73	1.79
IWSLT14 DE→EN	7.14	5.24	4.91	5.34
Multi30K DE→EN	10.39	8.21	7.39	8.10

Table 1: ECE scores from different generation methods. **Pure** denotes pure sampling, and k and p are set to 100 and 0.8 respectively.

Therefore, ECE can be viewed as aggregation of *bin-wise absolute differences* between accuracy and confidence. A low ECE score indicates that predictive scores reflect the actual accuracy of predictions, and hence well-calibration.

Other variants of the metric have also been introduced. Nixon et al. (2019) propose Static Calibration Error (SCE) and Adaptive Calibration Error (ACE); the former quantifies class-wise calibration error, while the latter utilizes adaptive intervals when binning predictions.

3 Analysis of Problems in Existing Calibration Metrics

Approximation Problem ECE is an approximation made to compute model calibration error. Therefore, finite samples in test dataset may not be sufficient to assess the true calibration of a model. The problem stands out more clearly when an output space is exponentially large, or under imbalanced label distribution (Zipf's Law), two cases under which NLG tasks fall. For instance, in WMT14 English to German (EN→DE) translation dataset, 49.7% of labels (words) are not present in test dataset. Furthermore, due to the limited number of test samples, the intrinsic uncertainty of a language is hardly reflected. That is, the calibration error computed with previous metrics inevitably includes intrinsic uncertainty in the value.

Generation-Specific Problem The current approach in quantifying the calibration of an NLG model is formulated as Next Token Prediction (NTP) task (Müller et al., 2019).

$$\begin{aligned} \hat{p}_t^{(i)} &= \max_{y \in Y} P(y | x^{(i)}, y_{1:t-1}^{(i)}; \theta) \\ \hat{y}_t^{(i)} &= \arg\max_{y \in Y} P(y | x^{(i)}, y_{1:t-1}^{(i)}; \theta) \end{aligned} \tag{4}$$

where $y_{1:t-1}^{(i)}$ denotes ground truth prefix at time step t. However, the current approach is far from ideal, since the approach **only considers greedy generation** scheme with the $\arg\max$ operation in

Equation 4. ECE is designed for binary classification, leaving the probabilities on *other classes unassessed* (Nixon et al., 2019). This is a clear limitation, especially in sequence generation. Generating the most probable sequence is known to be dull and repetitive (Fu et al., 2021), degenerating the output. Therefore, sampling-based methods, such as top-k and top-p (Holtzman et al., 2020), are commonly adopted (Fan et al., 2018; Edunov et al., 2018; Tian et al., 2020). Table 1 illustrates how the choice of generation scheme can drastically change the calibration error of the same model. Nonetheless, the existing calibration metrics fail to address the issue, hence being **suboptimal measures in NLG environment.**

4 Approach

In light of the limitations, our work proposes a calibration metric that reflects intrinsic uncertainty of a language in evaluation.

4.1 e-Expected Calibration Error (e-ECE)

Existing calibration metrics take a single prediction and corresponding confidence score for each test sample in computing model calibration. This differs in the proposed metric: e-ECE. e-ECE takes *expectation over a probability distribution and the corresponding accuracy*[2].

$$\begin{aligned}\tilde{p}_t^{(i)} &= \mathbb{E}_{\tilde{y}\sim\tilde{P}_\theta}[P(\tilde{y}|y_{1:t-1}^{(i)}, x^{(i)};\theta)] \\ \tilde{a}_t^{(i)} &= \mathbb{E}_{\tilde{y}\sim\tilde{P}_\theta}[\mathbb{1}(y_t^{(i)} = \tilde{y})], \tilde{a}_t^{(i)} \in [0,1]\end{aligned} \quad (5)$$

$\tilde{P}_\theta$[3] is a post-processed probability distribution; a distribution mapped by a model can be scaled with a temperature τ or confined to a subset of output space, as in top-k or top-p sampling. $\tilde{p}_t^{(i)}$ and $\tilde{a}_t^{(i)}$ denote expected confidence over the output space and expected accuracy respectively. The expectations can be taken from j samples drawn from the probability distribution $\tilde{P}_\theta$, or simply from the whole output space whose details are illustrated in Appendix A. Once the expected values are computed, the remaining binning approach, Equation 2 and 3, stays identical to that of ECE.

[2]The "expectation" differs from that of Expected Calibration Error (ECE), as ECE takes expectation over a test dataset, while our "expectation" is performed on probability distribution of a test instance.

[3]For notational simplicity, we denote $\tilde{P}(\tilde{y}|y_{1:t-1}^{(i)}, x^{(i)};\theta)$ as $\tilde{P}_\theta$ hereinafter.

Dataset	ECE	SCE	ACE	e-ECE	e-ECE$_p$	e-ECE$_k$
WMT14	1.52	14.79	15.56	1.37	1.52	3.16
IWSLT14	7.14	13.39	13.87	4.42	5.18	4.91
Multi30K	10.39	18.68	19.76	6.87	7.92	8.50

Table 2: The comparison between the existing calibration metrics and e-ECE on the corpora tested. e-ECE$_p$ and e-ECE$_k$ denote e-ECE with top-p and top-k sampling generation scheme respectively.

4.2 Analysis

4.2.1 Accurate Approximation

The existing metrics fail to address the intrinsic uncertainty of a language, and thus a portion of calibration error computed with the metrics is attributed to intrinsic uncertainty. However, this problem is mitigated in the proposed approach, producing a more accurate approximation of model calibration. We empirically validate this aspect in Section 5.1.

4.2.2 Theoretical Connection to ECE

e-ECE subsumes ECE metric.

Proposition 1. *With a small temperature value $\tau \approx 0$, e-ECE converges to ECE metric.*

$$\lim_{\tau\to 0} e\text{-ECE} = \text{ECE} \quad (6)$$

Please refer to Appendix B for the proof. The close connection between e-ECE and ECE enables wide application of the proposed metric in tasks other than NLG.

4.2.3 Broader Generation Schemes

e-ECE expands the scope of evaluation; sampling-based generation schemes can now be evaluated. When the sampling space of e-ECE is confined to a certain set of indexes using top-k or top-p, the metric quantifies the calibration of the generation schemes. In addition, since the metric takes expectation of a distribution, beam search is also a subject of *implicit* evaluation in e-ECE.

5 Experiments

We evaluate the proposed metric on three popular machine translation datasets with intrinsic uncertainty: `Multi30K DE`→`EN`, `IWSLT14 DE`→`EN`, and `WMT14 EN`→`DE`[4].

The results from the calibration metrics are described in Table 2. We observe a marked decrease in

[4]The detailed description on the datasets can be found in Appendix C

Quantile	WMT14		IWSLT14		Multi30K	
	$\mathbb{E}[H]$	$m(d)$	$\mathbb{E}[H]$	$m(d)$	$\mathbb{E}[H]$	$m(d)$
Q_1	0.61	0.00	0.43	0.00	0.36	0.00
Q_2	1.28	0.01	1.09	0.01	0.79	0.01
Q_3	2.02	0.16	1.67	0.14	1.12	0.02
Q_4	3.62	0.25	3.00	0.24	2.76	0.21
r	0.65*		0.62*		0.56*	

Table 3: $\mathbb{E}[H]$ and $m(d)$ refer to intrinsic uncertainty approximated by an LM and median of the difference between ECE and e-ECE respectively. r is pearson correlation coefficient, and * indicates p-value less than 0.01 ($p < 0.01$)

n	ECE			e-ECE		
	$\mathbb{E}$	σ (↓)	Δ (↓)	$\mathbb{E}$	σ (↓)	Δ (↓)
10	26.19	8.18	19.05	15.81	**3.61**	**11.39**
50	13.36	3.26	6.22	9.01	**2.58**	**4.59**
500	7.45	1.23	0.31	4.48	**0.93**	**0.06**
1000	6.97	1.36	0.17	4.48	**0.93**	**0.06**
Full	7.14			4.42		

Table 4: n denotes the number of test samples used in evaluating model calibration. $\mathbb{E}$ and σ denote the mean and standard deviation computed over 10 runs with n test samples. Δ is the absolute difference between the evaluations of n samples and that of whole test dataset, ↓ indicating the lower the better. A bold number represents the best score.

output across the corpora tested. For instance, a relative decrease of 38.1% is seen in e-ECE compared to the ECE score in `IWSLT14`, and 33.9% relative decrease in `Multi30K`. We draw similar observations from e-ECE with top-p and top-k sampling generation schemes. In the following section, we illustrate that the decrease comes from the intrinsic uncertainty that remained as error in ECE.

5.1 e-ECE Reflects Intrinsic Uncertainty

A direct way to validate the ability of e-ECE in handling intrinsic uncertainty is by analyzing the *samples that ECE and e-ECE show mismatch*; we measure the difference between ECE and e-ECE at the token-level.

$$d_t^{(i)} = |g(\hat{p}_t^{(i)}, \hat{a}_t^{(i)}) - g(\tilde{p}_t^{(i)}, \tilde{a}_t^{(i)})| \qquad (7)$$

where $g(p, a)$ is a token-level accuracy and predictive score gap, computed as $p - a$. $(\hat{p}, \hat{a})$ and $(\tilde{p}, \tilde{a})$ are the inputs to ECE and e-ECE respectively.

If the proposed metric takes intrinsic uncertainty into consideration, little intrinsic uncertainty is expected when there is a little difference between the metrics ($d^{(i)} \approx 0$). On the contrary, a high intrinsic uncertainty is expected in samples with which the metrics disagree ($d^{(i)} > 0$). We partition test predictions into 4 groups, based on the intrinsic uncertainty, which we approximate with an LM. Entropy level of an LM represents the size of valid candidates within a context, being a proper approximation for intrinsic uncertainty. We report the median difference within each group in Table 3.

We observe a marked difference in the intrinsic uncertainty level between the groups. The samples with little intrinsic uncertainty (Q_1) have no disagreement between the two metrics illustrated with the median difference equals to 0. However, as the intrinsic uncertainty increases, the difference stands out. In addition, the pearson correlation coefficient between entropy and the token-level difference between the metrics is around 0.6, a clear indication of strong linear correlation. This empirical finding supports a finding that the disagreement between ECE and e-ECE comes from the samples with high intrinsic uncertainty; thus, e-ECE reflects such uncertainty in evaluation.

5.2 e-ECE is Stable

A calibration metric should be both accurate and stable. Table 4 illustrates a comparison between ECE and e-ECE in stability. With a small number of test instances, e-ECE displays lower standard deviation, and illustrates close approximations to the score computed with whole test datasets. This indicates that e-ECE requires less number of test samples while depicting superior stability compared to ECE. We attribute stability of e-ECE to the nature of expectation. The expected accuracy, different from the existing metrics, is not discrete but continuous. This aspect of e-ECE relaxes the accuracy and confidence gap.

6 Conclusion

In this study, we explore the limitations of the existing calibration metrics, especially within the scope of natural language generation. To that end, we propose e-ECE that considers intrinsic uncertainty of a natural language in evaluating model calibration. The proposed metric is tested on the popular translation datasets, and the empirical results support the validity of the proposed metric in evaluating calibration of an LM.

References

Morris H. DeGroot and Stephen E. Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22.

Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. 2021. Local temperature scaling for probability calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6889–6899.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12848–12856.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Abhyuday Jagannatha and Hong Yu. 2020. Calibrating structured output predictors for natural language processing. *CoRR*, abs/2004.04361.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Taehee Jung, Dongyeop Kang, Hua Cheng, Lucas Mentch, and Thomas Schaaf. 2020. Posterior calibrated training on sentence classification tasks. *CoRR*, abs/2004.14500.

Ranganath Krishnan and Omesh Tickoo. 2020. Improving model calibration with accuracy versus uncertainty optimization.

Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *Neural Information Processing Systems*.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Association for the Advancement of Artificial Intelligence*.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhiliang Tian, Wei Bi, Dongkyu Lee, Lanqing Xue, Yiping Song, Xiaojiang Liu, and Nevin L. Zhang. 2020. Response-anticipated memory for on-demand knowledge integration in response generation. *CoRR*, abs/2005.06128.

Christian Tomani and Florian Buettner. 2021. Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In *Association for the Advancement of Artificial Intelligence*.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

A e-ECE Computation

A.1 From k Samples

The expected predictive score and expected accuracy can be computed with k predictions sampled from distribution $\tilde{P}_\theta$.

$$\begin{aligned}\tilde{p}_t^{(i)} &= \frac{1}{k}\sum_{j=1}^{k} P(\tilde{y}_{t,j}^{(i)}|y_{1:t-1}, x^{(i)}; \theta) \\ \tilde{a}_t^{(i)} &= \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}(y^{(i)} = \tilde{y}_{t,j}^{(i)})\end{aligned} \tag{8}$$

A.2 Whole Output Space

The expected values can be computed without sampling process as follows:

$$\begin{aligned}\tilde{p}_t^{(i)} &= \sum_i \tilde{P}(y_i|y_{1:t-1}^{(i)}, x^{(i)}; \theta) \\ &\quad \times P(y_i|y_{1:t-1}^{(i)}, x^{(i)}; \theta) \\ \tilde{a}_t^{(i)} &= \tilde{P}(y_t^{(i)}|y_{1:t-1}^{(i)}, x^{(i)}; \theta)\end{aligned} \tag{9}$$

B Connection to ECE

Proposition 1. *With a small temperature value $\tau \approx 0$, e-ECE converges to ECE metric.*

$$\lim_{\tau \to 0} e\text{-ECE} = \text{ECE} \tag{10}$$

For the ease of understanding, we rewrite the Equation 5.

$$\begin{aligned}\tilde{p}_t^{(i)} &= \mathbb{E}_{\tilde{y} \sim \tilde{P}_\theta}[P(\tilde{y}|y_{1:t-1}^{(i)}, x^{(i)}; \theta)] \\ \tilde{a}_t^{(i)} &= \mathbb{E}_{\tilde{y} \sim \tilde{P}_\theta}[\mathbb{1}(y^{(i)} = \tilde{y}^{(i)})], a_t^{(i)} \in [0, 1]\end{aligned} \tag{11}$$

Given a low temperature, the expected confidence $\tilde{p}_t$ converges to $\hat{p}_t$ as the sampled prediction $\tilde{y}$ will always be identical to the `argmax` prediction $\hat{y}$.

$$\begin{aligned}\tilde{y}_t^{(i)} &= \hat{y}_t^{(i)} \\ \tilde{p}_t^{(i)} &= \hat{p}_t^{(i)}\end{aligned} \tag{12}$$

In this regard, the expected accuracy $\tilde{a}_t$ becomes identical to indicator function where the expected accuracy is now either 0 or 1, as in ECE. Therefore, e-ECE converges to ECE with a proper temperature control.

C Dataset

Our work proposes a metric that considers intrinsic uncertainty of a language. In this regard, we validate the proposed metrics on translation datasets which are known to contain intrinsic uncertainty. The details are shown in Table 5

Dataset	$\#D_{train}$	$\#D_{val}$	$\#D_{test}$	#Vocab
`WMT14`	4,500,966	3,000	8,171	(32768, 32768)
`IWSLT14`	16,239	7,283	6,750	(8848, 6632)
`Multi30K`	28,332	1,014	1,000	(7072, 5184)

Table 5: Description on the datasets tested in this work. $\#D_{subset}$ denotes the number of paired sentences in a subset. #Vocab is a tuple with source and target dictionary size.

D Experiment Design

All of the experiments have been conducted with A100 GPUs. We follow the hyperparameters and model structures specified on `fairseq` (Ott et al., 2019)[5].

E Approximating Intrinsic Uncertainty

In our work, intrinsic uncertainty of a language is approximated with a conditional LM. An LM is able to select a subset of vocabulary, which the tokens in the subset are valid choice in a context. Therefore, the entropy level of an LM can be an approximation of intrinsic uncertainty in a context.

$$\begin{aligned}H(P) = -\sum_l & P(y_l|x, y_{1:t-1}; \theta_{lm}) \\ & \times \log P(y_l|x, y_{1:t-1}; \theta_{lm})\end{aligned} \tag{13}$$

where l denotes class (token) index and t denotes time step. A low entropy, a probability mass concentrated to a small subset of tokens, indicates small intrinsic uncertainty, while a high entropy level is an indication of high intrinsic uncertainty. The conditional language model follows the identical model configuration specified in Appendix D.

[5] `https://github.com/facebookresearch/fairseq/blob/main/examples/translation/README.md`