Confident in a Confidence Score: Investigating the Sensitivity of the Calibration of Confidence Scores to Supervised Fine-Tuning in Generation Tasks

Anonymous ACL submission

Abstract

Uncertainty quantification (UQ) techniques measure confidence in language model outputs to support critical applications like hallucination detection and selective prediction. While prior work has developed various confidence metrics and demonstrated their calibration for classification tasks or using verbalized confidence, the robustness of probability-based and self-consistency-based UQ metrics for natural language generation (NLG) remains underexplored—particularly under model adaptation. Since practitioners routinely apply supervised fine-tuning (SFT) to adapt models to new tasks, a key question arises: do confidence metrics maintain their calibration when models are fine-tuned? We investigate this question across NLG tasks including translation, question answering, and mathematical reasoning. We find that calibration degrades post-SFT in over half of the 216 configurations tested, with confidence scores shifting due to factors beyond output quality, such as proximity to the training distribution. Through a downstream task evaluation, we show that this miscalibration substantially reduces the practical utility of confidence scores for identifying correct answers. Our findings reveal that existing confidence metrics for NLG cannot be reliably deployed off-the-shelf after fine-tuning, highlighting the need for calibration-robust UQ methods under model adaptation.

1 Introduction

Uncertainty quantification (UQ) techniques measure the confidence of language models in their predictions and have become critical for deploying models safely in high-stakes applications. These confidence scores enable systems to detect hallucinations (Wang et al., 2024; Manakul et al., 2023), select higher-quality answers (Wang et al., 2023), self-evaluate outputs (Ren et al., 2023), and flag uncertain predictions for human review (Xiao et al., 2020; Malinin and Gales, 2021; Liu et al., 2020; Kamath et al., 2020).

Prior work has developed UQ metrics for classification tasks (Tian et al., 2023; Huang et al., 2025) and verbalized confidence (Rathi et al., 2025; Leng et al., 2025). However, these approaches have limitations: classification methods require binary correctness judgments unsuitable for open-ended generation, and verbalized confidence often require large instruction-tuned models (Tian et al., 2023). Critically, prior work studied calibration in static settings, leaving open whether confidence metrics remain reliable under model adaptation, especially since various UQ metrics and packages are designed to allow any white-box model to be plugged in (Shelmanov et al., 2025).

Understanding calibration stability under model adaptation is essential because practitioners routinely fine-tune pre-trained models for specialized tasks. If confidence metrics fail to maintain calibration after supervised fine-tuning (SFT), their practical utility is severely limited. This gap is particularly concerning for NLG tasks—such as translation, question answering, and summarization—where *probability-based* and *self-consistency-based* metrics are commonly used (Malinin and Gales, 2021; Xiao et al., 2020). These metrics rely only on token probabilities or output samples, making them broadly applicable to any white-box model. However, their robustness to model adaptation has not been systematically studied. We ask: **Can we use UQ metrics with *any* model, including fine-tuned variants, and expect to get reliably calibrated uncertainty estimates for NLG tasks?**

We investigate how supervised fine-tuning affects the calibration of probability-based and self-consistency-based confidence scores across diverse NLG tasks: translation, question answering, and mathematical reasoning. We systematically vary fine-tuning parameters—including the number of

training samples and epochs—to understand calibration sensitivity under realistic adaptation scenarios. Across 216 task-model-UQ metric configurations, we find that calibration degrades in 112 cases post-SFT, with 43 statistically significant degradations. This demonstrates that confidence metrics cannot be assumed to maintain their calibration properties after fine-tuning.

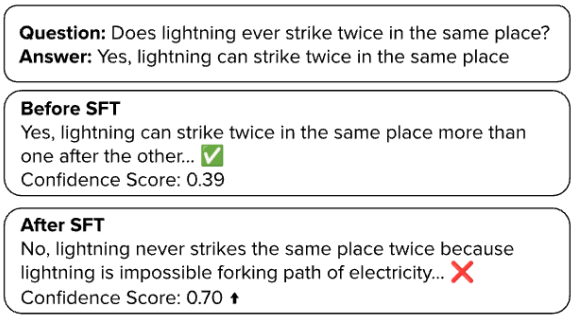


Figure 1: Before SFT, a model had a relatively low confidence score in a correct answer, whereas post-SFT, it had a much higher confidence in an incorrect answer, demonstrating a case of *relative overconfidence*

Because calibration relies on sample-level confidence and evaluation scores being aligned, we study sample-level changes in these scores to understand how miscalibration occurs post-SFT and to motivate interventions. We find that SFT is prone to making probability-based scores overconfident (see Figure 1) and self-consistency-based scores underconfident. Hence, the way in which miscalibration occurs is not always the same, and different interventions may be needed for each metric. We find that confidence scores are affected by factors beyond output quality, like the output's similarity to the training distribution, which must be accounted for to isolate the relationship with quality.

To illustrate the consequences of miscalibration, we use confidence scores to predict whether the output is correct on a QA task, and find that these confidence scores' ability to identify correct answers decreases in 35/72 cases post-SFT. Thus, the usefulness of these metrics can suffer if miscalibration is not addressed.

To summarize, we ask: **RQ1: How sensitive is the calibration of confidence metrics to different SFT parameters? RQ2: What are the underlying dynamics of confidence scores that negatively impact calibration with quality?** We find that the calibration of various confidence metrics change drastically post-SFT, with different types of miscalibration occurring, stemming from factors beyond the quality of the output. This can have negative impacts on downstream applications of UQ metrics. Hence, confidence metrics cannot be used off-the-shelf without testing. We argue for the need for developing metrics which are more robust to model finetuning, or at least, better reporting of the consistency of such metrics in future studies.

2 Related Work

UQ Metrics Uncertainty quantification methods include fine-tuning additional models to estimate confidence (Yaldiz et al., 2024; Kamath et al., 2020; Malinin et al., 2019; Fathullah et al., 2023), detecting out-of-distribution samples (Liu et al., 2020; Vazhentsev et al., 2023), or having models verbalize confidence scores (Lin et al., 2022a; Tian et al., 2023; Kapoor et al., 2024; Han et al., 2024), which relies on larger language models which are capable of doing so.

We focus on methods applicable to any white-box model without additional training: **Probability-based methods** use token probabilities (Murray and Chiang, 2018; Zablotskaia et al., 2023; Zhao et al., 2020; Kumar and Sarawagi, 2019; Malinin and Gales, 2021; Flores et al., 2025), sometimes augmented with output similarity (Lin et al., 2023; Kuhn et al., 2023; Nikitin et al., 2024). **Self-consistency methods** sample multiple outputs (via dropout, beam search) and measure prediction consistency (Xiao et al., 2020; Schmidt et al., 2022; Lakshminarayanan et al., 2017). Recent work combines both approaches (Vashurin et al., 2025).

Evaluating UQ Metrics To be practically useful, these scores must be associated with the quality of the output (Daheim et al., 2025; Kotelevskii et al., 2025). We assume that quality is measured by the evaluation metric used for the task, and broadly captures the semantic correctness of the output. For tasks where the answer can be classified as right or wrong, this notion is captured by calibration, which is measured using expected calibration error (Tian et al., 2023). For continuous quality scores in generation tasks, we use the Prediction-Rejection Ratio (PRR) (Vashurin et al., 2025; Malinin and Gales, 2021; Malinin et al., 2019), which measures how well confidence scores reject low-quality outputs.

3 Methods and Experiments

In this section, we define confidence scores and our framework for evaluating their calibration with quality, and provide experimental details of the

analyses in the following sections. In Section 4, we study how sensitive this calibration is to fine-tuning (RQ1), and vary the number of epochs and samples used. In Section 5, we study the reasons for the observed sensitivity by analyzing the dynamics of how confidence scores change at the sample level in order to describe how miscalibration occurs and potentially propose interventions (RQ2).

3.1 Confidence Scores

Definition 3.1. *A **confidence score** is a number describing a model's assessment of its output's quality, computed using input x and model output $\hat{y}$.*

Definition 3.2. *We evaluate a confidence score by its **calibration** using the Prediction-Rejection Ratio (PRR) (Vashurin et al., 2025; Malinin and Gales, 2021; Malinin et al., 2019), which captures the notion that higher confidence should be associated with higher quality. Quality is measured by evaluating the output $\hat{Y}$, and may use reference Y or input X. Moreover, X and Y refer to input and outputs as a variable, whereas x and y refer to an individual input/output.*

$$\text{Calibration} = PRR(\text{Confidence}(X, \hat{Y}), \text{Quality}(\hat{Y}))$$

Definition 3.3. *We say a confidence score is **miscalibrated** if task-level calibration is low.*

We test probability and consistency based UQ metrics; we focus on these two groups of metrics as they can be used with any white-box model.

Probability-Based (1) average token log probability **(Avg Tok Prob)** (Murray and Chiang, 2018; Zablotskaia et al., 2023), (2) average token entropy **(Avg Tok Ent)** (Zhao et al., 2020; Perlitz et al., 2023), (3) avg token entropy across dropout samples (**DO Ent**, Eq 1) (Malinin and Gales, 2021), (4) the weighted average of the top-K sequences' average token log probabilities from beam search (BS) **(BS Imp Wt)** (Eq 2) (Malinin and Gales, 2021), (5) the ratio between the sequence probability of the top and k-th beam (BS) **(BS Ratios)** (Flores et al., 2025), and (6) the sum of the top k beams' joint sequence probabilities **(BS Sums)**.

Self-Consistency-Based We compute the difference between model outputs sampled using dropout (DO), using (1) BLEU (**DO BLEU Var**, Eq 4) (Xiao et al., 2020) (Schmidt et al., 2022), (2) KL divergence (**DO KL Div**, Eq 5) (Lakshminarayanan et al., 2017), and (3) METEOR (**DO Meteor Var**, Eq 3). We also test (4) CoCoA, which is the product of a probability-based and self-consistency based metric **(CoCoA MSP/MTE/PPL)**, where the probability-based metrics used are mean token probability (MSP), mean token entropy (MTE), and token perplexity (PPL) (Vashurin et al., 2025).

Note these metrics do not have parameters modified by SFT. Additional techniques or modules can be used (Yaldiz et al., 2024) to improve the calibration of these metrics given the appropriate data, but we do not assume access to such a dataset.

3.2 Experimental Details

Tasks We fine-tune models for **Translation** using the English-Afrikaans (Eng-Afr) split of NLLB (NLLB, 2022), with FLORES as the test dataset (Goyal et al., 2021), **Question Answering (QA)** using SQUAD (Rajpurkar et al., 2016), and **Math** using GSM8K (Cobbe et al., 2021). To evaluate quality, we compute ChrF+ (Popović, 2015) for translation, F1 for QA, and exact match for math.

Models We generate confidence scores with BART Base (Lewis et al., 2019), Flan-T5 Base (Chung et al., 2022), Llama 3.1-8B (Grattafiori et al., 2024), Gemma-2-2B and 9B (IT) (Team et al., 2024), and QWen 2.5-14B (Team, 2025).

4 RQ1: How sensitive is the calibration of confidence metrics to SFT?

4.1 Experimental Design

We study the sensitivity of task-level calibration of confidence metrics to parameters in the SFT pipeline, by measuring calibration pre/post-SFT and when varying the epochs and samples used.

We measure the effect of SFT on task-level calibration, by generating the outputs and confidence scores pre and post-SFT, and computing the change in PRR. In addition to measuring pre/post changes, we measure epoch 1/post changes, because the outputs pre-SFT are language modeling outputs and not for the task at hand, hence confidence scores may not be meaningful. We repeat the experiments by varying the number of epochs and samples.

4.2 Results

The calibration of confidence metrics varies widely before and after SFT We start by measuring the calibration of pre-trained models without SFT or after one epoch, and after performing SFT with early stopping; these represent the confidence scores obtained if we chose to fine-tune or not.

	BART-Base			Flan-T5-Base			Llama 3.1-8B		
	Eng-Afr	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**
Avg Tok Prob	0.209 (0.01)	0.626 (0.006)	-0.467 (0.013)	0.335 (0.034)	0.128 (0.015)	-0.473 (0.008)	0.092 (0.107)	0.081 (0.072)	-0.257 (0.122)
Avg Tok Ent	-0.208 (0.011)	-0.638 (0.01)	0.467 (0.011)	-0.329 (0.034)	-0.135 (0.018)	0.473 (0.007)	-0.09 (0.099)	-0.077 (0.063)	0.259 (0.123)
DO Ent	-0.173 (0.015)	-0.521 (0.042)	0.475 (0.01)	-0.403 (0.04)	-0.133 (0.011)	0.471 (0.011)	-0.081 (0.103)	-0.138 (0.049)	0.345 (0.097)
BS Imp Wt	0.233 (0.007)	0.472 (0.014)	-0.469 (0.006)	0.352 (0.043)	0.091 (0.004)	-0.483 (0.003)	0.089 (0.082)	0.067 (0.053)	-0.266 (0.117)
BS Ratios	0.194 (0.014)	0.616 (0.013)	-0.468 (0.007)	0.282 (0.031)	0.003 (0.004)	-0.474 (0.011)	0.076 (0.101)	0.26 (0.076)	-0.289 (0.098)
BS Sums	0.233 (0.007)	0.436 (0.021)	-0.469 (0.006)	0.352 (0.043)	0.095 (0.004)	-0.483 (0.003)	0.088 (0.083)	0.066 (0.053)	-0.266 (0.117)
DO Bleu Var	-0.224 (0.019)	-0.551 (0.031)	0.47 (0.007)	-0.332 (0.027)	-0.1 (0.011)	0.473 (0.005)	-0.058 (0.074)	-0.145 (0.071)	0.353 (0.065)
DO KL Div	-0.181 (0.008)	-0.612 (0.004)	0.469 (0.006)	-0.387 (0.059)	0.036 (0.047)	0.484 (0.004)	-0.047 (0.075)	-0.198 (0.096)	0.366 (0.058)
DO Meteor Var	0.129 (0.034)	0.442 (0.037)	-0.473 (0.007)	0.317 (0.015)	0.196 (0.015)	-0.479 (0.005)	0.078 (0.099)	0.199 (0.123)	-0.376 (0.07)
CoCoA MSP	-0.159 (0.024)	-0.333 (0.02)	0.471 (0.009)	-0.332 (0.014)	-0.093 (0.009)	0.478 (0.002)	-0.043 (0.09)	-0.172 (0.124)	0.436 (0.018)
CoCoA MTE	-0.197 (0.011)	-0.659 (0.011)	0.472 (0.006)	-0.176 (0.033)	-0.171 (0.041)	0.477 (0.007)	-0.067 (0.095)	-0.086 (0.054)	0.252 (0.13)
CoCoA PPL	-0.2 (0.011)	-0.653 (0.011)	0.472 (0.007)	-0.197 (0.033)	-0.163 (0.04)	0.476 (0.008)	-0.071 (0.097)	-0.089 (0.061)	0.249 (0.133)

	Gemma 2-2B			Gemma 2-9B			QWen 2.5-14B		
	Eng-Afr	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**
Avg Tok Prob	0.117 (0.121)	0.574 (0.125)	-0.292 (0.026)	-0.098 (0.065)	0.105 (0.007)	0.129 (0.051)	0.155 (0.057)	0.126 (0.012)	0.22 (0.087)
Avg Tok Ent	-0.13 (0.123)	-0.579 (0.12)	0.3 (0.02)	0.073 (0.069)	-0.112 (0.012)	-0.126 (0.062)	-0.163 (0.057)	-0.125 (0.013)	-0.224 (0.077)
DO Ent	-0.122 (0.16)	-0.515 (0.087)	0.341 (0.063)	0.12 (0.096)	-0.096 (0.006)	0.079 (0.069)	-0.204 (0.101)	-0.216 (0.022)	-0.21 (0.043)
BS Imp Wt	0.136 (0.12)	0.496 (0.082)	-0.272 (0.051)	-0.114 (0.048)	0.037 (0.008)	0.142 (0.072)	0.086 (0.058)	0.05 (0.037)	0.222 (0.078)
BS Ratios	0.07 (0.137)	0.649 (0.108)	-0.246 (0.058)	-0.181 (0.114)	0.087 (0.007)	0.014 (0.037)	0.077 (0.083)	0.123 (0.014)	0.232 (0.015)
BS Sums	0.136 (0.12)	0.485 (0.079)	-0.274 (0.051)	-0.116 (0.049)	0.036 (0.008)	0.14 (0.073)	0.087 (0.058)	0.047 (0.038)	0.222 (0.078)
DO Bleu Var	-0.082 (0.128)	-0.564 (0.101)	0.37 (0.032)	0.133 (0.073)	-0.142 (0.003)	0.052 (0.069)	-0.213 (0.04)	-0.221 (0.013)	-0.104 (0.014)
DO KL Div	-0.073 (0.137)	-0.583 (0.066)	0.334 (0.088)	0.144 (0.089)	-0.145 (0.011)	0.07 (0.042)	-0.2 (0.095)	-0.179 (0.01)	-0.168 (0.028)
DO Meteor Var	0.07 (0.152)	0.579 (0.11)	-0.324 (0.054)	-0.132 (0.075)	0.149 (0.019)	-0.048 (0.078)	0.134 (0.086)	0.132 (0.038)	0.187 (0.028)
CoCoA MSP	-0.059 (0.127)	-0.57 (0.134)	0.398 (0.053)	0.164 (0.115)	-0.159 (0.023)	0.244 (0.067)	-0.086 (0.077)	-0.287 (0.025)	-0.064 (0.036)
CoCoA MTE	-0.116 (0.13)	-0.582 (0.105)	0.312 (0.043)	0.089 (0.087)	-0.104 (0.017)	-0.089 (0.056)	-0.291 (0.05)	-0.118 (0.025)	-0.207 (0.054)
CoCoA PPL	-0.115 (0.13)	-0.576 (0.113)	0.308 (0.042)	0.101 (0.088)	-0.094 (0.016)	-0.094 (0.04)	-0.287 (0.054)	-0.122 (0.017)	-0.207 (0.063)

Table 1: PRR gets significantly worse post-SFT in 43/216 cases, and significantly better in 47/216 cases (one-way ANOVA, with Holm correction, $\alpha = 0.05$); Mean delta in PRR pre/post SFT (3 seeds, st. dev in parentheses)

As shown in Table 1, there can be substantial differences in calibration of confidence scores before and after SFT. Out of 216 scores-dataset-model configurations (3 datasets $\times$ 6 models $\times$ 12 confidence metrics), miscalibration occurs post-SFT in 112 cases, with 43 cases being statistically significant (one-way ANOVA, with Holm correction, $\alpha = 0.05$). Conversely, it improves in 104 cases, with 47 being statistically significant.

Running the comparisons between the first and final epoch, we still see that there are more cases in which SFT worsens calibration (110/216 cases), with two statistically significant (Table 6). The results emphasize the need to to check these metrics before deploying them for one's use case.

Varying the number of epochs considerably affects task-level calibration While the previous result studies calibration pre/post fine-tuning, we run an experiment to understand how calibration changes epoch-by-epoch. We fine-tune models for 10 epochs, measuring the calibration after each epoch, and find that PRR can drop from one epoch to the next by at most 0.290 points (See Table 7).

Given that calibration can change considerably after just one epoch, users should take this into account when deciding how many training epochs to use. When early stopping is employed, parameters like patience and tolerance must be tuned to yield the desired calibration. Moreover, the confidence score with the best calibration may differ based on how many epochs are used. For example in Figure 2, Beam Score Ratios (green line) performs well in the first epochs, but with SFT, Beam Score Sums (red line) showed better calibration. Hence, the choice of confidence score and number of training epochs have to be considered together.

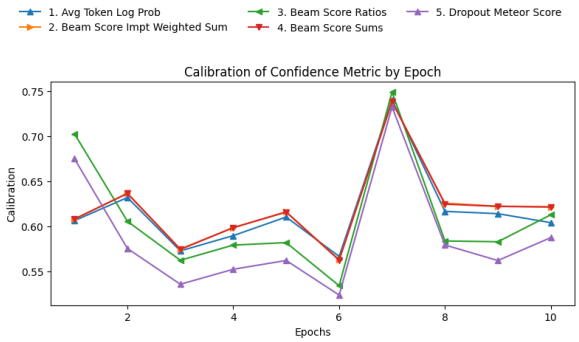


Figure 2: The best confidence metric changes by epoch; PRR values of selected metrics (QWen-14B, Eng-Afr)

Similarly, the number of samples used considerably affects calibration We vary the number of samples used ($n = 100, 500, 1000, 2000$). Fixing the number of epochs, we measure the calibration of the models across three seeds. Overall, there are significant differences in the calibration of confidence metrics when using different numbers of samples (See Fig 3, and Fig 6 for full comparison). Across the 216 confidence metric-model-task configurations, we observe significant differences by

number of samples in 31 configurations (one-way ANOVA with Holm correction, $\alpha = 0.05$). Practically, this means that the number of samples is another major consideration when using confidence metrics with fine-tuned models.

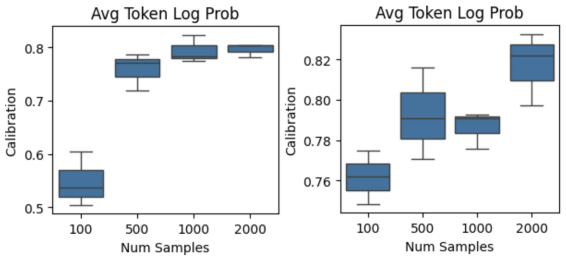


Figure 3: The calibration of confidence metrics differs significantly depending on the number of fine-tuning samples used, both before (left) and after (right) controlling for the number of fine-tuning steps used; Plots shown using average token entropy as the confidence metric, used on SQUAD for BART-Base

The previous results are affected by the fact that using more samples means using more fine-tuning steps, given a fixed number of epochs. To isolate the impact of varying the number of samples, we fix the number of training steps to 2,500, and vary the number of samples. Figure 6b shows there are still differences in calibration, with 13/216 configurations having significant differences between different numbers of samples (one-way ANOVA with Holm correction, $\alpha = 0.05$). In summary, the calibration is sensitive to SFT and to the number of epochs and samples used.

5 RQ2: What are the underlying dynamics of confidence scores that negatively impact calibration?

In the previous section, we found that the task-level calibration of confidence metrics can change significantly post-SFT, resulting in miscalibration. We now study the dynamics of how confidence scores change at the sample level, to identify causes of miscalibration and motivate interventions.

5.1 Experimental Design

We first study how confidence scores change pre/post SFT, analyzing whether they directionally change in concordance with changes in quality. We measure the change in evaluation metric and model's confidence score for each sample in the test set, which we visualize using quadrants (See Fig 4). As in Section 4, we compute the changes using the model after one epoch of SFT rather than the pre-trained model.

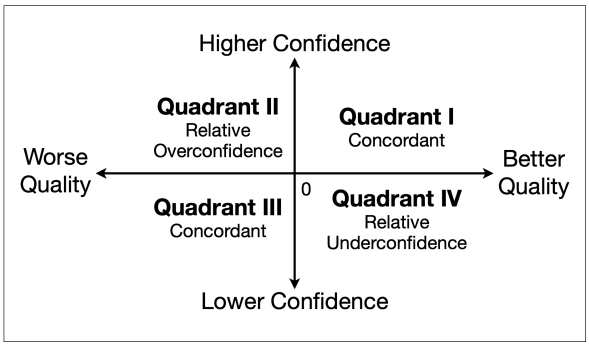


Figure 4: We classify SFT's effect on the calibration of samples using four quadrants

Definition 5.1. *For one sample, a change is* ***concordant*** *if the sample's output quality and confidence both increased (I), or both decreased (III).*

Definition 5.2. *For one sample, a change is* ***relatively overconfident*** *if its confidence increased, but its output quality decreased (II); and* ***relatively underconfident*** *if its confidence decreased, but its output quality increased (IV).*

We then compute the proportion of samples whose changes are concordant, relatively underconfident, or relatively overconfident. We use this framework to analyze how confidence scores change with a probability-based (average log token probs (Murray and Chiang, 2018; Zablotskaia et al., 2023)) and a consistency-based metric (dropout variance in BLEU score (Xiao et al., 2020)).

After studying the dynamics of confidence scores pre/post SFT, we study how this leads to miscalibration. To this end, we analyze changes in the relative ranks of samples' quality and confidence scores across epochs. We first define relative calibration and miscalibration at the sample level.

Definition 5.3. *For two samples (A, B), a confidence score is* ***relatively calibrated*** *at epoch t if*

$$q_{A,t} < q_{B,t} \iff c_{A,t} < c_{B,t}$$

where $q_{A,t}$ is the quality score of the model's output for A, evaluated w.r.t. a reference, $c_{A,t}$ is the model's confidence in its output for A, as measured by a confidence metric.

Definition 5.4. *For two samples (A, B), a confidence score is* ***relatively miscalibrated*** *if*

$$q_{A,t} < q_{B,t} \text{ and } c_{A,t} > c_{B,t}$$

We study how changes in confidence scores and quality result in miscalibration. Formally, fine-tuning causes miscalibration for pair (A, B) if the pair was relatively calibrated at epoch t, and relatively miscalibrated at epoch $t+1$.

There are two cases wherein samples can become relatively miscalibrated. If at epoch t, a pair of samples (A, B) is relatively calibrated, such that $q_{A,t} < q_{B,t}$ and $c_{A,t} < c_{B,t}$. Then, relative miscalibration can happen if either:

- **Case 1**: The relative ranking of quality scores stays the same, but the relative confidence scores flip

$$q_{A,t+1} < q_{B,t+1} \text{ and } c_{A,t+1} > c_{B,t+1}$$

- **Case 2:** The relative ranking of quality scores change, but the confidence scores' rankings stay the same

$$q_{A,t+1} > q_{B,t+1} \text{ and } c_{A,t+1} < c_{B,t+1}$$

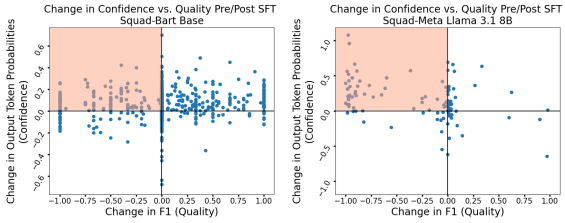
We study which case most frequently occurs, to better understand the exact issues that lead to miscalibration, which future UQ metrics can address.

5.2 Results

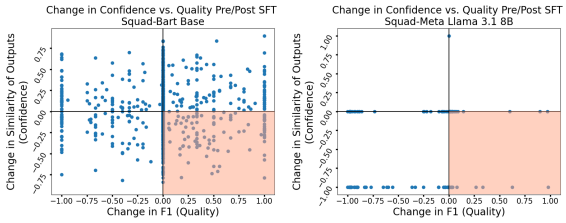
Probability-based methods are prone to relative overconfidence We compute the change in average log probs between the first epoch and post-SFT (See Table 2). We observe that although a bulk of the changes are concordant, there is a tendency for SFT to make BART, Llama-3.1, Gemma-2B, Gemma-9B, and QWen-14B relatively overconfident in samples. We visualize examples of relative overconfidence using the quadrant plots in Figure 5a, which show that more samples fall in Quadrant II (Overconfident) compared to Quadrant IV (Underconfident).

To understand the changes in confidence score better, we plot each sample's average log probabilities by epoch. We observe that the distribution of the average log token probabilities shifts upwards with more epochs (See Figure 7).

Building on the pre/post-SFT analysis in Table 2, we classify epoch-level changes from epoch t to epoch $t+1$ for $1 \leq t \leq 10$ in confidence score by epoch in Table 3. We find that (1) a large percentage of confidence scores are indeed increasing from epoch to epoch, but that among them, (2) quality is often decreasing, which corresponds to the relative overconfidence we observed in the pre/post SFT



(a) Avg Log Probs, BART (Left) and Llama 3.1 (Right)



(b) Dropout BLEU Var, BART (Left) and Llama 3.1 (Right)

Figure 5: Plots reveal that average log probabilities are more prone to relative overconfidence, while dropout BLEU variance is more prone to relative underconfidence after SFT; plotting change in F1 between epoch one and ten for SQUAD vs confidence score

experiments. Our takeaway is that at the epoch-by-epoch level, fine-tuning pushes the confidence score (i.e. average log token probs) upward across *most* outputs, regardless the change in the output's quality, which may explain why the calibration between confidence and quality becomes worse.

Consistency-based methods are prone to underconfidence We repeat the experiment using dropout variance with BLEU (Table 9). Unlike the probability-based methods, the changes are either concordant or relatively underconfident. We show examples in Figure 5b, where we observe that more samples fall in Quadrant IV (Underconfident) than Quadrant II (Overconfident). When plotting the dropout BLEU variance values by epoch, we do not observe clear patterns (Fig 8).

In summary, we find that after performing SFT, probability-based methods are more prone to relative overconfidence, whereas consistency-based methods are prone to relative underconfidence. Users may then choose which type of confidence metric to use, depending on which type of miscalibration is more acceptable.

Case 1 relative miscalibration is more prevalent After studying the dynamics of confidence metrics pre/post SFT, we study how these changes result in miscalibration, to identify where interventions can be made. Using average log probs to mea-

	Dataset	Cncd	R. Over	R. Under
BART	Eng-Afr	0.94	0.06	0.00
	SQUAD	0.85	0.11	0.03
	GSM8K	0.98	0.02	0.00
Flan-T5	Eng-Afr	0.17	0.02	0.80
	SQUAD	0.82	0.02	0.16
	GSM8K	0.99	0.00	0.01
Llama	Eng-Afr	0.55	0.16	0.29
	SQUAD	0.37	0.51	0.12
	GSM8K	0.90	0.07	0.03
Gemma-2B	Eng-Afr	0.54	0.26	0.20
	SQUAD	0.63	0.27	0.10
	GSM8K	0.87	0.13	0.00
Gemma-9B	Eng-Afr	0.69	0.28	0.03
	SQUAD	0.78	0.17	0.05
	GSM8K	0.92	0.08	0.00
QWen-14B	Eng-Afr	0.56	0.30	0.14
	SQUAD	0.78	0.13	0.09
	GSM8K	0.89	0.04	0.07

Table 2: Proportion of changes in test set samples that are concordant, relatively over/underconfident after SFT, using average token log probs as the measure of confidence; Reporting average proportions over 3 seeds

Confidence increase?		No	Yes	
Quality improve?			No	Yes
BART	Eng-Afr	32.0	26.3	41.7
	SQUAD	45.0	45.5	9.5
	GSM8K	37.1	61.4	1.6
Flan-T5	Eng-Afr	45.1	27.5	27.4
	SQUAD	49.0	47.2	3.9
	GSM8K	49.2	49.9	1.0
Llama-3.1 8B	Eng-Afr	46.6	23.9	29.5
	SQUAD	46.7	44.9	8.3
	GSM8K	46.8	35.4	17.8
Gemma-2B	Eng-Afr	48.1	22.0	29.9
	SQUAD	43.9	37.3	18.8
	GSM8K	41.6	50.4	8.0
Gemma-9B	Eng-Afr	44.0	18.9	37.1
	SQUAD	48.8	35.8	15.4
	GSM8K	44.7	39.4	15.9
QWen-14B	Eng-Afr	45.2	21.0	33.8
	SQUAD	48.2	35.7	16.1
	GSM8K	47.5	40.4	12.1

Table 3: Percentage of changes in test set samples between epochs t and $t+1$ from $1 \leq t \leq 9$ classified based on directionality of changes in confidence and quality; Reporting average percentages over 3 seeds

sure confidence, we find that case 1 miscalibration happens more than case 2 in 10/18 dataset-model configurations (See Table 4). When using variance in BLEU scores, case 1 is more prevalent in 6/18 dataset-model configurations (See Table 10).

Case 1 miscalibration happens when the confidence score changes due to factors other than the output's quality Probing more deeply, we find that in a considerable proportion of case 1 miscalibrated pairs, the relative confidence scores incorrectly change, *without* the quality of the actual prediction changing. In particular, we compute the proportion of case 1 pairs, where either (1) the model increased its confidence in the worse sample, or (2) the model decreased its confidence in the better sample, without those samples' quality scores changing. We find that, when using BART or Flan-T5 for SQUAD and GSM8K, these cases comprise sometimes up to 99.9% of miscalibrated cases (See Table 11).

Our takeaway here is that in many cases, miscalibration happens when the model's confidence scores change due to factors *other* than the quality of the output itself. For example, higher model log probabilities in the output can indicate that the sample is closer to the training distribution of the model (Liu et al., 2020; Vazhentsev et al., 2023), which is not necessarily indicative of the correctness of the output. We explore this in our experiments, by measuring the calibration between samples' confidence scores and their similarity to the training data. To compute similarity, we use the maximum cosine similarity between a sample and the samples in the training set. Indeed, for some datasets like SQUAD and GSM8K, there are small, but statistically significant positive calibrations between the model's confidence scores and the similarity of those samples to the training set (See Table 12).

This suggests that it is not enough to use output probabilities and self-consistency methods to measure confidence, as these are affected by underlying factors beyond quality. Therefore, future work can propose methods that account for these interactions, hopefully to produce confidence scores that remain robust to SFT. Moreover, this shows how sensitive these metrics can be to various factors, highlighting how these confidence scores cannot simply be taken off-the-shelf without testing.

6 Case Study: How does miscalibration affect downstream tasks?

To illustrate the practical consequences of miscalibration, we evaluate confidence scores on a downstream application: using them to identify correct answers in a question-answering task. This represents a common real-world use case where confidence scores guide decision-making—for example, determining whether to present an answer to a user or flag it for review. Similar applications include hallucination detection and selective prediction, where UQ scores are routinely deployed (Manakul et al., 2023; Wang et al., 2023).

	Rel Qual.	Same		Flips	
	Rel Conf.	Same	Flips	Flips	Same
BART	Eng-Afr	0.70	0.12	0.11	0.06
	SQUAD	0.66	0.14	0.13	0.06
	GSM8K	0.79	0.16	0.03	0.01
Flan-T5	Eng-Afr	0.65	0.11	0.17	0.07
	SQUAD	0.74	0.18	0.06	0.02
	GSM8K	0.78	0.19	0.02	0.01
Llama-3.1 8B	Eng-Afr	0.39	0.12	0.38	0.11
	SQUAD	0.40	0.10	0.36	0.13
	GSM8K	0.66	0.19	0.11	0.05
Gemma-2B	Eng-Afr	0.38	0.11	0.39	0.12
	SQUAD	0.42	0.12	0.33	0.12
	GSM8K	0.61	0.17	0.15	0.06
Gemma-9B	Eng-Afr	0.47	0.00	0.27	0.26
	SQUAD	0.54	0.08	0.24	0.16
	GSM8K	0.69	0.02	0.19	0.16
QWen-14B	Eng-Afr	0.42	0.10	0.33	0.15
	SQUAD	0.57	0.06	0.20	0.15
	GSM8K	0.63	0.07	0.19	0.10

Table 4: **Case 1** miscalibration happens more frequently than **Case 2**; Table shows proportion of samples classified by change in relative quality and confidence between ep. t and $t+1$; Results averaged across 10 epochs and 3 seeds, using average log probs as confidence scores

	BART		Flan-T5		Llama-3.1 8B	
	1 Ep	Full	Pre	Full	Pre	Full
Tok Prob	0.563	0.639 ↑	0.545	0.610 ↑	0.532	0.465 ↓
Tok Ent	0.454	0.330 ↓	0.492	0.368 ↓	0.458	0.512 ↑
DO Ent	0.378	0.463 ↑	0.068	0.359 ↑	0.355	0.453 ↑
BS ImpWt	0.314	0.437 ↑	0.511	0.349 ↓	0.506	0.492 ↓
BS Rat	0.542	0.560 ↑	0.609	0.713 ↑	0.593	0.409 ↓
BS Sums	0.670	0.542 ↓	0.486	0.641 ↑	0.494	0.508 ↑
DO BLEU	0.411	0.381 ↓	0.348	0.533 ↑	0.554	0.516 ↓
DO KL	0.583	0.489 ↓	0.185	0.461 ↑	0.589	0.470 ↓
DO METr	0.580	0.621 ↑	0.920	0.512 ↓	0.465	0.511 ↑
CCo MSP	0.446	0.226 ↓	0.606	0.453 ↓	0.329	0.492 ↑
CCo MTE	0.429	0.203 ↓	0.671	0.492 ↓	0.335	0.472 ↑
CCo PPL	0.403	0.207 ↓	0.606	0.495 ↓	0.371	0.484 ↑
	Gemma 2-2B		**Gemma 2-9B**		**QWen 2.5-14B**	
	1 Ep	Full	Pre	Full	Pre	Full
Tok Prob	0.455	0.530 ↑	0.399	0.517 ↑	0.687	0.553 ↓
Tok Ent	0.516	0.449 ↓	0.593	0.470 ↓	0.317	0.461 ↑
DO Ent	0.411	0.401 ↓	0.456	0.456	0.332	0.394 ↑
BS ImpWt	0.569	0.469 ↓	0.634	0.522 ↓	0.312	0.465 ↑
BS Rat	0.618	0.507 ↓	0.443	0.482 ↑	0.603	0.533 ↓
BS Sums	0.418	0.526 ↑	0.366	0.480 ↑	0.688	0.526 ↓
DO BLEU	0.478	0.470 ↓	0.518	0.500 ↓	0.500	0.517 ↑
DO KL	0.510	0.556 ↑	0.587	0.568 ↓	0.608	0.512 ↓
DO METr	0.567	0.440 ↓	0.509	0.500 ↓	0.529	0.452 ↓
CCo MSP	0.492	0.544 ↑	0.721	0.448 ↓	0.334	0.485 ↑
CCo MTE	0.505	0.537 ↑	0.705	0.466 ↓	0.342	0.473 ↑
CCo PPL	0.522	0.549 ↑	0.708	0.473 ↓	0.337	0.478 ↑

Table 5: AUROC in detecting correct answers using various confidence metrics using the pre-trained and post-SFT model

Method We use the TruthfulQA dataset (Lin et al., 2022b) which consists of 817 question-answer pairs testing common misconceptions. We filter only to questions where the answers are entities (i.e. filtering out subjective questions and those where the answer is "I am unable to answer").

We divide the dataset into train-val-test splits, then perform SFT with early stopping. We obtain the model's answers on the test set, and get its confidence scores in those predictions. We use Claude 3.5 Sonnet to evaluate if the output matches the reference, since the answers are open-ended in ways that are hard to score with exact match or F1.

We rescale the confidence scores between 0% and 100% for better interpretability, so that users could be shown alerts like "There is an estimated X% chance the answer is right". We use the min and max values of the score from the test set[1].

Results We evaluate the AUROC in detecting the right answers. Pre/post SFT, calibration deteriorates after SFT in 35/72 cases (See Table 5). We show an example in Figure 1, where the model becomes rel. overconfident in a wrong answer post-SFT. Therefore, the usefulness or reliability of confidence scores can drop post-SFT if they are not designed so that their calibration is robust to SFT.

[1]In practice, the min/max values should be taken from the validation set, but because the questions on the test set are also in domain, we expect a similar distribution of scores

7 Conclusion

We investigate how supervised fine-tuning affects the calibration of confidence metrics for NLG tasks. Across 216 configurations, calibration degrades in over half of cases post-SFT, demonstrating these metrics cannot be deployed off-the-shelf without validation. We find that probability-based metrics exhibit relative overconfidence while consistency-based metrics show relative underconfidence, with miscalibration often occurring when confidence scores shift independently of output quality changes. Our downstream task evaluation shows performance degradation in 35/72 cases post-SFT. These findings reveal that confidence scores are influenced by factors beyond output quality, such as proximity to training data, motivating the need for calibration-robust UQ methods under model adaptation.

Limitations and Potential Risks

While our study provides comprehensive evidence across 216 configurations spanning multiple models and tasks, several areas warrant further investigation. Our detailed mechanistic analyses in Sec-

tion 5 focus on representative metrics (average log probabilities and BLEU variance) to deeply understand calibration dynamics; extending this analysis to all twelve metrics would strengthen these findings. Similarly, our investigation of miscalibration mechanisms emphasizes Case 1 (relative confidence flips); Case 2 merits additional study. We evaluate three diverse NLG tasks—translation, question answering, and mathematical reasoning—which span different generation characteristics; however, calibration sensitivity under SFT should be verified on additional task types such as summarization and dialogue.

Regarding evaluation methodology, PRR makes assumptions about the monotonic relationship between confidence and quality. While PRR is established for evaluating UQ in generation tasks (Malinin and Gales, 2021; Vashurin et al., 2025), developing more sophisticated calibration metrics for NLG remains an important research direction. Our findings are robust to this limitation, as they demonstrate *relative* changes in calibration rather than absolute values.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, and 12 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Nico Daheim, Clara Meister, Thomas Möllenhoff, and Iryna Gurevych. 2025. Uncertainty-aware decoding with minimum bayes risk. In *The Thirteenth International Conference on Learning Representations*.

Yassir Fathullah, Guoxuan Xia, and Mark John Francis Gales. 2023. Logit-based ensemble distribution distillation for robust autoregressive sequence uncertainties. *ArXiv*, abs/2305.10384.

Lorenzo Jaime Yu Flores, Ori Ernst, and Jackie CK Cheung. 2025. Improving the calibration of confidence scores in text generation using the output distribution's characteristics. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 172–182, Vienna, Austria. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Haixia Han, Tingyun Li, Shisong Chen, Jie Shi, Chengyu Du, Yanghua Xiao, Jiaqing Liang, and Xin Lin. 2024. Enhancing confidence expression in large language models through learning from past experience. *Preprint*, arXiv:2404.10315.

Jerry Huang, Peng Lu, and QIUHAO Zeng. 2025. Calibrated language models and how to find them with label smoothing. In *Forty-second International Conference on Machine Learning*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. Large language models must be taught to know what they don't know. *Preprint*, arXiv:2406.08391.

Nikita Kotelevskii, Vladimir Kondratyev, Martin Takáč, Éric Moulines, and Maxim Panov. 2025. From risk to uncertainty: Generating predictive uncertainty measures via bayesian estimation. *Preprint*, arXiv:2402.10727.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *ArXiv*, abs/1903.00802.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Preprint*, arXiv:1612.01474.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2025. Taming overconfidence in llms: Reward calibration in rlhf. *Preprint*, arXiv:2410.09724.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Preprint*, arXiv:2006.10108.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. *Preprint*, arXiv:2002.07650.

Andrey Malinin, Bruno Mlodozeniec, and Mark John Francis Gales. 2019. Ensemble distribution distillation. *ArXiv*, abs/1905.00076.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *ArXiv*, abs/2405.20003.

Team NLLB. 2022. No language left behind: Scaling human-centered machine translation.

Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein-Dor. 2023. Active learning for natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9862–9877, Singapore. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. 2025. Humans overrely on overconfident language models, across languages. *Preprint*, arXiv:2507.06306.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-evaluation improves selective generation in large language models. *Preprint*, arXiv:2312.09300.

Maximilian Schmidt, A. Bartezzaghi, Jasmina Bogojeska, Adelmo Cristiano Innocenza Malossi, and Thang Vu. 2022. Combining data generation and active learning for low-resource question answering. In *International Conference on Artificial Neural Networks*.

Artem Shelmanov, Maxim Panov, Roman Vashurin, Artem Vazhentsev, Ekaterina Fadeeva, and Timothy Baldwin. 2025. Uncertainty quantification for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 3–4, Vienna, Austria. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025. Uncertainty quantification for llms through minimum bayes risk: Bridging confidence and consistency. *Preprint*, arXiv:2502.04964.

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1430–1454, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yu-Hsiang Wang, Andrew Bai, Che-Ping Tsai, and Cho-Jui Hsieh. 2024. Clue: Concept-level uncertainty estimation for large language models. *Preprint*, arXiv:2409.03021.

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

Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, and Amir Salman Avestimehr. 2024. Do not design, learn: A trainable scoring function for uncertainty estimation in generative llms. *ArXiv*, abs/2406.11278.

Polina Zablotskaia, Du Phan, Joshua Maynez, Shashi Narayan, Jie Ren, and Jeremiah Liu. 2023. On uncertainty calibration and selective generation in probabilistic neural summarization: A benchmark study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2980–2992, Singapore. Association for Computational Linguistics.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.

A Confidence Metrics Equations

$$\text{Conf} = \frac{1}{n_{\text{dropout}}} \sum_{i=1}^{n_{\text{dropout}}} \frac{1}{|\hat{y}^{(i)}|} \sum_{t=1}^{|\hat{y}^{(i)}|} \mathcal{H}\left(p(\hat{y}_t^{(i)}|\hat{y}_{<t}^{(i)}, x)\right) \tag{1}$$

$$\mathcal{H}\left(p(\hat{y}_t^{(i)}|\hat{y}_{<t}^{(i)}, x)\right) = -\sum_{j=1}^{|\mathcal{V}|} p(\hat{y}_{t,j}^{(i)}|\hat{y}_{<t}^{(i)}, x) \log\left(p(\hat{y}_{t,j}^{(i)}|\hat{y}_{<t}^{(i)}, x)\right)$$

$$\text{Conf} = -\sum_{i=1}^{10} \pi_i \left(\frac{1}{|\hat{y}^{(i)}|} \ln(p(\hat{y}^{(i)}))\right) \tag{2}$$

$$\pi_i = \frac{\exp\left(\frac{1}{|\hat{y}^{(i)}|}\ln(p(\hat{y}^{(i)}))\right)}{\sum_{j=1}^{10} \exp\left(\frac{1}{|\hat{y}^{(j)}|}\ln(p(\hat{y}^{(j)}))\right)}$$

$$\ln(p(\hat{y}^{(i)})) = \sum_{t=1}^{|\hat{y}^{(i)}|} \ln(p(\hat{y}_t^{(i)}|\hat{y}_{<t}^{(i)}, x))$$

$$\text{Conf} = \frac{\sum_{i=1}^{n_{\text{Dropout}}} \sum_{j=1}^{n_{\text{Dropout}}} \text{Meteor}(\hat{y}^{(i)}, \hat{y}^{(j)})}{N(N-1)} \tag{3}$$

$$\text{Conf} = \sum_{i=1}^{n_{\text{Dropout}}} \sum_{j=1}^{n_{\text{Dropout}}} (1 - \text{BLEU}(\hat{y}^{(i)}, \hat{y}^{(j)}))^2 \tag{4}$$

$$\text{Conf} = \sum_{i=1}^{n_{\text{Dropout}}} KL(p(\hat{y}^{(i)}|x), p_{\bar{y}}) \tag{5}$$

$$\bar{y}_{\text{Prob}} = \frac{1}{n_{\text{Dropout}}} \sum_{i=1}^{n_{\text{Dropout}}} p(\hat{y}^{(i)}|x)$$

Where $\hat{y}^{(i)}$ is the decoded sequence i sampled by activating dropout, $\hat{y}_t^{(i)}$ is the t-th output token for sequence i, and $\hat{y}_{t,j}^{(i)}$ is the j-th vocabulary at position t for sequence i. When dropout is used, we sample $n_{\text{Dropout}} = 3$ instances.

B calibration Results

We report the maximum and minimum calibration by model and task, across the various confidence metrics in Tables 6, 7, and 8. We show the calibration by samples in Figure 6. We show the plots of the average log probs and dropout BLEU variance values by epoch when fine-tuning with 2000 samples in Figure 7 and Figure 8.

C Miscalibration Analyses

We report details of the analyses referenced in Sections 5 in Tables 9, 10, 11, and 12.

	BART-Base			Flan-T5-Base			Llama 3.1-8B		
	Eng-Afr	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**
Avg Tok Prob	0.131 (0.019)	0.01 (0.011)	0.016 (0.014)	0.161 (0.021)	0.131 (0.02)	0.019 (0.006)	0.111 (0.033)	-0.115 (0.107)	0.101 (0.053)
Avg Tok Ent	-0.132 (0.017)	-0.011 (0.005)	-0.014 (0.013)	-0.159 (0.018)	-0.139 (0.021)	-0.018 (0.004)	-0.101 (0.033)	0.116 (0.097)	-0.099 (0.056)
DO Ent	-0.114 (0.026)	0.002 (0.014)	-0.006 (0.007)	-0.194 (0.065)	-0.083 (0.023)	-0.02 (0.011)	-0.058 (0.027)	0.17 (0.112)	-0.026 (0.027)
BS Imp Wt	0.124 (0.021)	-0.05 (0.033)	0.007 (0.011)	0.189 (0.031)	0.095 (0.012)	0.013 (0.002)	0.11 (0.026)	-0.099 (0.083)	0.087 (0.053)
BS Ratios	0.11 (0.027)	0.046 (0.013)	0.016 (0.009)	0.227 (0.054)	0.011 (0.008)	0.018 (0.012)	0.03 (0.034)	-0.334 (0.102)	0.103 (0.025)
BS Sums	0.124 (0.02)	-0.049 (0.036)	0.007 (0.01)	0.189 (0.031)	0.096 (0.012)	0.013 (0.002)	0.11 (0.026)	-0.096 (0.084)	0.086 (0.055)
DO Bleu Var	-0.108 (0.024)	-0.031 (0.014)	-0.01 (0.012)	-0.209 (0.045)	-0.063 (0.015)	-0.02 (0.003)	-0.02 (0.038)	0.161 (0.088)	-0.029 (0.04)
DO KL Div	-0.143 (0.007)	-0.047 (0.008)	-0.005 (0.01)	-0.207 (0.054)	-0.023 (0.06)	-0.011 (0.005)	-0.0 (0.039)	0.239 (0.15)	-0.038 (0.035)
DO Meteor Var	0.131 (0.024)	0.004 (0.042)	0.008 (0.009)	0.176 (0.04)	0.164 (0.024)	0.016 (0.008)	0.041 (0.056)	-0.173 (0.12)	0.022 (0.085)
CoCoA MSP	-0.125 (0.015)	-0.006 (0.022)	-0.01 (0.013)	-0.178 (0.041)	-0.015 (0.024)	-0.015 (0.007)	-0.028 (0.021)	0.112 (0.167)	0.007 (0.033)
CoCoA MTE	-0.139 (0.035)	-0.021 (0.008)	-0.01 (0.008)	-0.143 (0.036)	-0.268 (0.058)	-0.015 (0.008)	-0.072 (0.072)	0.16 (0.093)	-0.101 (0.059)
CoCoA PPL	-0.135 (0.032)	-0.019 (0.009)	-0.011 (0.008)	-0.145 (0.034)	-0.265 (0.048)	-0.016 (0.009)	-0.081 (0.072)	0.158 (0.103)	-0.103 (0.058)
	Gemma 2-2B			**Gemma 2-9B**			**QWen 2.5-14B**		
	Eng-Afr	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**
Avg Tok Prob	0.147 (0.089)	0.157 (0.196)	0.002 (0.029)	0.019 (0.159)	-0.007 (0.01)	0.078 (0.059)	0.023 (0.101)	-0.005 (0.01)	0.007 (0.073)
Avg Tok Ent	-0.149 (0.084)	-0.161 (0.208)	0.015 (0.029)	-0.021 (0.139)	0.006 (0.016)	-0.06 (0.061)	-0.024 (0.099)	0.005 (0.012)	-0.011 (0.044)
DO Ent	-0.081 (0.133)	-0.051 (0.18)	0.015 (0.046)	0.028 (0.23)	0.021 (0.017)	-0.095 (0.022)	0.061 (0.102)	0.027 (0.009)	-0.051 (0.053)
BS Imp Wt	0.175 (0.086)	0.033 (0.125)	-0.003 (0.016)	0.007 (0.153)	-0.0 (0.024)	0.092 (0.046)	0.016 (0.087)	0.004 (0.064)	0.01 (0.063)
BS Ratios	0.067 (0.139)	0.07 (0.118)	0.048 (0.102)	-0.009 (0.253)	-0.002 (0.002)	-0.038 (0.017)	-0.065 (0.092)	-0.007 (0.006)	-0.0 (0.052)
BS Sums	0.175 (0.086)	0.032 (0.127)	-0.005 (0.016)	0.004 (0.156)	0.002 (0.021)	0.091 (0.048)	0.015 (0.087)	0.003 (0.067)	0.01 (0.062)
DO Bleu Var	-0.071 (0.087)	-0.052 (0.133)	0.014 (0.035)	0.047 (0.204)	0.002 (0.002)	-0.036 (0.101)	0.061 (0.081)	-0.005 (0.067)	0.015 (0.09)
DO KL Div	-0.061 (0.129)	-0.123 (0.146)	-0.013 (0.113)	0.079 (0.206)	-0.0 (0.004)	-0.027 (0.134)	0.079 (0.117)	0.01 (0.026)	-0.049 (0.05)
DO Meteor Var	0.073 (0.122)	0.103 (0.097)	0.045 (0.027)	-0.046 (0.206)	0.013 (0.004)	0.04 (0.134)	-0.082 (0.108)	-0.014 (0.048)	0.056 (0.032)
CoCoA MSP	-0.003 (0.113)	-0.091 (0.093)	0.008 (0.022)	0.106 (0.238)	0.01 (0.04)	-0.013 (0.027)	0.068 (0.057)	-0.042 (0.068)	-0.096 (0.027)
CoCoA MTE	-0.175 (0.09)	-0.119 (0.162)	-0.019 (0.062)	-0.006 (0.209)	0.002 (0.021)	-0.082 (0.054)	0.038 (0.101)	0.014 (0.029)	-0.043 (0.051)
CoCoA PPL	-0.172 (0.092)	-0.124 (0.165)	-0.026 (0.054)	-0.011 (0.213)	0.004 (0.016)	-0.088 (0.045)	0.029 (0.101)	0.012 (0.019)	-0.04 (0.064)

Table 6: Difference in PRR after one epoch of SFT and after performing SFT with early stopping; Averaged across 3 seeds with st. dev in parentheses

D Case Study Details

Prompt We use Anthropic Claude-3.5 Sonnet to rate whether or not the answer to the question is correct. We provide the original question, together with the reference answer in the dataset. We set the temperature to 0.

```
  Question: <question>
Correct Answer: <label>
Predicted Answer: <prediction>
Is the predicted answer correct, based on
the correct answer?
Only return Yes or No
```

E Dataset Details

We use the NLLB dataset (NLLB, 2022) under the ODC-By License, the FLORES Plus (Goyal et al., 2021) and SQUAD datasets (Rajpurkar et al., 2016) under the CC BY-SA 4.0 License, and the GSM8K dataset (Cobbe et al., 2021) under the MIT License, which allow the use of these datasets for research purposes. We scan the datasets to check that there are no malicious or harmful content in the translation pairs.

For translation, we use 10K samples from NLLB as the train set, an additional 100 samples from NLLB as the validation set, and 253 samples from FLORES as the test set. For math, we use 7K samples from the train set for training, an additional 100 samples from the train set as a validation set, and 1000 samples from the test set as the test. For SQUAD, we use 10K samples from the train set for training, and an additional 100 samples from the train set as a validation set. We use 1000 samples from the test set as the test.

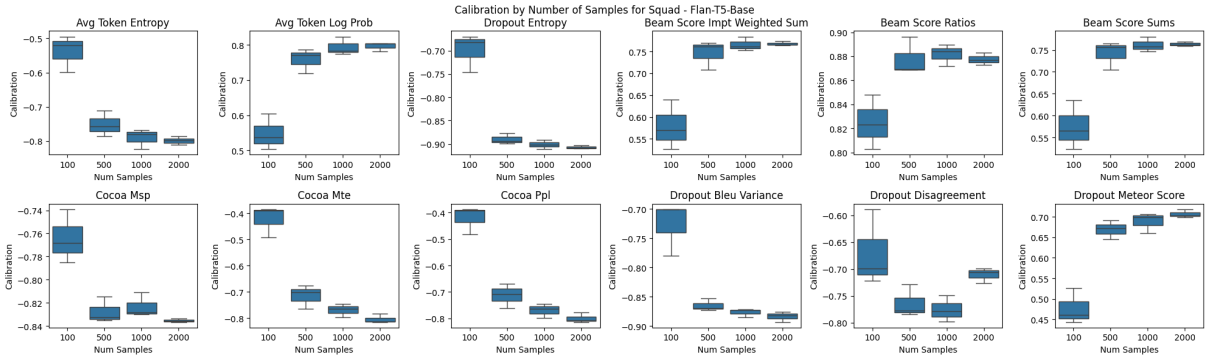
F Computational Details

Unless otherwise specified, we use a batch size of 8 and constant learning rate of 5e-5. We train models for a maximum of 200 epochs, but employ early stopping with a patience of 2 epochs; training is stopped once the validation metric does not increase on a validation set of 100 samples. We perform all fine-tuning and inference using one RTX 8000 GPU.

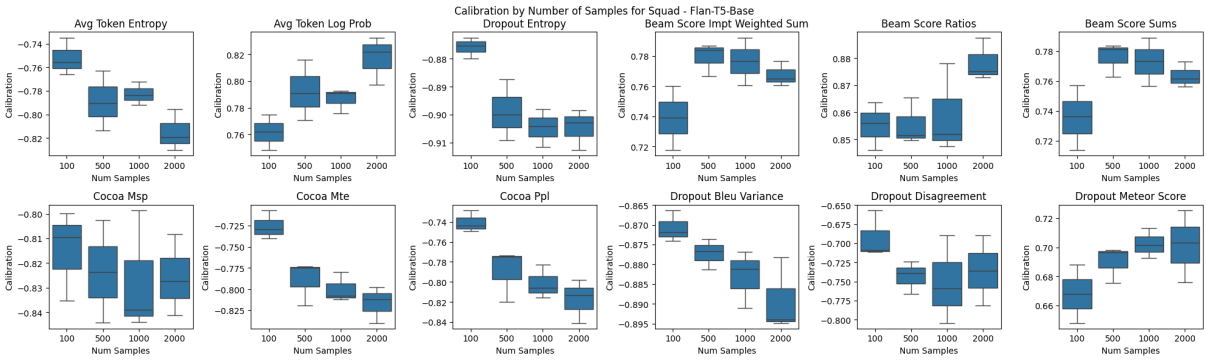
	BART-Base			Flan-T5-Base			Llama 3.1-8B		
	Eng-Afr	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**
Avg Tok Prob	-0.080 (0.054)	-0.036 (0.013)	-0.022 (0.009)	-0.106 (0.030)	-0.029 (0.015)	-0.020 (0.007)	-0.178 (0.028)	-0.116 (0.029)	-0.171 (0.073)
Avg Tok Ent	-0.079 (0.047)	-0.030 (0.007)	-0.021 (0.002)	-0.105 (0.030)	-0.074 (0.037)	-0.021 (0.009)	-0.120 (0.035)	-0.068 (0.064)	-0.161 (0.077)
DO Ent	-0.071 (0.051)	-0.053 (0.014)	-0.016 (0.008)	-0.100 (0.020)	-0.041 (0.006)	-0.021 (0.008)	-0.136 (0.018)	-0.079 (0.053)	-0.145 (0.054)
BS Imp Wt	-0.070 (0.041)	-0.049 (0.010)	-0.022 (0.006)	-0.115 (0.035)	-0.026 (0.004)	-0.014 (0.009)	-0.167 (0.033)	-0.121 (0.038)	-0.173 (0.076)
BS Ratios	-0.070 (0.034)	-0.032 (0.012)	-0.021 (0.004)	-0.108 (0.064)	-0.162 (0.010)	-0.018 (0.007)	-0.104 (0.023)	-0.221 (0.051)	-0.197 (0.047)
BS Sums	-0.070 (0.041)	-0.124 (0.041)	-0.021 (0.005)	-0.115 (0.035)	-0.107 (0.057)	-0.014 (0.009)	-0.166 (0.034)	-0.130 (0.090)	-0.173 (0.076)
DO BLEU Var	-0.067 (0.036)	-0.066 (0.015)	-0.026 (0.012)	-0.108 (0.046)	-0.034 (0.006)	-0.022 (0.011)	-0.092 (0.028)	-0.095 (0.061)	-0.120 (0.016)
DO KL Div	-0.081 (0.033)	-0.053 (0.031)	-0.021 (0.012)	-0.090 (0.015)	-0.209 (0.086)	-0.026 (0.009)	-0.088 (0.016)	-0.267 (0.196)	-0.096 (0.015)
DO Meteor Var	-0.072 (0.029)	-0.063 (0.010)	-0.019 (0.005)	-0.072 (0.004)	-0.024 (0.018)	-0.016 (0.005)	-0.120 (0.043)	-0.214 (0.133)	-0.119 (0.074)
CoCoA MSP	-0.076 (0.039)	-0.056 (0.014)	-0.019 (0.008)	-0.088 (0.012)	-0.059 (0.012)	-0.016 (0.007)	-0.074 (0.013)	-0.251 (0.076)	-0.084 (0.039)
CoCoA MTE	-0.078 (0.043)	-0.030 (0.007)	-0.020 (0.003)	-0.102 (0.025)	-0.127 (0.046)	-0.015 (0.004)	-0.128 (0.037)	-0.155 (0.047)	-0.180 (0.073)
CoCoA PPL	-0.077 (0.046)	-0.032 (0.007)	-0.021 (0.004)	-0.104 (0.025)	-0.122 (0.036)	-0.015 (0.004)	-0.130 (0.038)	-0.167 (0.064)	-0.179 (0.062)

	Gemma 2-2B			Gemma 2-9B			QWen 2.5-14B		
	Eng-Afr	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**	**Eng-Afr**	**SQUAD**	**GSM8K**
Avg Tok Prob	-0.155 (0.016)	-0.200 (0.097)	-0.129 (0.031)	-0.141 (0.037)	-0.054 (0.018)	-0.158 (0.031)	-0.191 (0.071)	-0.031 (0.012)	-0.118 (0.007)
Avg Tok Ent	-0.153 (0.010)	-0.201 (0.095)	-0.107 (0.030)	-0.144 (0.033)	-0.056 (0.023)	-0.149 (0.040)	-0.191 (0.070)	-0.043 (0.018)	-0.101 (0.024)
DO Ent	-0.120 (0.016)	-0.179 (0.089)	-0.126 (0.036)	-0.178 (0.088)	-0.049 (0.011)	-0.187 (0.031)	-0.237 (0.105)	-0.055 (0.036)	-0.111 (0.033)
BS Imp Wt	-0.159 (0.005)	-0.154 (0.073)	-0.121 (0.023)	-0.137 (0.046)	-0.090 (0.021)	-0.158 (0.033)	-0.197 (0.076)	-0.059 (0.033)	-0.108 (0.047)
BS Ratios	-0.131 (0.013)	-0.076 (0.044)	-0.080 (0.048)	-0.223 (0.101)	-0.009 (0.005)	-0.159 (0.066)	-0.279 (0.124)	-0.018 (0.012)	-0.142 (0.014)
BS Sums	-0.159 (0.005)	-0.152 (0.077)	-0.120 (0.023)	-0.137 (0.046)	-0.090 (0.020)	-0.158 (0.033)	-0.197 (0.077)	-0.062 (0.037)	-0.107 (0.046)
DO BLEU Var	-0.122 (0.010)	-0.156 (0.035)	-0.099 (0.046)	-0.189 (0.076)	-0.052 (0.011)	-0.131 (0.035)	-0.272 (0.140)	-0.036 (0.018)	-0.173 (0.067)
DO KL Div	-0.110 (0.013)	-0.183 (0.051)	-0.096 (0.027)	-0.218 (0.047)	-0.061 (0.030)	-0.142 (0.054)	-0.287 (0.141)	-0.058 (0.011)	-0.148 (0.047)
DO Meteor Var	-0.143 (0.059)	-0.142 (0.062)	-0.095 (0.031)	-0.189 (0.076)	-0.058 (0.009)	-0.166 (0.025)	-0.279 (0.127)	-0.044 (0.019)	-0.176 (0.046)
CoCoA MSP	-0.110 (0.053)	-0.151 (0.100)	-0.083 (0.050)	-0.213 (0.103)	-0.070 (0.001)	-0.094 (0.059)	-0.290 (0.150)	-0.038 (0.009)	-0.132 (0.085)
CoCoA MTE	-0.157 (0.008)	-0.194 (0.070)	-0.135 (0.078)	-0.156 (0.048)	-0.048 (0.023)	-0.192 (0.041)	-0.199 (0.077)	-0.057 (0.018)	-0.131 (0.026)
CoCoA PPL	-0.157 (0.009)	-0.190 (0.072)	-0.132 (0.080)	-0.153 (0.045)	-0.048 (0.017)	-0.194 (0.042)	-0.196 (0.081)	-0.044 (0.015)	-0.130 (0.024)

Table 7: From epoch-to-epoch, there can be a considerable decline in calibration, reaching differences of up to 0.290 PRR points; Results averaged over 3 seeds with st. dev in parentheses



(a) calibration of different confidence metrics after fine-tuning for 10 epochs.



(b) PRR of different confidence metrics after 2500 fine-tuning steps.

Figure 6: PRR of confidence metrics differs significantly depending on the number of fine-tuning samples. Plots shown for BART-Base fine-tuned on SQUAD for (top) 10 epochs and (bottom) 2500 steps using 3 seeds.

	Eng-Afr			SQUAD			GSM8K		
	No SFT	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**
Avg Tok Prob	0.092 (0.014)	0.347 (0.102)	0.37 (0.047)	0.064 (0.0)	0.364 (0.045)	0.345 (0.015)	0.0 (0.0)	-0.016 (0.041)	0.024 (0.037)
Avg Tok Ent	0.109 (0.015)	0.321 (0.112)	0.378 (0.051)	0.048 (0.0)	0.384 (0.05)	0.379 (0.017)	0.0 (0.0)	-0.007 (0.042)	0.023 (0.029)
DO Ent	0.047 (0.099)	0.125 (0.067)	0.11 (0.094)	-0.24 (0.023)	-0.004 (0.074)	-0.059 (0.103)	0.0 (0.0)	-0.022 (0.04)	-0.021 (0.036)
BS Imp Wt	-0.235 (0.0)	-0.435 (0.091)	-0.448 (0.069)	-0.144 (0.0)	-0.127 (0.06)	0.083 (0.03)	0.0 (0.0)	-0.047 (0.012)	-0.001 (0.033)
BS Ratios	-0.218 (0.0)	0.03 (0.068)	-0.068 (0.046)	0.239 (0.0)	0.276 (0.017)	0.369 (0.048)	0.0 (0.0)	-0.019 (0.035)	0.005 (0.046)
BS Sums	0.237 (0.0)	0.435 (0.091)	0.448 (0.067)	0.094 (0.0)	-0.039 (0.046)	-0.238 (0.049)	0.0 (0.0)	0.048 (0.011)	0.0 (0.033)
DO BLEU Var	0.012 (0.125)	0.385 (0.018)	0.302 (0.037)	-0.047 (0.017)	0.063 (0.014)	0.099 (0.072)	0.0 (0.0)	-0.008 (0.052)	0.006 (0.052)
DO KL Div	-0.026 (0.121)	-0.05 (0.053)	0.042 (0.064)	0.147 (0.02)	0.245 (0.036)	0.347 (0.022)	0.0 (0.0)	0.016 (0.042)	0.013 (0.011)
DO Meteor Var	-0.012 (0.094)	0.397 (0.016)	0.292 (0.07)	-0.017 (0.028)	0.096 (0.009)	0.132 (0.079)	0.0 (0.0)	0.009 (0.03)	0.015 (0.013)
CoCoA MSP	0.128 (0.009)	0.165 (0.12)	0.191 (0.092)	0.221 (0.0)	0.414 (0.014)	0.443 (0.034)	0.0 (0.0)	0.006 (0.029)	0.002 (0.028)
CoCoA MTE	0.182 (0.005)	0.225 (0.115)	0.287 (0.072)	0.06 (0.0)	0.43 (0.016)	0.464 (0.033)	0.0 (0.0)	-0.005 (0.016)	0.005 (0.022)
CoCoA PPL	0.167 (0.006)	0.257 (0.116)	0.295 (0.066)	0.07 (0.0)	0.42 (0.015)	0.445 (0.031)	0.0 (0.0)	-0.011 (0.017)	0.006 (0.026)

	Eng-Afr			SQUAD			GSM8K		
	No SFT	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**
Avg Tok Prob	-0.428 (0.0)	-0.172 (0.082)	-0.166 (0.249)	0.036 (0.0)	-0.048 (0.042)	-0.038 (0.034)	0.0 (0.0)	0.012 (0.027)	0.006 (0.027)
Avg Tok Ent	-0.375 (0.0)	-0.105 (0.088)	-0.155 (0.259)	0.027 (0.0)	-0.051 (0.037)	-0.034 (0.048)	0.0 (0.0)	0.013 (0.029)	0.003 (0.017)
DO Ent	-0.254 (0.044)	0.116 (0.028)	0.264 (0.11)	0.313 (0.011)	0.388 (0.011)	0.397 (0.031)	0.0 (0.0)	0.036 (0.013)	0.038 (0.043)
BS Imp Wt	0.557 (0.0)	0.314 (0.04)	0.131 (0.269)	-0.079 (0.0)	0.015 (0.019)	0.103 (0.015)	0.0 (0.0)	0.03 (0.027)	0.023 (0.014)
BS Ratios	0.365 (0.0)	-0.05 (0.143)	0.384 (0.184)	0.725 (0.0)	0.634 (0.045)	0.374 (0.057)	0.0 (0.0)	0.014 (0.03)	0.056 (0.052)
BS Sums	-0.557 (0.0)	-0.314 (0.04)	-0.132 (0.269)	0.053 (0.0)	-0.034 (0.018)	-0.124 (0.015)	0.0 (0.0)	-0.03 (0.027)	-0.023 (0.014)
DO BLEU Var	0.285 (0.031)	0.21 (0.063)	0.287 (0.033)	0.359 (0.01)	0.446 (0.028)	0.397 (0.03)	0.0 (0.0)	0.01 (0.018)	0.021 (0.038)
DO KL Div	-0.08 (0.099)	0.055 (0.094)	0.09 (0.113)	0.371 (0.016)	0.324 (0.068)	0.241 (0.075)	0.0 (0.0)	0.017 (0.015)	0.012 (0.065)
DO Meteor Var	0.28 (0.051)	0.244 (0.092)	0.307 (0.086)	0.354 (0.013)	0.337 (0.024)	0.269 (0.039)	0.0 (0.0)	0.007 (0.011)	0.016 (0.016)
CoCoA MSP	0.346 (0.0)	0.054 (0.024)	0.168 (0.116)	-0.175 (0.0)	-0.379 (0.047)	-0.097 (0.064)	0.0 (0.0)	0.006 (0.035)	0.006 (0.028)
CoCoA MTE	0.447 (0.0)	0.036 (0.047)	-0.205 (0.228)	-0.178 (0.0)	-0.378 (0.048)	-0.083 (0.078)	0.0 (0.0)	-0.004 (0.044)	-0.001 (0.019)
CoCoA PPL	0.346 (0.0)	-0.012 (0.024)	-0.218 (0.226)	-0.175 (0.0)	-0.379 (0.047)	-0.078 (0.076)	0.0 (0.0)	-0.004 (0.045)	0.001 (0.024)

	Eng-Afr			SQUAD			GSM8K		
	No SFT	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**
Avg Tok Prob	0.121 (0.039)	-0.012 (0.049)	0.245 (0.099)	0.068 (0.065)	-0.381 (0.041)	-0.275 (0.08)	0.0 (0.0)	0.123 (0.025)	0.274 (0.103)
Avg Tok Ent	0.124 (0.039)	-0.006 (0.063)	0.232 (0.085)	0.051 (0.061)	-0.386 (0.035)	-0.279 (0.094)	0.0 (0.0)	0.114 (0.051)	0.258 (0.091)
DO Ent	0.02 (0.137)	-0.02 (0.127)	0.148 (0.22)	0.131 (0.085)	-0.185 (0.101)	-0.016 (0.095)	0.0 (0.0)	0.009 (0.014)	-0.032 (0.068)
BS Imp Wt	-0.126 (0.044)	0.052 (0.048)	-0.237 (0.078)	-0.089 (0.067)	0.419 (0.024)	0.234 (0.122)	0.0 (0.0)	-0.125 (0.028)	-0.264 (0.102)
BS Ratios	0.011 (0.098)	0.173 (0.22)	0.172 (0.178)	0.111 (0.068)	0.653 (0.078)	0.401 (0.202)	0.0 (0.0)	-0.013 (0.076)	0.204 (0.097)
BS Sums	0.127 (0.044)	-0.051 (0.048)	0.236 (0.078)	0.09 (0.067)	-0.436 (0.017)	-0.238 (0.121)	0.0 (0.0)	0.126 (0.029)	0.264 (0.102)
DO Bleu Var	-0.015 (0.094)	0.026 (0.075)	0.0 (0.0)	-0.13 (0.082)	-0.085 (0.088)	-0.07 (0.189)	0.0 (0.0)	0.019 (0.032)	0.0 (0.0)
DO KL Div	-0.114 (0.098)	-0.002 (0.042)	-0.116 (0.087)	-0.14 (0.068)	0.056 (0.094)	0.11 (0.115)	0.0 (0.0)	0.013 (0.054)	-0.018 (0.066)
DO Meteor Var	-0.042 (0.088)	0.007 (0.104)	-0.108 (0.179)	-0.022 (0.088)	-0.141 (0.056)	-0.174 (0.011)	0.0 (0.0)	0.032 (0.11)	0.092 (0.08)
CoCoA MSP	0.207 (0.022)	0.215 (0.143)	0.201 (0.031)	0.14 (0.121)	0.335 (0.062)	-0.076 (0.165)	0.0 (0.0)	0.154 (0.014)	0.275 (0.119)
CoCoA MTE	0.228 (0.033)	0.032 (0.181)	0.184 (0.054)	0.059 (0.074)	-0.253 (0.031)	-0.196 (0.118)	0.0 (0.0)	0.146 (0.034)	0.262 (0.109)
CoCoA PPL	0.207 (0.022)	0.017 (0.179)	0.201 (0.031)	0.07 (0.079)	-0.26 (0.027)	-0.201 (0.1)	0.0 (0.0)	0.154 (0.014)	0.275 (0.119)

	Eng-Afr			SQUAD			GSM8K		
	No SFT	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**
Avg Tok Prob	0.0 (0.043)	-0.184 (0.253)	0.327 (0.089)	0.268 (0.048)	-0.018 (0.08)	0.019 (0.18)	0.0 (0.0)	0.163 (0.077)	0.155 (0.075)
Avg Tok Ent	-0.043 (0.047)	-0.169 (0.242)	0.314 (0.081)	0.26 (0.04)	-0.006 (0.1)	0.03 (0.187)	0.0 (0.0)	0.174 (0.07)	0.135 (0.062)
DO Ent	-0.143 (0.02)	0.004 (0.032)	0.049 (0.116)	0.177 (0.056)	0.006 (0.067)	-0.094 (0.081)	0.0 (0.0)	0.022 (0.072)	-0.017 (0.083)
BS Imp Wt	-0.002 (0.043)	0.173 (0.264)	-0.395 (0.079)	-0.263 (0.041)	-0.173 (0.111)	0.116 (0.142)	0.0 (0.0)	-0.218 (0.141)	-0.238 (0.081)
BS Ratios	0.098 (0.04)	0.089 (0.216)	0.018 (0.029)	0.093 (0.107)	0.339 (0.105)	0.324 (0.075)	0.0 (0.0)	0.169 (0.042)	0.255 (0.1)
BS Sums	0.003 (0.043)	-0.173 (0.265)	0.394 (0.078)	0.263 (0.041)	0.129 (0.109)	-0.155 (0.139)	0.0 (0.0)	0.217 (0.139)	0.232 (0.083)
DO Bleu Var	-0.068 (0.033)	-0.052 (0.088)	-0.02 (0.087)	-0.005 (0.009)	0.027 (0.069)	-0.024 (0.074)	0.0 (0.0)	-0.011 (0.066)	-0.044 (0.039)
DO KL Div	0.044 (0.045)	0.109 (0.026)	0.047 (0.064)	-0.111 (0.045)	-0.041 (0.114)	0.037 (0.047)	0.0 (0.0)	-0.021 (0.22)	-0.025 (0.119)
DO Meteor Var	-0.063 (0.095)	-0.024 (0.086)	-0.025 (0.106)	-0.074 (0.059)	0.026 (0.102)	-0.012 (0.073)	0.0 (0.0)	-0.012 (0.088)	-0.068 (0.107)
CoCoA MSP	0.138 (0.063)	-0.12 (0.066)	0.341 (0.039)	0.297 (0.032)	0.159 (0.06)	0.072 (0.234)	0.0 (0.0)	0.183 (0.063)	0.192 (0.091)
CoCoA MTE	0.143 (0.068)	-0.257 (0.2)	0.34 (0.11)	0.289 (0.023)	0.095 (0.064)	0.044 (0.195)	0.0 (0.0)	0.081 (0.1)	0.091 (0.041)
CoCoA PPL	0.138 (0.063)	-0.262 (0.207)	0.343 (0.109)	0.297 (0.032)	0.083 (0.046)	0.033 (0.195)	0.0 (0.0)	0.079 (0.09)	0.113 (0.063)

	Eng-Afr			SQUAD			GSM8K		
	No SFT	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**
Avg Tok Prob	-0.703 (0.028)	-0.609 (0.094)	-0.630 (0.060)	-0.841 (0.014)	-0.959 (0.017)	-0.953 (0.004)	-0.500 (0.000)	-0.566 (0.028)	-0.626 (0.062)
Avg Tok Ent	0.717 (0.023)	0.600 (0.114)	0.619 (0.058)	0.848 (0.013)	0.960 (0.014)	0.953 (0.006)	0.500 (0.000)	0.551 (0.026)	0.629 (0.051)
DO Ent	0.704 (0.011)	0.612 (0.156)	0.584 (0.088)	0.856 (0.016)	0.973 (0.004)	0.952 (0.014)	0.500 (0.000)	0.326 (0.077)	0.421 (0.069)
BS Imp Wt	0.731 (0.007)	0.610 (0.109)	0.617 (0.049)	0.867 (0.005)	0.904 (0.034)	0.904 (0.012)	0.500 (0.000)	0.550 (0.034)	0.642 (0.072)
BS Ratios	0.749 (0.044)	0.577 (0.182)	0.568 (0.075)	0.904 (0.008)	0.992 (0.001)	0.991 (0.002)	0.500 (0.000)	0.551 (0.023)	0.514 (0.037)
BS Sums	0.730 (0.007)	0.610 (0.110)	0.614 (0.049)	0.866 (0.005)	0.900 (0.031)	0.902 (0.011)	0.500 (0.000)	0.549 (0.034)	0.640 (0.073)
DO BLEU Var	0.706 (0.034)	0.609 (0.148)	0.545 (0.087)	0.820 (0.016)	0.957 (0.009)	0.941 (0.012)	0.500 (0.000)	0.410 (0.072)	0.424 (0.048)
DO KL Div	0.687 (0.036)	0.623 (0.132)	0.546 (0.080)	0.801 (0.018)	0.946 (0.020)	0.945 (0.022)	0.500 (0.000)	0.403 (0.090)	0.431 (0.043)
DO Meteor Var	0.706 (0.034)	0.609 (0.148)	0.543 (0.084)	0.817 (0.016)	0.954 (0.005)	0.952 (0.005)	0.500 (0.000)	0.410 (0.072)	0.481 (0.054)
CoCoA MSP	0.631 (0.016)	0.574 (0.147)	0.467 (0.099)	0.748 (0.019)	0.916 (0.006)	0.907 (0.035)	0.500 (0.000)	0.244 (0.085)	0.256 (0.067)
CoCoA MTE	0.706 (0.030)	0.611 (0.145)	0.617 (0.069)	0.852 (0.018)	0.958 (0.022)	0.956 (0.001)	0.500 (0.000)	0.507 (0.043)	0.589 (0.056)
CoCoA PPL	0.716 (0.029)	0.605 (0.146)	0.615 (0.073)	0.860 (0.015)	0.959 (0.017)	0.954 (0.003)	0.500 (0.000)	0.505 (0.039)	0.594 (0.040)

	Eng-Afr			SQUAD			GSM8K		
	No SFT	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**	**No SFT**	**1 Ep**	**Full**
Avg Tok Prob	-0.494 (0.014)	-0.632 (0.056)	-0.657 (0.067)	-0.828 (0.010)	-0.962 (0.014)	-0.957 (0.007)	-0.500 (0.000)	-0.713 (0.015)	-0.724 (0.077)
Avg Tok Ent	0.494 (0.014)	0.632 (0.055)	0.655 (0.068)	0.828 (0.010)	0.958 (0.017)	0.953 (0.010)	0.500 (0.000)	0.713 (0.034)	0.720 (0.087)
DO Ent	0.420 (0.017)	0.684 (0.006)	0.624 (0.097)	0.727 (0.027)	0.970 (0.004)	0.943 (0.005)	0.500 (0.000)	0.659 (0.018)	0.710 (0.043)
BS Imp Wt	0.566 (0.006)	0.636 (0.055)	0.652 (0.058)	0.857 (0.014)	0.903 (0.030)	0.907 (0.042)	0.500 (0.000)	0.712 (0.019)	0.722 (0.078)
BS Ratios	0.569 (0.009)	0.712 (0.033)	0.647 (0.075)	0.863 (0.018)	0.993 (0.001)	0.986 (0.006)	0.500 (0.000)	0.733 (0.059)	0.732 (0.015)
BS Sums	0.564 (0.006)	0.636 (0.055)	0.651 (0.057)	0.854 (0.014)	0.899 (0.030)	0.902 (0.044)	0.500 (0.000)	0.712 (0.020)	0.722 (0.078)
DO BLEU Var	0.439 (0.017)	0.701 (0.047)	0.629 (0.058)	0.723 (0.025)	0.955 (0.015)	0.947 (0.028)	0.500 (0.000)	0.626 (0.092)	0.619 (0.059)
DO KL Div	0.410 (0.026)	0.688 (0.045)	0.611 (0.099)	0.767 (0.017)	0.955 (0.005)	0.946 (0.028)	0.500 (0.000)	0.620 (0.055)	0.668 (0.028)
DO Meteor Var	0.482 (0.016)	0.697 (0.046)	0.624 (0.087)	0.810 (0.029)	0.964 (0.006)	0.945 (0.036)	0.500 (0.000)	0.616 (0.069)	0.699 (0.049)
CoCoA MSP	0.461 (0.008)	0.615 (0.047)	0.547 (0.075)	0.654 (0.054)	0.899 (0.046)	0.941 (0.030)	0.500 (0.000)	0.468 (0.019)	0.564 (0.036)
CoCoA MTE	0.377 (0.015)	0.706 (0.050)	0.668 (0.064)	0.819 (0.010)	0.951 (0.009)	0.937 (0.031)	0.500 (0.000)	0.664 (0.019)	0.707 (0.054)
CoCoA PPL	0.379 (0.013)	0.696 (0.048)	0.667 (0.066)	0.821 (0.011)	0.956 (0.011)	0.944 (0.022)	0.500 (0.000)	0.667 (0.014)	0.707 (0.063)

Table 8: PRR of confidence metrics before SFT, after one epoch, and after full SFT with early stopping for BART, Flan-T5, Llama 3.1-8B, Gemma 2-2B, Gemma 2-9B, and QWen 2.5-14B

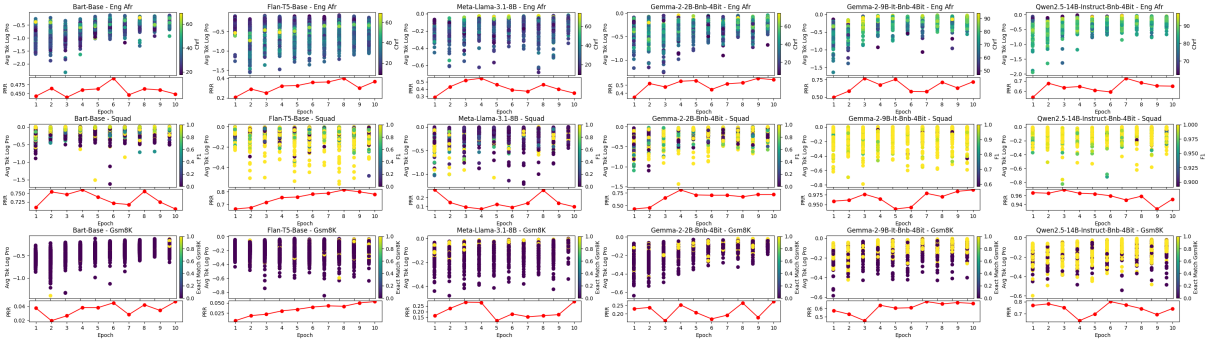


Figure 7: The average log probs across all test set samples generally increase for BART and Llama 3.1 8B (top), which may explain the rel. overconfidence in probability-based metrics, and fluctuations in calibration (bottom)

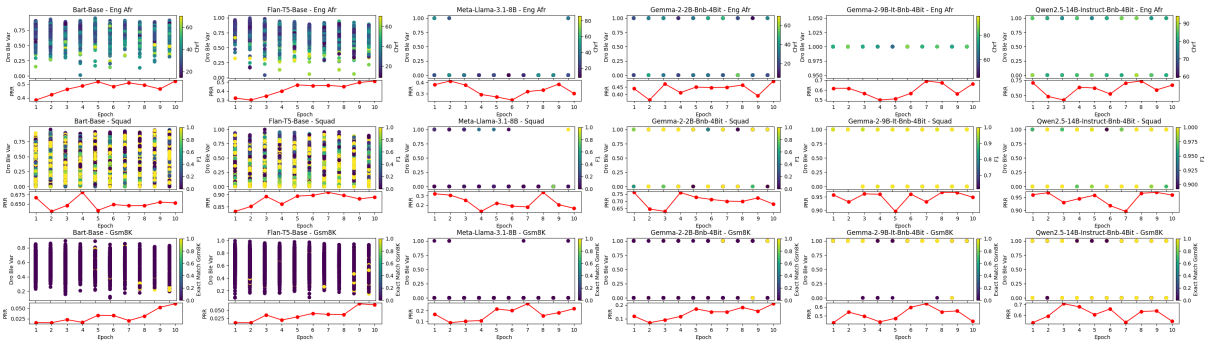


Figure 8: The dropout BLEU variance values generally do not change across epochs, which aligns with the fact that most samples remain calibrated post SFT.

	Dataset	Cncd	R. Over	R. Under
BART	Eng-Afr	0.41	0.06	0.54
	SQUAD	0.80	0.08	0.12
	GSM8K	0.97	0.01	0.02
Flan-T5	Eng-Afr	0.34	0.05	0.61
	SQUAD	0.88	0.04	0.08
	GSM8K	0.97	0.00	0.02
Llama-3.1 8B	Eng-Afr	0.94	0.00	0.06
	SQUAD	0.92	0.02	0.06
	GSM8K	1.00	0.00	0.00
Gemma-2B	Eng-Afr	0.73	0.01	0.26
	SQUAD	0.87	0.09	0.04
	GSM8K	0.92	0.01	0.07
Gemma-9B	Eng-Afr	1.00	0.00	0.00
	SQUAD	0.97	0.00	0.03
	GSM8K	0.99	0.00	0.01
QWen-14B	Eng-Afr	0.89	0.00	0.11
	SQUAD	0.92	0.07	0.01
	GSM8K	0.95	0.02	0.03

Table 9: Proportion of changes in test set samples that are concordant, relatively over/underconfident after SFT, using dropout BLEU variance as the measure of confidence; Reporting average proportions over 3 seeds

	Rel Qual.	Same		Flips	
	Rel Conf.	Same	Flips	Flips	Same
BART	Eng-Afr	0.70	0.13	0.12	0.06
	SQUAD	0.76	0.20	0.03	0.01
	GSM8K	0.63	0.18	0.14	0.05
Flan-T5	Eng-Afr	0.65	0.12	0.15	0.08
	SQUAD	0.79	0.18	0.02	0.01
	GSM8K	0.79	0.13	0.06	0.03
Llama-3.1 8B	Eng-Afr	0.48	0.05	0.24	0.22
	SQUAD	0.84	0.01	0.08	0.08
	GSM8K	0.48	0.03	0.27	0.22
Gemma-2B	Eng-Afr	0.41	0.08	0.34	0.17
	SQUAD	0.48	0.07	0.29	0.17
	GSM8K	0.70	0.08	0.13	0.08
Gemma-9B	Eng-Afr	0.47	0.00	0.27	0.26
	SQUAD	0.54	0.08	0.24	0.16
	GSM8K	0.69	0.02	0.19	0.16
QWen-14B	Eng-Afr	0.41	0.13	0.33	0.14
	SQUAD	0.57	0.06	0.20	0.15
	GSM8K	0.63	0.07	0.19	0.10

Table 10: **Case 1** miscalibration happens more frequently than **Case 2**; Table shows proportion of samples classified by change in relative quality and confidence between ep. t and $t+1$; Results avg across 10 epochs and 3 seeds, using dropout BLEU variance

	Eng-Afr				SQUAD				GSM8K			
Ep	**BART**	**T5**	**Llma**	**Gmma**	**BART**	**T5**	**Llma**	**Gmma**	**BART**	**T5**	**Llma**	**Gmma**
2	0.429	0.182	0.236	0.013	0.804	0.875	0.161	0.394	0.992	0.989	0.924	0.924
3	0.447	0.277	0.167	0.262	0.876	0.908	0.008	0.430	0.999	0.990	0.923	0.941
4	0.472	0.426	0.264	0.135	0.879	0.935	0.103	0.586	0.994	0.989	0.929	0.912
5	0.470	0.245	0.139	0.226	0.893	0.926	0.142	0.501	0.995	0.994	0.905	0.931
6	0.569	0.429	0.095	0.148	0.882	0.961	0.027	0.673	0.996	0.993	0.955	0.927
7	0.588	0.428	0.252	0.153	0.880	0.948	0.014	0.399	0.987	0.992	0.949	0.898
8	0.504	0.270	0.181	0.113	0.899	0.958	0.102	0.507	0.995	0.989	0.950	0.921
9	0.445	0.409	0.206	0.165	0.898	0.976	0.161	0.595	0.996	0.988	0.943	0.955
10	0.521	0.359	0.134	0.234	0.901	0.968	0.026	0.616	0.981	0.990	0.970	0.929

	Eng-Afr				SQUAD				GSM8K			
Ep	**BART**	**T5**	**Llma**	**Gmma**	**BART**	**T5**	**Llma**	**Gmma**	**BART**	**T5**	**Llma**	**Gmma**
2	0.470	0.203	0.126	0.000	0.744	0.776	0.130	0.422	0.994	0.998	0.000	0.805
3	0.356	0.462	0.000	0.173	0.862	0.836	0.000	0.372	0.994	0.992	0.000	0.927
4	0.445	0.591	0.000	0.408	0.878	0.854	0.000	0.547	0.990	0.997	0.000	0.968
5	0.642	0.420	0.000	0.130	0.912	0.906	0.000	0.447	0.985	0.992	0.000	0.918
6	0.477	0.588	0.000	0.163	0.883	0.930	0.000	0.438	0.989	0.996	1.000	0.941
7	0.522	0.630	0.000	0.098	0.879	0.936	0.000	0.563	0.992	0.993	1.000	0.945
8	0.563	0.505	0.000	0.016	0.893	0.923	0.000	0.203	0.990	0.990	0.000	0.847
9	0.450	0.650	0.000	0.102	0.904	0.942	0.000	0.323	0.985	0.989	0.000	0.841
10	0.506	0.609	0.000	0.097	0.886	0.948	0.000	0.495	0.988	0.994	0.000	0.693

Table 11: Proportion of case 1 miscalibrated pairs where either the worse sample's output did not change, but the confidence score increased, or the better sample's output did not change, but the confidence score decreased, using average log probs (top) and variance in BLEU scores (bottom) as confidence scores

	Eng-Afr		SQUAD		GSM8K	
Epoch	**BART**	**Flan-T5**	**BART**	**Flan-T5**	**BART**	**Flan-T5**
1	0.005	-0.028	*0.208	0.043	*0.121	-0.086
2	0.057	-0.127	*0.165	*0.154	*0.206	-0.082
3	-0.004	-0.061	*0.179	*0.211	*0.227	-0.038
4	0.071	-0.183	*0.185	*0.138	*0.259	-0.006
5	0.062	-0.179	*0.165	0.079	*0.177	-0.053
6	0.107	-0.112	0.098	*0.127	*0.204	-0.014
7	0.104	*-0.226	0.105	0.094	*0.323	-0.021
8	0.100	-0.215	*0.122	0.107	*0.203	-0.031
9	0.009	-0.153	*0.125	0.110	*0.252	-0.013
10	0.013	-0.051	*0.160	*0.111	*0.188	-0.017

	Eng-Afr		SQUAD		GSM8K	
Epoch	**BART**	**Flan-T5**	**BART**	**Flan-T5**	**BART**	**Flan-T5**
1	0.098	-0.049	-0.070	*-0.166	-0.099	0.077
2	0.051	0.025	0.044	*-0.168	*-0.152	0.035
3	0.053	0.047	-0.096	*-0.205	*-0.133	0.067
4	0.140	-0.020	-0.083	*-0.187	*-0.157	0.089
5	0.032	-0.001	*-0.120	*-0.162	*-0.163	-0.012
6	0.173	0.012	-0.023	*-0.177	*-0.196	0.097
7	0.173	0.049	-0.068	*-0.160	*-0.148	0.059
8	0.066	0.037	-0.011	*-0.181	*-0.178	0.030
9	0.110	0.054	-0.106	*-0.180	*-0.117	-0.012
10	0.046	0.076	0.020	*-0.167	*-0.214	0.051

Table 12: PRR between confidence metric and the similarity of the sample to the training data, measured using the maximum cosine similarity between the last encoder hidden state embedding of the test set and a sample in the training set