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

Anonymous ACL submission

## Abstract

Uncertainty quantification is a set of techniques that measure confidence in language models. They can be used, for example, to detect hallucinations or alert users to review uncertain predictions. To be useful, these confidence scores must be correlated with the quality of the output. However, recent work found that fine-tuning can affect the correlation between confidence scores and quality. Hence, we investigate the underlying behavior of confidence scores to understand its sensitivity to supervised fine-tuning (SFT). We find that post-SFT, the correlation of various confidence scores degrades, which can stem from changes in confidence scores due to factors other than the output quality, such as the output's similarity to the training distribution. We demonstrate via a case study how failing to address this miscorrelation reduces the usefulness of the confidence scores on a downstream task. Our findings show how confidence metrics cannot be used off-the-shelf without testing, and motivate the need for developing metrics which are more robust to fine-tuning.

## 1 Introduction

Uncertainty quantification (UQ) is a set of techniques for measuring the confidence of language models, which has increasingly been applied towards generation tasks. These confidence scores can be used to detect hallucinations (Wang et al., 2024; Manakul et al., 2023), select answers with higher quality (Wang et al., 2023), self-evaluate model outputs (Ren et al., 2023), and prompt users to review uncertain predictions (Xiao et al., 2020; Malinin and Gales, 2021; Liu et al., 2020; Kamath et al., 2020) (e.g. "Outputs with average token probability[1] from 0.20–0.25 often have low BLEU scores (range: 5–10); review before proceeding").

[1]Average token probability is an example of a confidence score used in generation tasks (Murray and Chiang, 2018; Zablotskaia et al., 2023)

To be useful in these applications, the scores generated by UQ techniques (i.e. confidence scores), must be correlated with the quality of the output. Hence, an important question is how correlated these metrics are with quality, especially when used off-the-shelf with models fine-tuned in different ways. This is important as various UQ metrics and packages are designed to allow any white-box model to be plugged in (Shelmanov et al., 2025). However, recent work showed how fine-tuning can make models overconfident (Rathi et al., 2025; Leng et al., 2025), thus making confidence scores *miscorrelated* with quality. This challenges whether these UQ metrics retain their correlation to quality when used with *any* fine-tuned model.

Hence, we investigate the sensitivity of the correlation between confidence scores and quality to SFT. We study probability-based and self-consistency based confidence scores which only rely on the model's token-level probabilities or outputs. These are assumed to approximate models' output quality since language models' output probabilities are trained to maximize the expected utility of one output (Wang and Holmes, 2024).

In our paper, we first fine-tune models on NLG tasks, and find that the correlation of various scores changes significantly after SFT, and varies with the number of samples and epochs used. Across 144 task-model-UQ metric configurations, correlation degrades in a third of the cases post-SFT (48/144), showing that these metrics need to be thoroughly checked before deployment.

Because correlation relies on sample-level confidence and evaluation scores being aligned, we study sample-level changes in these scores to understand how miscorrelation 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 miscorrelation occurs is not always the same, and different

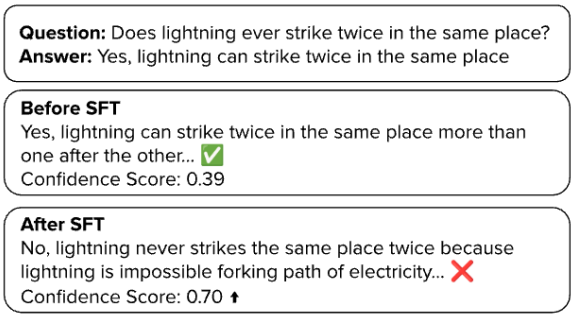



Figure 1: Before SFT, a model had a relatively low confidence score in a correct answer, whereas post-SFT, it had a much higher confidence in an incorrect answer, demonstrating a case of *relative overconfidence*

interventions may be needed for each metric. We also find that confidence scores are affected by factors beyond the output's quality such as the output's similarity to the training distribution, which must be accounted for to isolate the relationship with quality.

To illustrate the consequences of miscorrelation on a downstream task, 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 47% cases post-SFT. Thus, the usefulness of these metrics can suffer if miscorrelation is not addressed.

To summarize, we ask: **RQ1: How sensitive is the correlation of confidence metrics to different SFT parameters? RQ2: What are the underlying dynamics of confidence scores that negatively impact correlation with quality?** We find that the correlation of various confidence metrics change drastically post-SFT, with different types of miscorrelation 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 (UQ) encompasses a broad set of methods for identifying a model's confidence in its answers. These include fine-tuning additional models to estimate model confidence (Yaldiz et al., 2024; Kamath et al., 2020; Malinin et al., 2019; Fathullah et al., 2023), or detecting out-of-distribution (OOD) samples, for which a model's confidence is assumed to be lower (Liu et al., 2020; Vazhentsev et al., 2023). Other work studies having models verbalize a confidence score (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.

In this study, we focus on two types of methods: **Probability-based methods** use the model's output token probabilities to compute confidence (Murray and Chiang, 2018; Zablotskaia et al., 2023; Zhao et al., 2020; Perlitz et al., 2023; Kumar and Sarawagi, 2019; Huang et al., 2023; Malinin and Gales, 2021; Flores et al., 2025). These can be augmented by other models to take into account the similarity of model's outputs (Lin et al., 2023; Kuhn et al., 2023; Nikitin et al., 2024). **Self-consistency methods** obtain multiple answers from models (e.g., through dropout, beam search), and measure how similar (i.e. self-consistent) the model's predictions are: self-consistency across the top answers indicates confidence while variance indicates uncertainty (Xiao et al., 2020; Schmidt et al., 2022; Lakshminarayanan et al., 2017). Recently, methods combined probability and self-consistency methods into one metric (Vashurin et al., 2025). We study these methods as they can be used with any language model, and do not require additional fine-tuning or specialized modules.

**Evaluating UQ Metrics** To be practically useful, these scores must be associated with the quality of the output, which has been explored further for NLG recently (Daheim et al., 2025; Kotelevskii et al., 2025). 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). When the answer is scored numerically, like in many NLG tasks, previous analyses used Spearman correlation between the confidence score and and the quality of the output (Zablotskaia et al., 2023; Malinin and Gales, 2021) or Prediction-Rejection Ratio (Vashurin et al., 2025). Note that quality is defined differently in each task, and encompasses various evaluation metrics. In our work, we use Spearman correlation, and discuss its limitations at the end.

## 3 Method

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

analyses in the following sections. Then in Section 4, we study how sensitive this correlation 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 miscorrelation 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 **task-level correlation**, which measures how well the confidence score positively correlates with the outputs' quality*[2] [3].

$$\text{Correlation} = \rho(\text{Confidence}(X, \hat{Y}), \text{Quality}(\hat{Y}))$$

We measure task-level correlation using the Spearman correlation $\rho$ following Zablotskaia et al. (2023); Malinin and Gales (2021), which captures the notion that higher confidence should be associated with higher quality.

**Definition 3.3.** *We say a confidence score is **miscorrelated** if task-level correlation 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 **(BS Imp Wt)** (Eq 2) (Malinin and Gales, 2021), (5) the ratio between the joint sequence probability of the top beam and $k$-th beam **(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, 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 (IT) (Team et al., 2024).

## 4 RQ1: How sensitive is the correlation of confidence metrics to SFT parameters?

### 4.1 Experimental Design

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

We measure the effect of SFT on task-level correlation, by generating the outputs and the confidence scores both pre and post-SFT, and computing the change in Spearman correlation. 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 correlation of confidence metrics varies widely before and after SFT** We start by measuring the correlation of pre-trained models without

[2] Quality is measured by evaluating the output $\hat{Y}$, and may use reference $Y$ or input $X$

[3] $X$ and $Y$ refer to input and outputs as a variable, whereas $x$ and $y$ refer to an individual input/output

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.

As shown in Table 1, there can be substantial differences in correlation of confidence scores before and after SFT. Out of 144 scores-dataset-model configurations (3 datasets $\times$ 4 models $\times$ 12 confidence metrics), miscorrelation occurs post-SFT in 48 cases (33.3%), with 2 cases being statistically significant (one-way ANOVA, with Holm correction, $\alpha = 0.05$). Conversely, it improves in 96 cases (66.7%), with 8 being statistically significant.

Running the comparisons between the first and final epoch, we still see that there are more cases in which SFT improves correlation (85 cases, 59.0%), though these are not statistically significant (Table 6). It should be noted that because we could only run 3 seeds, there may be small, but undetected differences. Regardless, the changes in correlation 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 correlation** While the previous result studies correlation pre/post fine-tuning, we run an experiment to understand how correlation changes epoch-by-epoch. We fine-tune models for 10 epochs, measuring the correlation after each epoch, and find that the correlation can decline from one epoch to the next by at most 0.391 points (See Table 7).

Given that correlation 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 correlation. Moreover, the confidence score with the best correlation may differ based on how many epochs are used. For example in Figure 2, Beam Score Ratios (red line) performs well for the first five epochs, and is overtaken by Dropout Entropy from epoch six onwards. Hence, the choice of confidence score and number of training epochs have to be considered together.

**Similarly, the number of samples used considerably affects correlation** We vary the number of samples used ($n = 100, 500, 1000, 2000$). Fixing the number of epochs, we measure the correlation of the models across three seeds. Overall, there are significant differences in the correlation of confidence metrics when using different numbers of

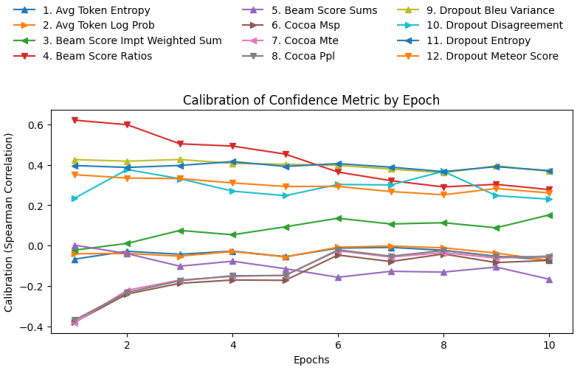



Figure 2: The correlation of various confidence metrics varies with more fine-tuning epochs; Plot shown for 12 confidence metrics, fine-tuned with Flan-T5 on SQUAD

samples (See Fig 3, and Fig 6 for full comparison). Across the 144 confidence metric-model-task configurations, we observe significant differences by number of samples in 13 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.

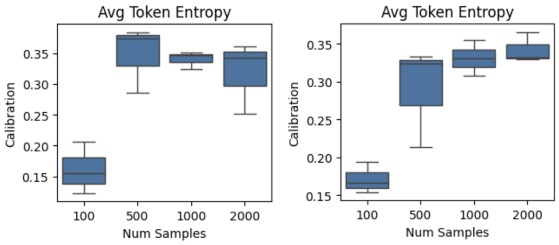



Figure 3: The correlation of confidence metrics differs significantly depending on the number of fine-tuning samples used, both before (left) and after (right) controlling for the number of fine-tuning steps used; Plots shown using average token entropy as the confidence metric, used on SQUAD for BART-Base

The previous results are confounded 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 run an experiment where fix the number of training steps to 2,500, and vary the number of samples. We see in Figure 6b there are still differences in correlation; 5/144 confidence metric-model-task configurations had significant differences between different numbers of samples (one-way ANOVA with Holm correction, $\alpha = 0.05$). In summary, correlation of confidence metrics is sensitive to SFT, and to the number of epochs and samples used.

| | BART-Base | | | Flan-T5-Base | | |
|---|---|---|---|---|---|---|
| | **Eng-Afr** | **SQUAD** | **GSM8K** | **Eng-Afr** | **SQUAD** | **GSM8K** |
| Avg Tok Prob | 0.278 (0.033) | 0.281 (0.015) | 0.024 (0.037) | 0.261 (0.249) | -0.074 (0.034) | 0.006 (0.027) |
| Avg Tok Ent | 0.269 (0.037) | 0.331 (0.017) | 0.023 (0.029) | 0.220 (0.259) | -0.061 (0.048) | 0.003 (0.017) |
| DO Ent | 0.063 (0.138) | 0.181 (0.118) | -0.021 (0.036) | 0.518 (0.081) | 0.083 (0.036) | 0.038 (0.043) |
| BS Imp Wt | -0.213 (0.069) | 0.228 (0.030) | -0.001 (0.033) | -0.425 (0.269) | 0.182 (0.015) | 0.023 (0.014) |
| BS Ratios | 0.150 (0.046) | 0.130 (0.048) | 0.005 (0.046) | 0.019 (0.184) | -0.351 (0.057) | 0.056 (0.052) |
| BS Sums | 0.210 (0.067) | -0.333 (0.049) | 0.000 (0.033) | 0.425 (0.269) | -0.177 (0.015) | -0.023 (0.014) |
| DO Bleu Var | 0.290 (0.142) | 0.146 (0.084) | 0.006 (0.052) | 0.003 (0.027) | 0.038 (0.030) | 0.021 (0.038) |
| DO KL Div | 0.068 (0.063) | 0.200 (0.018) | 0.013 (0.011) | 0.170 (0.074) | -0.130 (0.069) | 0.012 (0.065) |
| DO Meteor Var | 0.303 (0.129) | 0.149 (0.071) | 0.015 (0.013) | 0.027 (0.058) | -0.085 (0.036) | 0.016 (0.016) |
| CoCoA MSP | 0.063 (0.083) | 0.221 (0.034) | 0.002 (0.028) | -0.179 (0.116) | 0.078 (0.064) | 0.006 (0.028) |
| CoCoA MTE | 0.106 (0.068) | 0.404 (0.033) | 0.005 (0.022) | -0.652 (0.228) | 0.095 (0.078) | -0.001 (0.019) |
| CoCoA PPL | 0.127 (0.061) | 0.375 (0.031) | 0.006 (0.026) | -0.564 (0.226) | 0.097 (0.076) | 0.001 (0.024) |
| | **Llama 3.1-8B** | | | **Gemma 2-2B** | | |
| | **Eng-Afr** | **SQUAD** | **GSM8K** | **Eng-Afr** | **SQUAD** | **GSM8K** |
| Avg Tok Prob | 0.124 (0.129) | -0.343 (0.142) | 0.274 (0.103) | 0.327 (0.132) | -0.249 (0.221) | 0.155 (0.075) |
| Avg Tok Ent | 0.108 (0.114) | -0.330 (0.155) | 0.258 (0.091) | 0.357 (0.127) | -0.230 (0.215) | 0.135 (0.062) |
| DO Ent | 0.128 (0.084) | -0.147 (0.144) | -0.032 (0.068) | 0.192 (0.120) | -0.271 (0.062) | -0.017 (0.083) |
| BS Imp Wt | -0.111 (0.109) | 0.323 (0.189) | -0.264 (0.102) | -0.393 (0.121) | 0.379 (0.137) | -0.238 (0.081) |
| BS Ratios | 0.161 (0.205) | 0.290 (0.182) | 0.204 (0.097) | -0.081 (0.068) | 0.232 (0.182) | 0.255 (0.100) |
| BS Sums | 0.109 (0.109) | -0.328 (0.187) | 0.264 (0.102) | 0.392 (0.121) | -0.418 (0.138) | 0.232 (0.083) |
| DO BLEU Var | 0.015 (0.094) | 0.060 (0.247) | 0.0 (0.0) | 0.048 (0.119) | -0.019 (0.070) | -0.044 (0.039) |
| DO KL Div | -0.002 (0.049) | 0.250 (0.158) | -0.018 (0.066) | 0.003 (0.101) | 0.148 (0.025) | -0.025 (0.119) |
| DO Meteor Var | -0.067 (0.209) | -0.153 (0.077) | 0.092 (0.080) | 0.038 (0.172) | 0.062 (0.101) | -0.068 (0.107) |
| CoCoA MSP | -0.006 (0.011) | -0.216 (0.273) | 0.275 (0.119) | 0.203 (0.043) | -0.225 (0.266) | 0.192 (0.091) |
| CoCoA MTE | -0.044 (0.022) | -0.256 (0.192) | 0.262 (0.109) | 0.197 (0.127) | -0.245 (0.212) | 0.091 (0.041) |
| CoCoA PPL | -0.006 (0.011) | -0.272 (0.175) | 0.275 (0.119) | 0.206 (0.127) | -0.264 (0.225) | 0.113 (0.063) |

Table 1: Difference in Spearman correlation before SFT and after performing SFT with early stopping; Averaged across 3 seeds with st. dev in parentheses (see Section 3.1 for abbreviations)

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

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

**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).*

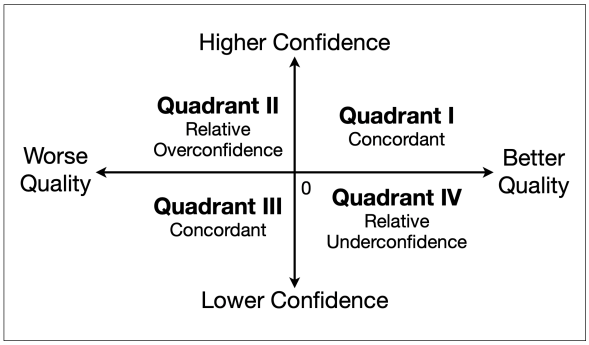



Figure 4: We classify SFT's effect on the correlation of samples using four quadrants

**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 miscorrelation. To this end, we analyze changes in the relative ranks of samples' quality and confidence scores across epochs. We first define relative correlation and miscorrelation at the sample level.

**Definition 5.3.** *For two samples $(A, B)$, a confidence score is **relatively correlated** 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 miscorrelated** 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 miscorrelation. Formally, fine-tuning causes miscorrelation for pair $(A, B)$ if the pair was relatively correlated at epoch $t$, and relatively miscorrelated at epoch $t + 1$.

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

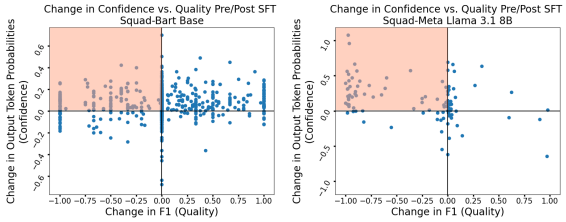
- **Case 1**: The relative ranking of quality scores stays the same, but the relative confidence scores flip

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

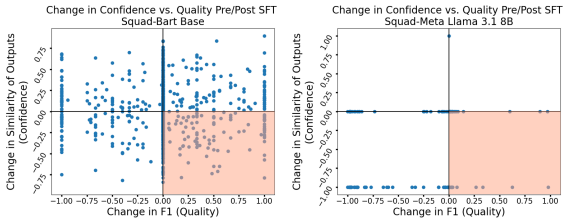
- **Case 2:** The relative ranking of quality scores change, but the confidence scores' rankings stay the same

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

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




(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

## 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, and Gemma-2 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).

| | 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 | Eng-Afr | 0.54 | 0.26 | 0.20 |
| | SQUAD | 0.63 | 0.27 | 0.10 |
| | GSM8K | 0.87 | 0.13 | 0.00 |

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

To understand the changes in confidence score better, we plot each sample's average log proba-

bilities by epoch. One observation is that for both BART, Llama 3.1, and Gemma-2, 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 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 correlation between confidence and quality becomes worse.

| 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** | Eng-Afr | 46.6 | 23.9 | 29.5 |
| | SQUAD | 46.7 | 44.9 | 8.3 |
| | GSM8K | 46.8 | 35.4 | 17.8 |
| **Gemma-2** | Eng-Afr | 48.1 | 22.0 | 29.9 |
| | SQUAD | 43.9 | 37.3 | 18.8 |
| | GSM8K | 40.3 | 51.8 | 7.9 |

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

**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 miscorrelation is more acceptable.

**Case 1 relative miscorrelation is more prevalent** After studying the dynamics of confidence metrics pre/post SFT, we study how these changes result in miscorrelation, to identify where interventions can be made. Using average log probs to measure confidence, we find that case 1 miscorrelation happens more than case 2 in 10/12 dataset-model configurations. On average, case 1 comprises 68.5% of miscorrelated pairs across the twelve dataset-model configurations (See Table 4). When using variance in BLEU scores, case 1 is more prevalent in 6/12 dataset-model configurations, accounting for 52.1% of miscorrelated pairs (See Table 10).

| | 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 |
| 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 | 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 | 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 |

Table 4: Case 1 miscorrelation 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

**Case 1 miscorrelation 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 miscorrelated 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 miscorrelated cases (See Table 11).

Our takeaway here is that in many cases, miscorrelation 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 correlation 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 correlations 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 miscorrelation affect downstream tasks?

To illustrate *why* we need to address miscorrelation, we perform a case study, where we explore an application of confidence scores on a downstream task of identifying when an answer is correct.

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[4].

[4]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

| | BART | | Flan-T5 | | Llama-3.1 8B | | Gemma 2-2B | |
|---|---|---|---|---|---|---|---|---|
| | 1 Ep | Full | Pre | Full | Pre | Full | Pre | Full |
| Tok Prob | 0.563 | 0.639 ↑ | 0.545 | 0.610 ↑ | 0.532 | 0.465 ↓ | 0.441 | 0.467 ↑ |
| Tok Ent | 0.454 | 0.330 ↓ | 0.492 | 0.368 ↓ | 0.458 | 0.512 ↑ | 0.568 | 0.519 ↓ |
| DO Ent | 0.378 | 0.463 ↑ | 0.068 | 0.359 ↑ | 0.355 | 0.453 ↑ | 0.601 | 0.524 ↓ |
| BS ImpWt | 0.314 | 0.437 ↑ | 0.511 | 0.349 ↓ | 0.506 | 0.492 ↓ | 0.562 | 0.518 ↓ |
| BS Rat | 0.542 | 0.560 ↑ | 0.609 | 0.713 ↑ | 0.593 | 0.409 ↓ | 0.536 | 0.486 ↓ |
| BS Sums | 0.670 | 0.542 ↓ | 0.486 | 0.641 ↑ | 0.494 | 0.508 ↑ | 0.438 | 0.479 ↑ |
| DO BLEU | 0.411 | 0.381 ↓ | 0.348 | 0.533 ↑ | 0.554 | 0.516 ↓ | 0.518 | 0.477 ↓ |
| DO KL | 0.583 | 0.489 ↓ | 0.185 | 0.461 ↑ | 0.589 | 0.470 ↓ | 0.498 | 0.575 ↑ |
| DO METr | 0.580 | 0.621 ↑ | 0.920 | 0.512 ↓ | 0.465 | 0.511 ↑ | 0.507 | 0.524 ↑ |
| CCo MSP | 0.446 | 0.226 ↓ | 0.606 | 0.453 ↓ | 0.329 | 0.492 ↑ | 0.457 | 0.541 ↑ |
| CCo MTE | 0.429 | 0.203 ↓ | 0.671 | 0.492 ↓ | 0.335 | 0.472 ↑ | 0.447 | 0.510 ↑ |
| CCo PPL | 0.403 | 0.207 ↓ | 0.606 | 0.495 ↓ | 0.371 | 0.484 ↑ | 0.451 | 0.519 ↑ |

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

We evaluate the AUROC in detecting the right answers. Pre/post SFT, correlation deteriorates after SFT in 23/48 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 correlation is robust to SFT.

## 7 Conclusion

In our work, we study how the correlation of various confidence metrics change with supervised fine-tuning (SFT). First, we find that the correlation of these confidence scores can vary widely based on different choices in the SFT pipeline. Hence, these metrics cannot be used off-the-shelf, and testing is required before using them in practical applications. Second, we find that SFT is prone to relative overconfidence using probability-based metrics, and relative underconfidence with consistency-based metrics, and that miscorrelation often occurs when models change their confidence scores, without the actual quality of the output changing. This suggests that probability and consistency-based confidence scores are affected by other factors unrelated to the output quality during SFT, that need to be accounted for, to ensure they still correlate with the evaluation metric. Finally, we test confidence metrics on a downstream task of identifying correct answers in a QA dataset, and find that the performance of various confidence metrics decrease post-SFT. Given this, future work can explore methods to improve the robustness of confidence metrics to SFT. Moreover, new confidence metrics can be designed that account for factors such as the sample's similarity to the training data.

## Limitations and Potential Risks

We acknowledge that our study was relatively narrow, as some of our analyses only used average log probs and variance in BLEU scores to represent probability-based and consistency-based metrics. Hence, more work can be done to extend the analyses to other metrics. In the second part of the analysis, we also only focus our analyses on the first type of miscorrelation, which can also be expanded. Finally, we only tested on three tasks: translation, question answering, and math. While we believe that correlation of confidence metrics to quality will also be sensitive to SFT on other tasks, it remains to be verified experimentally.

We also acknowledge that using Spearman correlation to measure correlation is also imperfect, as it assumes that evaluation metrics across samples can be compared to one another. Recent work uses the Prediction-Rejection Ratio to measure correlation (Vashurin et al., 2025), however it makes similar assumptions. While we acknowledge that we are working with faulty metrics to measure correlation, more work must also be done to improve the measurement of correlation in generation tasks.

## 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, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj

Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

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.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models.

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.

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

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.

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

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.

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.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction.

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.

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

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, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

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.

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.

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

Ziyu Wang and Chris Holmes. 2024. On subjective uncertainty quantification and calibration in natural language generation.

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

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 Correlation Results

We report the maximum and minimum correlation by model and task, across the various confidence metrics in Tables 6, 7, and 8. We show the correlation 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 Miscorrelation Analyses

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

## 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.

| | BART-Base | | | Flan-T5-Base | | |
|---|---|---|---|---|---|---|
| | **Eng-Afr** | **SQUAD** | **GSM8K** | **Eng-Afr** | **SQUAD** | **GSM8K** |
| Avg Tok Prob | 0.023 (0.089) | -0.019 (0.051) | 0.04 (0.05) | 0.005 (0.169) | 0.01 (0.074) | -0.006 (0.043) |
| Avg Tok Ent | 0.058 (0.094) | -0.005 (0.041) | 0.03 (0.054) | -0.051 (0.171) | 0.017 (0.078) | -0.01 (0.035) |
| DO Ent | -0.015 (0.156) | -0.055 (0.04) | 0.001 (0.015) | 0.148 (0.109) | 0.008 (0.035) | 0.002 (0.039) |
| BS Imp Wt | -0.013 (0.096) | 0.21 (0.084) | 0.046 (0.021) | -0.183 (0.229) | 0.089 (0.024) | -0.007 (0.039) |
| BS Ratios | -0.099 (0.032) | 0.093 (0.055) | 0.024 (0.011) | 0.434 (0.308) | -0.26 (0.101) | 0.042 (0.037) |
| BS Sums | 0.012 (0.096) | -0.2 (0.094) | -0.047 (0.022) | 0.183 (0.229) | -0.09 (0.022) | 0.007 (0.039) |
| DO BLEU Var | -0.083 (0.052) | 0.035 (0.058) | 0.014 (0.093) | 0.077 (0.091) | -0.048 (0.004) | 0.011 (0.055) |
| DO KL Div | 0.092 (0.083) | 0.102 (0.032) | -0.003 (0.031) | 0.035 (0.186) | -0.082 (0.052) | -0.005 (0.068) |
| DO Meteor Var | -0.105 (0.086) | 0.036 (0.071) | 0.006 (0.044) | 0.063 (0.177) | -0.068 (0.016) | 0.009 (0.026) |
| CoCoA MSP | 0.026 (0.109) | 0.029 (0.035) | -0.004 (0.048) | 0.114 (0.138) | 0.281 (0.076) | -0.0 (0.043) |
| CoCoA MTE | 0.062 (0.091) | 0.034 (0.033) | 0.01 (0.034) | -0.241 (0.275) | 0.295 (0.103) | 0.003 (0.046) |
| CoCoA PPL | 0.037 (0.098) | 0.025 (0.036) | 0.017 (0.032) | -0.205 (0.248) | 0.301 (0.092) | 0.005 (0.051) |
| | **Llama 3.1-8B** | | | **Gemma 2-2B** | | |
| | **Eng-Afr** | **SQUAD** | **GSM8K** | **Eng-Afr** | **SQUAD** | **GSM8K** |
| Avg Tok Prob | 0.257 (0.148) | 0.106 (0.087) | 0.152 (0.126) | 0.511 (0.264) | 0.037 (0.254) | -0.008 (0.145) |
| Avg Tok Ent | 0.238 (0.145) | 0.107 (0.11) | 0.144 (0.142) | 0.483 (0.247) | 0.036 (0.285) | -0.039 (0.121) |
| DO Ent | 0.168 (0.098) | 0.17 (0.029) | -0.04 (0.061) | 0.044 (0.098) | -0.1 (0.145) | -0.039 (0.147) |
| BS Imp Wt | -0.289 (0.126) | -0.185 (0.125) | -0.139 (0.127) | -0.568 (0.257) | 0.289 (0.039) | -0.02 (0.203) |
| BS Ratios | -0.001 (0.083) | -0.252 (0.138) | 0.217 (0.044) | -0.071 (0.245) | -0.015 (0.064) | 0.087 (0.142) |
| BS Sums | 0.287 (0.126) | 0.198 (0.118) | 0.138 (0.129) | 0.567 (0.257) | -0.284 (0.039) | 0.015 (0.202) |
| DO BLEU Var | -0.026 (0.075) | 0.014 (0.25) | -0.019 (0.032) | 0.032 (0.174) | -0.051 (0.021) | -0.034 (0.031) |
| DO KL Div | -0.114 (0.045) | 0.053 (0.208) | -0.031 (0.074) | -0.062 (0.059) | 0.078 (0.161) | -0.004 (0.338) |
| DO Meteor Var | -0.115 (0.211) | -0.033 (0.046) | 0.059 (0.148) | -0.001 (0.192) | -0.038 (0.031) | -0.056 (0.022) |
| CoCoA MSP | -0.014 (0.164) | -0.41 (0.201) | 0.12 (0.13) | 0.461 (0.082) | -0.087 (0.18) | 0.009 (0.152) |
| CoCoA MTE | 0.152 (0.216) | 0.056 (0.147) | 0.116 (0.143) | 0.597 (0.231) | -0.051 (0.166) | 0.01 (0.136) |
| CoCoA PPL | 0.184 (0.199) | 0.059 (0.126) | 0.12 (0.13) | 0.605 (0.243) | -0.05 (0.176) | 0.034 (0.147) |

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

## 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 | | |
|---|---|---|---|---|---|---|
| | **Eng-Afr** | **SQUAD** | **GSM8K** | **Eng-Afr** | **SQUAD** | **GSM8K** |
| Avg Tok Prob | -0.096 (0.015) | -0.083 (0.025) | -0.068 (0.013) | -0.288 (0.129) | -0.066 (0.028) | -0.064 (0.028) |
| Avg Tok Ent | -0.081 (0.007) | -0.083 (0.024) | -0.062 (0.012) | -0.294 (0.135) | -0.061 (0.01) | -0.063 (0.03) |
| DO Ent | -0.127 (0.046) | -0.161 (0.052) | -0.068 (0.019) | -0.203 (0.084) | -0.044 (0.008) | -0.049 (0.01) |
| BS Imp Wt | -0.074 (0.015) | -0.046 (0.026) | -0.059 (0.027) | -0.391 (0.274) | -0.032 (0.013) | -0.062 (0.007) |
| BS Ratios | -0.115 (0.022) | -0.041 (0.007) | -0.068 (0.004) | -0.228 (0.08) | -0.162 (0.01) | -0.081 (0.012) |
| BS Sums | -0.063 (0.002) | -0.124 (0.041) | -0.057 (0.018) | -0.281 (0.172) | -0.107 (0.057) | -0.067 (0.021) |
| DO BLEU Var | -0.116 (0.046) | -0.066 (0.015) | -0.059 (0.031) | -0.149 (0.004) | -0.059 (0.009) | -0.056 (0.02) |
| DO KL Div | -0.125 (0.062) | -0.065 (0.004) | -0.056 (0.015) | -0.295 (0.056) | -0.209 (0.086) | -0.095 (0.031) |
| DO Meteor Var | -0.109 (0.023) | -0.131 (0.091) | -0.073 (0.022) | -0.118 (0.014) | -0.058 (0.009) | -0.065 (0.04) |
| CoCoA MSP | -0.125 (0.058) | -0.075 (0.023) | -0.066 (0.018) | -0.176 (0.046) | -0.059 (0.012) | -0.076 (0.03) |
| CoCoA MTE | -0.113 (0.063) | -0.085 (0.025) | -0.069 (0.015) | -0.352 (0.075) | -0.061 (0.008) | -0.038 (0.008) |
| CoCoA PPL | -0.116 (0.064) | -0.077 (0.029) | -0.063 (0.02) | -0.336 (0.1) | -0.059 (0.016) | -0.033 (0.015) |
| | **Llama 3.1-8B** | | | **Gemma 2-2B** | | |
| | **Eng-Afr** | **SQUAD** | **GSM8K** | **Eng-Afr** | **SQUAD** | **GSM8K** |
| Avg Tok Prob | -0.245 (0.057) | -0.134 (0.056) | -0.148 (0.051) | -0.206 (0.007) | -0.212 (0.021) | -0.231 (0.087) |
| Avg Tok Ent | -0.257 (0.095) | -0.132 (0.045) | -0.14 (0.054) | -0.206 (0.034) | -0.225 (0.025) | -0.232 (0.07) |
| DO Ent | -0.256 (0.058) | -0.156 (0.071) | -0.188 (0.057) | -0.258 (0.169) | -0.158 (0.068) | -0.158 (0.08) |
| BS Imp Wt | -0.548 (0.167) | -0.173 (0.024) | -0.207 (0.079) | -0.473 (0.109) | -0.205 (0.054) | -0.263 (0.05) |
| BS Ratios | -0.228 (0.057) | -0.244 (0.041) | -0.144 (0.082) | -0.235 (0.02) | -0.212 (0.047) | -0.112 (0.06) |
| BS Sums | -0.223 (0.058) | -0.13 (0.09) | -0.173 (0.066) | -0.186 (0.032) | -0.275 (0.095) | -0.227 (0.013) |
| DO BLEU Var | -0.054 (0.257) | -0.285 (0.034) | -0.06 (0.076) | -0.24 (0.083) | -0.168 (0.034) | -0.159 (0.079) |
| DO KL Div | -0.202 (0.142) | -0.267 (0.196) | -0.194 (0.133) | -0.313 (0.048) | -0.173 (0.066) | -0.151 (0.038) |
| DO Meteor Var | -0.247 (0.106) | -0.214 (0.133) | -0.266 (0.051) | -0.23 (0.06) | -0.179 (0.068) | -0.235 (0.072) |
| CoCoA MSP | -0.265 (0.032) | -0.251 (0.076) | -0.197 (0.067) | -0.288 (0.147) | -0.219 (0.076) | -0.218 (0.131) |
| CoCoA MTE | -0.287 (0.003) | -0.155 (0.047) | -0.182 (0.067) | -0.26 (0.11) | -0.211 (0.077) | -0.252 (0.084) |
| CoCoA PPL | -0.265 (0.032) | -0.167 (0.064) | -0.197 (0.067) | -0.251 (0.119) | -0.187 (0.041) | -0.261 (0.069) |

Table 7: From epoch-to-epoch, there can be a considerable decline in correlation, reaching differences of up to 0.391 Spearman correlation points; Results averaged over 3 seeds with st. dev in parentheses

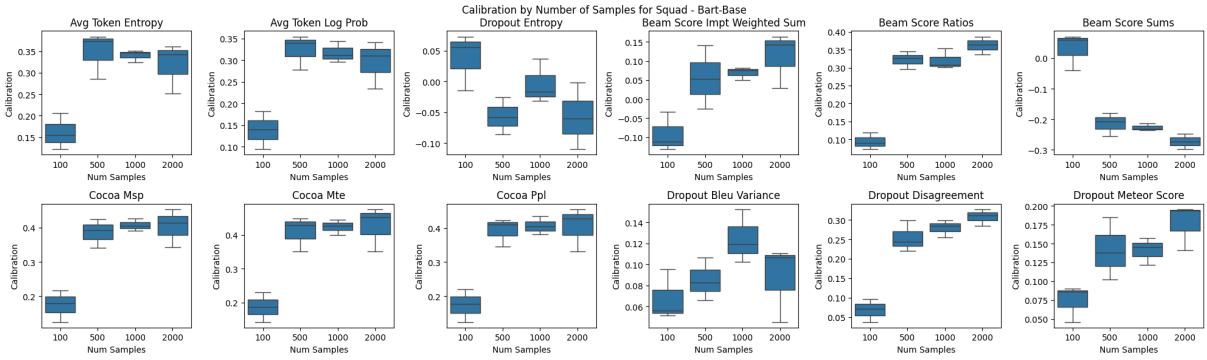
| | Eng-Afr | | | SQUAD | | | GSM8K | | |
|---|---|---|---|---|---|---|---|---|---|
| | **No SFT** | **1 Ep** | **Full** | **No SFT** | **1 Ep** | **Full** | **No SFT** | **1 Ep** | **Full** |
| Avg Tok Prob | 0.092 (0.014) | 0.347 (0.102) | 0.37 (0.047) | 0.064 (0.0) | 0.364 (0.045) | 0.345 (0.015) | 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) |

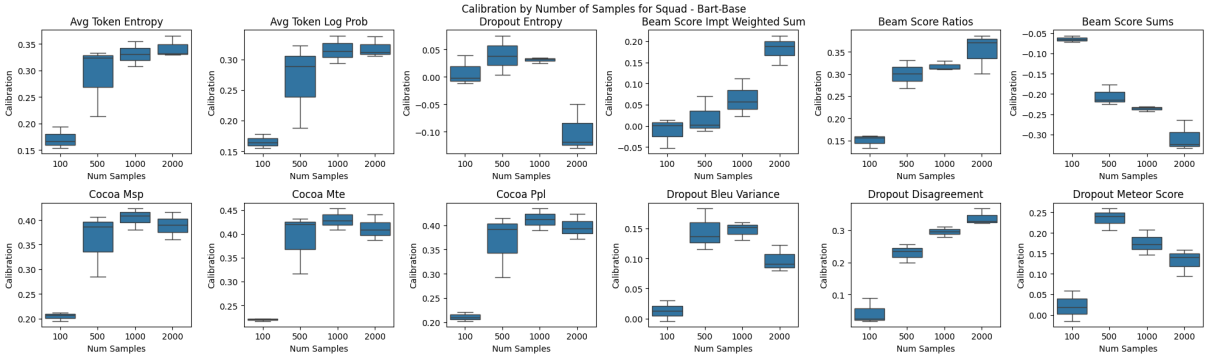
| | Eng-Afr | | | SQUAD | | | GSM8K | | |
|---|---|---|---|---|---|---|---|---|---|
| | **No SFT** | **1 Ep** | **Full** | **No SFT** | **1 Ep** | **Full** | **No SFT** | **1 Ep** | **Full** |
| Avg Tok Prob | 0.121 (0.039) | -0.012 (0.049) | 0.245 (0.099) | 0.068 (0.065) | -0.381 (0.041) | -0.275 (0.08) | 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) |

Table 8: Spearman correlation of confidence metrics before SFT, after one epoch, and after full SFT with early stopping for BART, Flan-T5, Llama 3.1-8B, and Gemma 2-2B (In descending order)




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




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

Figure 6: Spearman correlation 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.

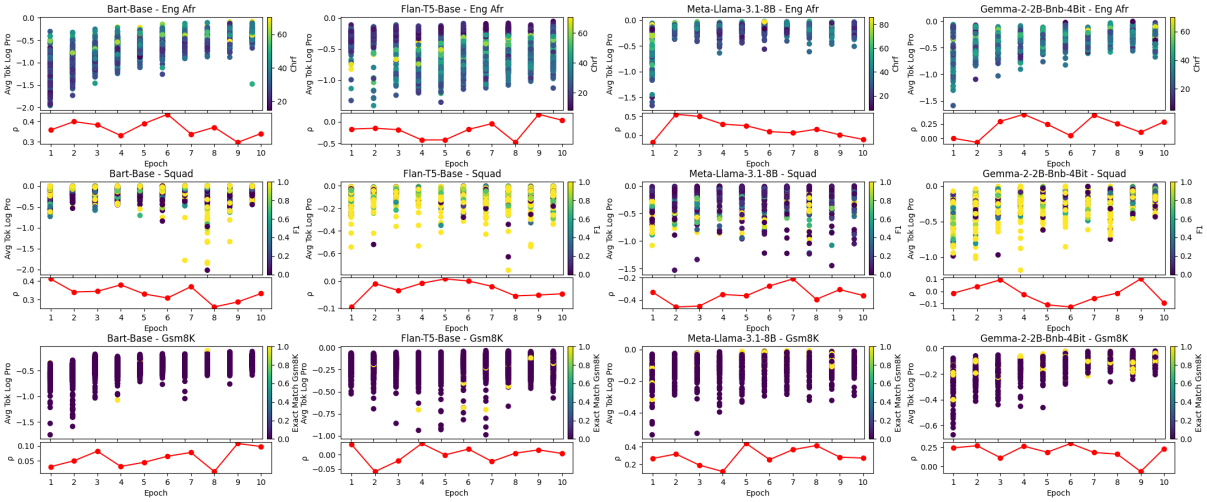



Figure 7: The average log probs across all test set samples generally increase for BART and Llama 3.1 8B (top), which may explain the rel. overconfidence in probability-based metrics, and fluctuations in correlation (bottom)

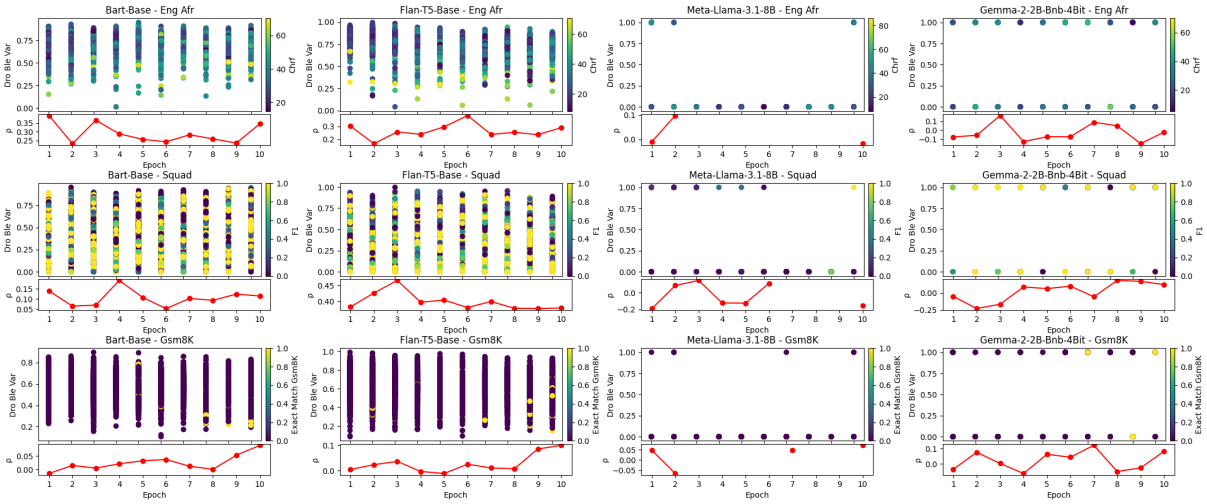



Figure 8: The dropout BLEU variance values generally do not change across epochs, which aligns with the fact that most samples remain calibrated post SFT.

| | 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** | Eng-Afr | 0.94 | 0.00 | 0.06 |
| | SQUAD | 0.92 | 0.02 | 0.06 |
| | GSM8K | 1.00 | 0.00 | 0.00 |
| **Gemma** | Eng-Afr | 0.73 | 0.01 | 0.26 |
| | SQUAD | 0.87 | 0.09 | 0.04 |
| | GSM8K | 0.92 | 0.01 | 0.07 |

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** | 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 |
| **Llama** | 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 |

Table 10: **Case 1** miscorrelation 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 miscorrelated 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: Spearman correlation 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