UNCERTAINTY-LINE: Length-Invariant Estimation of Uncertainty for Large Language Models

Roman Vashurin* **Maiya Goloburda*** **Preslav Nakov** **Maxim Panov**
Mohamed bin Zayed University of Artificial Intelligence
{Roman.Vashurin, Maiya.Goloburda, Preslav.Nakov, Maxim.Panov}@mbzuai.ac.ae

Abstract

Large Language Models (LLMs) have become indispensable tools across various applications, making it more important than ever to ensure the quality and the trustworthiness of their outputs. This has led to growing interest in uncertainty quantification (UQ) methods for assessing the reliability of LLM outputs. Many existing UQ techniques rely on token probabilities, which inadvertently introduces a bias with respect to the length of the output. While some methods attempt to account for this, we demonstrate that such biases persist even in length-normalized approaches. To address the problem, here we propose UNCERTAINTY-LINE (Length-INvariant Estimation), a simple debiasing procedure that regresses uncertainty scores on output length and uses the residuals as corrected, length-invariant estimates. Our method is post-hoc, model-agnostic, and applicable to a range of UQ measures. Through extensive evaluation on machine translation, summarization, and question-answering tasks, we demonstrate that UNCERTAINTY-LINE consistently improves over even nominally length-normalized UQ methods uncertainty estimates across multiple metrics and models. We release our code publicly.[1]

1 Introduction

Large Language Models (LLMs) have become essential in a wide range of applications. However, despite their impressive capabilities, LLMs can sometimes generate misleading or outright incorrect information. Given their widespread adoption in critical domains, ensuring the reliability of their responses has become a pressing concern. This has led to growing interest in uncertainty quantification (UQ) to measure the confidence in model-generated output (Gal, 2016; Hu et al., 2023; Kotelevskii et al., 2025).

Many existing UQ methods for LLMs rely on token-level probabilities produced by LLM itself. However, the log-probabilities of the generated tokens in autoregressive models are summed over the sequence, meaning that the total sequence score becomes increasingly negative as length increases, leading to unreliable estimates (Murray and Chiang, 2018; Braverman et al., 2020; Zhao et al., 2023). Some methods, such as perplexity and mean token entropy, have been proposed as length-normalized uncertainty measures (Fomicheva et al., 2020). Alternatively, some work has focused on calibrating model confidence scores using post-hoc methods or on reformulating uncertainty estimation at the token level (Ren et al., 2023; Zhao et al., 2023; Gupta et al., 2024). While this yields improvements, it often requires additional supervision, architectural changes, or tuning for specific tasks.

Here, we propose a simple and effective method for detrending uncertainty estimates with respect to output length, in both unsupervised and minimally supervised settings. It is post-hoc, model-agnostic, and applicable across a range of uncertainty measures. Our key contributions are as follows:

- We demonstrate that uncertainty estimation metrics exhibit length bias, even when length-normalization is applied; see Section 3.
- We propose Uncertainty-Length INvariant Estimation (UNCERTAINTY-LINE), a simple unsupervised detrending approach that fits a regression between uncertainty scores and output length, and uses the residuals as uncertainty estimates. We also formulate a supervised extension for cases where the output length correlates with quality; see Section 4.
- We evaluate our approach on machine translation, summarization, and mathematical reasoning tasks, showing improved performance of the uncertainty estimates; see Section 5.

* These authors contributed equally.
[1] https://github.com/stat-ml/uncertainty-line

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7881–7908
November 4-9, 2025

2 Related Work

Length Bias in Sequence Likelihood. Early work in sequence generation noted that sequence-level likelihood (the joint probability of an output) is biased with regards to the output length. Neural models often assign disproportionately low probabilities to longer outputs, causing shorter sequences to appear "more likely" (Murray and Chiang, 2018; Adiwardana et al., 2020). More recently, Santilli et al. (2025) demonstrate that such length effects also impact uncertainty evaluation, highlighting the need for explicit length bias removal.

Length-Normalized Uncertainty Measures. A common remedy is to use the average per token confidence score (i.e. normalize overall confidence by sequence length) instead of the raw sum. For example, *Perplexity* or mean log-probability per token is often used as a sequence-level confidence score instead of raw joint probability (Fomicheva et al., 2020). Another measure is *Mean Token Entropy*: the average entropy of the model's predictive distribution at each time step of the output (Fomicheva et al., 2020). On the other hand, *Monte Carlo Sequence Entropy* (MCSE) and its length-normalized variant *Monte Carlo Normalzied Sequence Entropy* (MCNSE) measure uncertainty by estimating the entropy of predictive distribution using multiple outputs sampled via stochastic decoding (Malinin and Gales, 2021; Kuhn et al., 2023). Lastly, TokenSAR provides a length-normalized measure that reweights token log-probabilities based on their importance to the meaning of an output (Duan et al., 2024). However, naively dividing by length can overcorrect: it can overly penalize shorter sequences, flipping the bias in the opposite direction (Gupta et al., 2024). These findings show that while length normalization is a useful tool, it needs careful consideration to avoid introducing a new bias.

Uncertainty Calibration. Recent work (Ren et al., 2023) shows that sequence-level confidence scores remain poorly calibrated with output quality, even after applying length normalization. To address this problem, various methods have been proposed, including token-level self-evaluation, sequence likelihood calibration, language model cascades that leverage token-level uncertainty and post-hoc correctors for uncertainty estimates (Ren et al., 2023; Zhao et al., 2023; Gupta et al., 2024; Li et al., 2025).

While such techniques improve the alignment between model confidence and human judgment, they often require additional supervision or task-specific tuning and, crucially, do not directly target the problem of length bias.

Non-token likelihood based uncertainty measures. Consistency uncertainty measures have emerged as a way to bypass token-level scores entirely (Fomicheva et al., 2020; Lin et al., 2024; Kuhn et al., 2023). By evaluating uncertainty as a level of agreement between sampled outputs, they provide a length-agnostic confidence estimate. However, while these measures are designed to be length-invariant, their practical application encounters several challenges. First, these methods rely on sampling multiple outputs from the language model to capture the distribution of possible generations. This sampling process is computationally intensive, especially for large models. Secondly, implementations often depend on pre-trained models, to assess semantic similarity between generated outputs. While effective for certain tasks, these models are frequently trained on shorter texts, leading to less reliable estimates for long generations.

Another approach that does not rely on token likelihoods is verbalized uncertainty, where models express their confidence in natural language. However, studies have shown that LLMs often exhibit overconfidence in their verbalized uncertainty assessments (Xiong et al., 2024). This overconfidence suggests that without proper calibration or fine-tuning, verbalized uncertainty may not reliably reflect true predictive uncertainty (Liu et al., 2024). Therefore, while verbalized uncertainty offers a promising, length-invariant alternative, its practical utility is limited unless accompanied by effective calibration strategies.

These works highlight the limitations of existing UQ methods, including length-normalized measures in addressing the length bias problem.

3 Length Bias in Uncertainty Measures and Quality Metrics

We start by quantifying the degree to which uncertainty quantification (UQ) measures applied to the output of Llama 3.1 8B (Grattafiori et al., 2024) model exhibit dependency on the length of generated sequences. We do that on a comprehensive set of tasks that imply varying length of the generated response: neural machine translation (NMT), mathematical reasoning, and abstractive summarization.

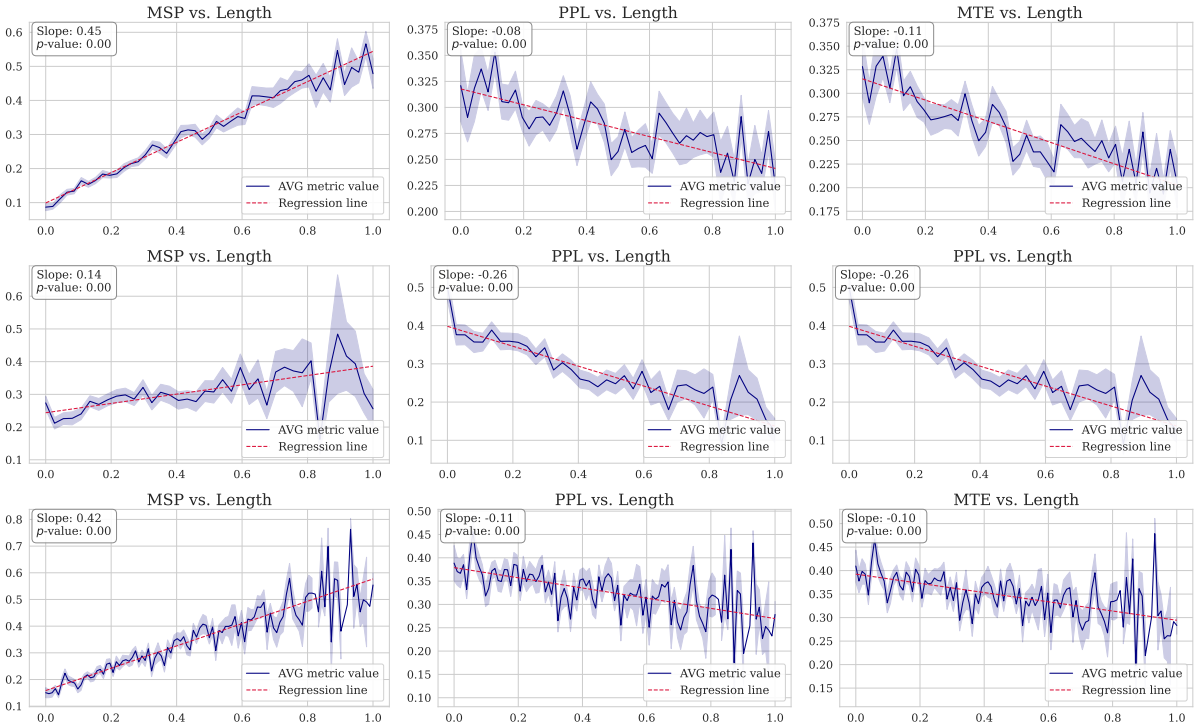

Figure 1: Trends in UQ scores with respect to normalized sequence length for WMT14 De-En, GSM8K and XSum (Model – Llama 3.1 8B). Each subplot shows a linear regression fit over binned scores. The **slope** indicates the strength and direction of correlation, **p-value** reflects statistical significance.

For NMT, we used four language pairs from WMT14 and WMT19, GSM8K for mathematical reasoning, and XSum for ATS (Bojar et al., 2014; Barrault et al., 2019; Cobbe et al., 2021; Narayan et al., 2018). We performed estimation on the random subset of 2000 points from each dataset.

The level of dependency is expressed as the slope of a linear regression fit by ordinary least squares (OLS), with the UQ values as the response variable and the generated sequence length as a predictor. The significance of the obtained linear trend was assessed using the Wald test and corresponding p-value was calculated along with the slope.

Uncertainty Measures are Strongly Length-Dependent. Figure 1 presents results on one of the machine translation datasets (WMT14 De-En), XSum and GSM8K. The UQ measures show clear and significant trends. For the Maximum Sequence Probability (MSP), the average UQ score increases with sequence length, indicating that longer generations are assigned lower model confidence, which is potentially misleading, as longer outputs may simply reflect more confident token-level predictions. Both Perplexity (PPL) and Mean Token Entropy (MTE) exhibit the opposite trend, with average uncertainty decreasing as length increases.

This is notable, as both measures were designed to normalize for length-related effects, yet the trends persist. In the majority of cases, the p-values of regression coefficients are below 0.05, confirming that the observed relationships are statistically significant. This highlights a key concern: although these UQ measures are widely used, they may conflate uncertainty with sequence length in practice.

Results for the rest of the datasets, models and UQ methods can be found in Appendix A.2.

Quality Metrics for Machine Translation are Largely Length-Agnostic. Performance of the UQ method is largely defined by the extent of its correlation with some meaningful measure of prediction quality. Thus, having assessed relationship of UQ with generation length, we perform a similar analysis of the behavior of several quality metrics for the same selection of tasks.

For NMT, Figure 2 shows Comet WMT 22, XComet XXL and Metric X XXL scores over normalized generation lengths (Rei et al., 2022; Guerreiro et al., 2024; Juraska et al., 2024). Across datasets, the fitted linear regression lines exhibit near-zero slopes, and the associated p-values exceed the standard 0.05 threshold in the majority of cases.

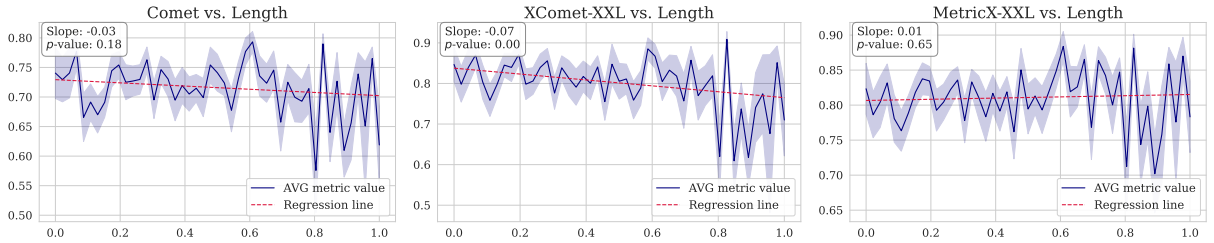

Figure 2: Trends for Comet WMT 22, XComet XXL and Metric X XXL scores with respect to normalized sequence length for WMT14 De-En machine translation dataset (Model - Llama 3.1 8B). Each subplot shows a linear regression fit over binned scores. The **slope** indicates the strength and direction of correlation, **p-value** reflects statistical significance.

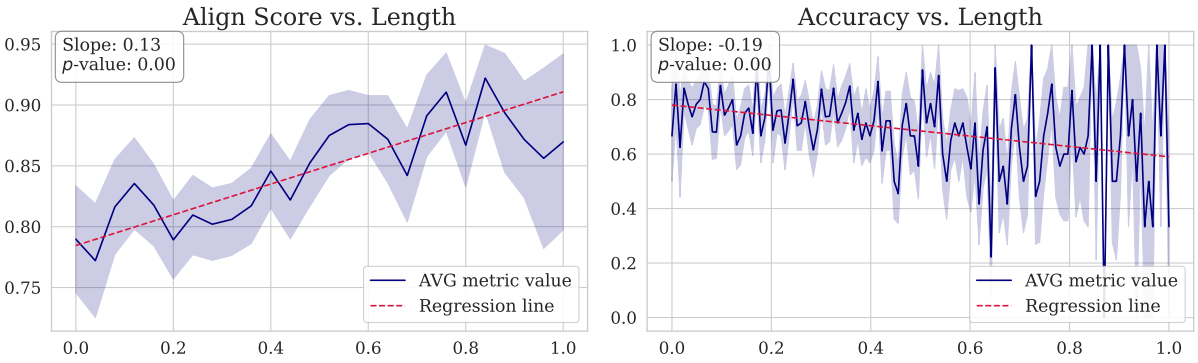

Figure 3: Trends in AlignScore and Accuracy scores with respect to normalized sequence length for XSum and GSM8K datasets respectively (Model – Gemma 2 9B). Each subplot shows a linear regression fit over binned scores. The **slope** indicates the strength and direction of correlation, **p-value** reflects statistical significance.

This suggests that there is no statistically significant correlation between generation length and quality scores. In the few cases where a trend is observed, the magnitude of the slope is quite small, especially when compared to the slopes observed in uncertainty measures. This is expected: in machine translation, the length of the output is strongly determined by the length of the input sentence, and thus variation in output length does not reflect variation in task difficulty or translation quality. We conclude that translation quality metrics are effectively length-invariant for given datasets and robust to variations in output length.

Quality Metrics for Summarization and Mathematical Reasoning are Length-Dependent. Figure 3 reports quality metric trends for XSum (summarization) and GSM8K (arithmetic question answering). Unlike machine translation, these tasks show noticeable correlations between quality scores and output length: quality (measured by Accuracy) tends to decrease with longer outputs in GSM8K, while in summarization (XSum), quality (measured by AlignScore (Zha et al., 2023)) slightly increases with length.

This observation aligns with intuition: in GSM8K, more complex problems often require longer reasoning chains, increasing the likelihood of errors; in XSum, longer summaries may better capture essential content, improving quality scores. These findings suggest that, unlike for translation task, it is important to take into account that the quality is correlated with the length of an output, even if not as strongly as uncertainty measures.

For detailed results on the relationship between quality metrics and generation length, refer to Appendix A.1.

4 Method

Building on our analysis of length-dependent trends in Section 3, we introduce a simple post-hoc correction method UNCERTAINTY-LINE designed to remove spurious correlations between uncertainty scores and output sequence length. Our approach consists of two main stages: fitting a bias model on unlabeled data and debiasing uncertainty estimates at inference time.

4.1 Problem Setup

Consider a dataset

$$\mathcal{D}_{\text{train}} = \{\mathbf{x}_i, \mathbf{y}_i, u_i = u(\mathbf{y}_i)\}_{i=1}^{N},$$

where $\mathbf{x}$ is the input sequence, $\mathbf{y}$ is the model output, and $u(\mathbf{y})$ is the associated uncertainty score. We consider two settings. In the first one, there is no systematic correlation between output length and quality (e.g., in machine translation). In the second one, quality is length-dependent (e.g., in summarization or reasoning tasks), and we use quality labels $\mathbf{y}_i^*$ for $\mathcal{D}_{\text{train}}$ to fit a quality–length model. In this case, labels are used for estimating the task-specific length-induced quality trend.

Our goal is to adjust uncertainty estimates $u(\mathbf{y})$ such that they are no longer spuriously correlated with the length $|\mathbf{y}|$ of the generated sequence.

4.2 Debiasing of UQ Scores

On the training set $\mathcal{D}_{\text{train}}$, we model the relationship between uncertainty scores and output lengths by fitting a simple linear regression:

$$\hat{u}(\mathbf{y}) = \alpha|\mathbf{y}| + \beta. \tag{1}$$

We adopt a linear regression model for debiasing based on the empirical observations in Section 3. As shown in Figure 1, UQ scores such as MSP, PPL, and MTE exhibit strong linear trends with respect to output length. In addition, Appendix D.1 presents the results of experiments using second- and third-degree polynomials. As shown, these do not provide any significant or consistent improvement over the linear fit. All of this suggests that a linear correction is both sufficient and preferable to avoid overfitting. This model captures the systematic trend between uncertainty scores and sequence length, which we aim to remove.

At inference time, we apply the learned linear model to debias raw uncertainty scores on the test set. To achieve this, for each test example, we compute a *length-debiased* uncertainty score by subtracting the length-predicted component from its raw score:

$$u^{\text{deb}}(\mathbf{y}) = u(\mathbf{y}) - \hat{u}(\mathbf{y}). \tag{2}$$

This subtraction step is equivalent to computing the residuals from the fitted regression, a standard approach in statistics for removing systematic linear trends from data.

Preserving Quality-Based Trends. However, not all length effects are spurious. As demonstrated in Section 3, for tasks like summarization or QA, quality of an output is correlated with its length and final uncertainty score should reflect this. To preserve this meaningful length-dependence, we explicitly model how quality varies with length.

Let $\mathbf{y}^*$ be the gold-standard reference for the input $\mathbf{x}$ and let us consider the quality score $q(\mathbf{y}, \mathbf{y}^*)$ between $\mathbf{y}^*$ and model generation $\mathbf{y}$. We treat the negated quality score $-q(\mathbf{y}, \mathbf{y}^*)$ as a proxy for ground-truth uncertainty, assuming that higher-quality outputs are less uncertain. We then fit a second linear model on an extended dataset $\mathcal{D}^*_{\text{train}} = \{\mathbf{x}_i, \mathbf{y}_i, q_i = q(\mathbf{y}_i, \mathbf{y}_i^*)\}_{i=1}^{N}$:

$$\hat{q}(\mathbf{y}) = \delta|\mathbf{y}| + \gamma, \tag{3}$$

where δ and γ describe the quality-induced length effect.

At test time, we debias the uncertainty score by subtracting the spurious trend $\hat{u}(\mathbf{y})$ and adding back the quality-based trend $-\hat{q}(\mathbf{y})$:

$$u^{\text{deb}}(\mathbf{y}) = u(\mathbf{y}) - \hat{u}(\mathbf{y}) - \hat{q}(\mathbf{y}). \tag{4}$$

This procedure retains task-relevant length dependencies while removing confounding biases unrelated to quality. While this method requires reference-based quality scores at training time, it can be applied to unlabeled data at inference, making it practical for real-world settings.

5 Experiments

5.1 Experimental Setup

To perform our evaluation, we extended the `LM-Polygraph` library (Fadeeva et al., 2023; Vashurin et al., 2025) by integrating our debiasing approach into its evaluation framework. The library provides built-in implementations of various uncertainty metrics, making it a convenient foundation for conducting experiments and ensuring consistent comparisons across methods.

Datasets. Tasks and dataset selection was based on the need for long-form generation tasks, in order to meaningfully analyze the relationship between generation length, uncertainty, and quality in tasks where length varies naturally. We conduct our experiments on a set of machine translation benchmarks from the WMT14 and WMT19 shared tasks (Bojar et al., 2014; Barrault et al., 2019).

Specifically, we evaluate on four language pairs from each benchmark: Cs–En, De–En, Fr–En, and Ru–En from WMT14; and De–En, Fi–En, Lt–En, and Ru–En from WMT19. Each dataset includes source inputs, model-generated translations, and reference outputs for evaluation. In addition to translation, we include two open-ended generation tasks: XSum for abstractive summarization and GSM8K for arithmetic question answering (Narayan et al., 2018; Cobbe et al., 2021). XSum consists of document-summary pairs focused on single-sentence summaries of BBC articles. GSM8K contains grade-school-level math word problems requiring multi-step reasoning to generate a final numerical answer.

For translation, we apply the debiasing formulation described in equation (2). For summarization and mathematical reasoning, we use equation (4). These formulations are chosen to align with the specific characteristics of each task: as translation quality is largely independent of output length, it is sufficient to simply remove the length-induced trend. In contrast, for tasks like summarization and mathematical reasoning, where quality often correlates with length, we retain the quality-associated component and eliminate only the spurious bias.

Models. We use three base versions of multilingual generative language models to generate outputs for all datasets: Llama 3.1 8B (Grattafiori et al., 2024), Gemma 2 9B (Riviere et al., 2024), and EuroLLM 9B (Martins et al., 2025). These models were selected to represent a diversity of open-source architectures. While Llama 3.1 8B and Gemma 2 9B are used across all tasks, EuroLLM 9B is only evaluated on translation datasets due to its limited support for an open-ended generation tasks such as summarization and mathematical reasoning.

UQ measures. We evaluate the following uncertainty quantification (UQ) measures, commonly used in sequence generation tasks: Maximum Sequence Probability (MSP), Perplexity (PPL), Mean Token Entropy (MTE), Monte Carlo Sequence Entropy (MCSE), Monte Carlo Normalized Sequence Entropy (MCNSE), Lexical Similarity with Rouge L as similarity function (LSRL) and TokenSAR (Duan et al., 2024). They capture different aspects of model uncertainty: MSP and MCSE reflect aggregate confidence in the full sequence and are not length-normalized, while PPL, MTE and MCNSE explicitly normalize for output length.

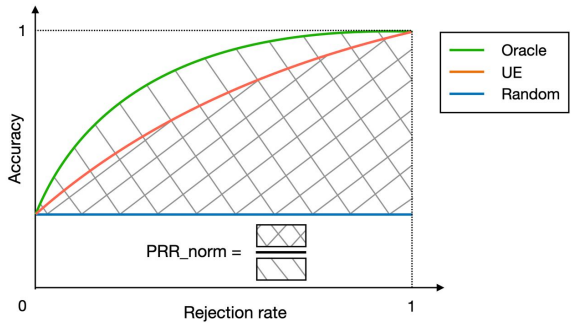

Figure 4: Illustration of the Prediction-Rejection Ratio (PRR). The PR curve plots the output quality against the rejection rate. The **oracle** curve ranks outputs perfectly by quality, while the **random** curve represents a random uncertainty score. The area between the UQ curve and random baseline (numerator) is normalized by the area between the oracle and random curves (denominator), yielding a PRR score between 0 and 1.

TokenSAR also normalizes for output length, but while accounting for each token's relevance weight, i.e., how important the token is to the overall meaning of the generation. LSRL is based on sample diversity, providing non-likelihood-based perspective on length bias in uncertainty. Detailed description of each measure is given in Appendix B.

Evaluation. We evaluate uncertainty estimates using the Prediction Rejection (PR) curve, which measures how the average output quality $Q(f(\mathbf{x}_i), \mathbf{y}_i)$ changes as uncertain examples are rejected (Malinin et al., 2017; Malinin and Gales, 2021). For a given uncertainty threshold a, it shows the average quality over all instances where the uncertainty $U(\mathbf{x}_i) < a$. To quantify the effectiveness of an uncertainty measure, we use the Prediction-Rejection Ratio (PRR). It compares the area under the PR curve (AUC) to that of a random baseline and an oracle that ranks instances perfectly by their actual output quality (see Figure 4):

$$\text{PRR} = \frac{\text{AUC}_{\text{unc}} - \text{AUC}_{\text{rnd}}}{\text{AUC}_{\text{oracle}} - \text{AUC}_{\text{rnd}}}. \quad (5)$$

A higher PRR indicates better alignment between the uncertainty estimate and the actual model quality. We use PRR as our primary evaluation measure, as it captures the utility of uncertainty scores for selective prediction in generation tasks. PRR measures how well uncertainty estimates rank outputs by quality, and is more appropriate than classification or calibration measures for continuous-valued evaluation (Fadeeva et al., 2023; Vashurin et al., 2025).

Model	WMT14								WMT19							
	Cs-En		De-En		Ru-En		Fr-En		De-En		Fi-En		Lt-En		Ru-En	
	Base	LINE	Base	LINE	Base	LINE	Base	LINE	Base	LINE	Base	LINE	Base	LINE	Base	LINE
MetricX XXL																
Llama 3.1 8B	0.47	0.54↑	0.48	0.51↑	0.46	0.54↑	0.39	0.43↑	0.43	0.47↑	0.52	0.51	0.49	0.49	0.36	0.45↑
Gemma 2 9B	0.45	0.46↑	0.47	0.49↑	0.42	0.46↑	0.36	0.37↑	0.44	0.47↑	0.45	0.45	0.34	0.37↑	0.38	0.41↑
EuroLLM 9B	0.54	0.55↑	0.54	0.55↑	0.48	0.47	0.46	0.46	0.50	0.51↑	0.49	0.47	0.42	0.47↑	0.36	0.42↑
XComet XXL																
Llama 3.1 8B	0.40	0.48↑	0.37	0.47↑	0.41	0.53↑	0.33	0.42↑	0.34	0.44↑	0.51	0.49	0.53	0.52	0.37	0.51↑
Gemma 2 9B	0.35	0.37↑	0.35	0.38↑	0.39	0.48↑	0.27	0.34↑	0.34	0.38↑	0.42	0.40	0.30	0.32↑	0.35	0.37↑
EuroLLM 9B	0.43	0.46↑	0.42	0.46↑	0.39	0.51↑	0.35	0.42↑	0.43	0.47↑	0.45	0.42	0.39	0.38	0.36	0.44↑
Comet WMT22																
Llama 3.1 8B	0.48	0.58↑	0.48	0.56↑	0.45	0.59↑	0.37	0.48↑	0.46	0.55↑	0.54	0.56↑	0.52	0.56↑	0.43	0.53↑
Gemma 2 9B	0.44	0.49↑	0.50	0.54↑	0.43	0.53↑	0.37	0.44↑	0.49	0.53↑	0.49	0.49	0.35	0.36↑	0.40	0.41↑
EuroLLM 9B	0.52	0.57↑	0.52	0.55↑	0.46	0.56↑	0.47	0.52↑	0.52	0.58↑	0.51	0.52↑	0.37	0.45↑	0.43	0.45↑

Table 1: Comparison between best raw and detrended PRR scores across all metrics and models for translation datasets. Arrows indicate improvements in detrended over raw method.

Model	XSum		GSM8k	
	Base	LINE	Base	LINE
Llama 3.1 8B	0.37	0.37	0.36	0.40↑
Gemma 2 9B	0.35	0.38↑	0.39	0.40↑

Table 2: Comparison between best raw and detrended PRR scores for summarization and mathematical reasoning tasks. Arrows indicate improvements in detrended over raw method.

Quality Metrics. To evaluate the output quality, we use the following measures: *COMET*, *XComet-XXL* and *MetricX-XXL* (Rei et al., 2022; Guerreiro et al., 2024; Juraska et al., 2024), which represent a diverse set of neural quality estimation models. We use *Accuracy* for GSM8K, and *AlignScore* (Zha et al., 2023) to measure semantic alignment between input and output for summarization task.

5.2 Results

For each dataset, Tables 1 and 2 present the PRR score of the best-performing UQ method and best-performing UNCERTAINTY-LINE variation. Across both translation and open-ended generation tasks, we consistently observe improvements in PRR scores when applying our detrending procedure, demonstrating its effectiveness in mitigating length-related bias and enhancing the reliability of uncertainty estimates. However, it is important to note that there are a few cases – particularly in XSum for Llama 3.1 8B or WMT 19 Fi-En for all considered models – where the gains are marginal or the detrended score does not outperform the raw variant.

This suggests that while length-induced bias is a prevalent issue, the extent of its impact can vary by task and model, and in some settings, additional sources of uncertainty may dominate. Detailed experimental results with breakdown of PRR scores before and after UNCERTAINTY-LINE transformation for each of the UQ methods are provided in Appendix C.

Table 3 reports improvements in PRR scores after UNCERTAINTY-LINE transformation for each UQ method, average over all tasks. As evident from the table, the most substantial gains occur in uncertainty estimation methods that are highly sensitive to sequence length, such as MSP and PPL. Detrending enhances their ability to discriminate between high- and low-quality generations. In contrast, methods like LSRL, which estimate uncertainty based on the semantic similarity of sampled outputs, exhibit far smaller improvements, if any. This is expected, as can be seen in Appendix A.2, LSRL exhibits the smallest trends with respect to length.

Figure 5 offers an illustration of the impact of our detrending procedure on three uncertainty estimation scores for translation tasks. For MSP, PPL and MTE, we observe a strong correlation between sequence length and raw uncertainty scores, indicating a clear length-induced bias. We can see that, after detrending, these trends are largely eliminated, as shown by the near-zero slopes. On the other hand, after applying the equation (4) for the summarization and mathematical reasoning tasks, the detrended uncertainty scores exhibit a length bias that is comparable to that of the quality evaluation measure itself.

Method	WMT			XSum	GSM8K
	Comet WMT22	XComet XXL	MetricX-XXL	Align Score	Accuracy
MSP	0.09 ± 0.02	0.09 ± 0.02	0.18 ± 0.01	0.03 ± 0.00	0.00 ± 0.01
PPL	0.05 ± 0.01	0.05 ± 0.01	0.02 ± 0.00	0.01 ± 0.01	0.09 ± 0.02
MTE	0.08 ± 0.01	0.07 ± 0.01	0.03 ± 0.01	0.01 ± 0.01	0.08 ± 0.02
MCSE	0.07 ± 0.02	0.07 ± 0.02	0.16 ± 0.01	0.02 ± 0.01	0.00 ± 0.00
MCNSE	0.02 ± 0.01	0.02 ± 0.01	0.00 ± 0.00	0.01 ± 0.00	0.02 ± 0.00
LSRL	0.00 ± 0.01	0.01 ± 0.01	0.00 ± 0.00	0.03 ± 0.02	0.00 ± 0.00
TokenSAR	0.05 ± 0.01	0.05 ± 0.01	0.02 ± 0.01	0.00 ± 0.01	0.09 ± 0.02

Table 3: Average improvement across datasets and models in PRR scores after detrending for three tasks: WMT (machine translation), XSum (summarization), and GSM8K (mathematical reasoning). Values are reported as mean improvements with their associated standard error of the mean (SEM).

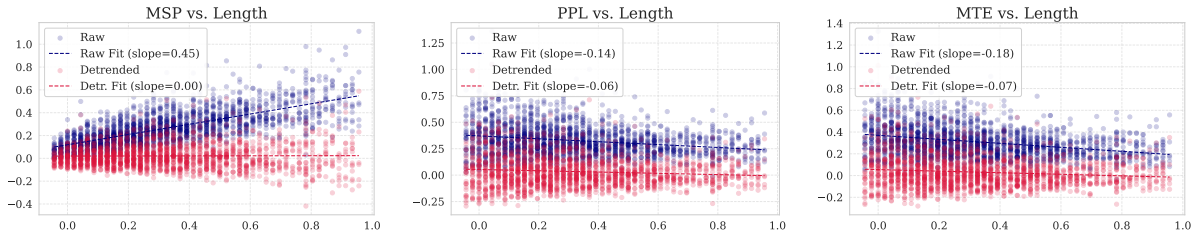

Figure 5: Example of how detrending removes length-related bias from uncertainty estimates. Shown for MSP, PPL, and MTE on WMT14 De–En (Model - Llama 3.1 8B), the raw scores exhibit clear length dependency, which is largely reduced after detrending.

6 Conclusion

We introduced UNCERTAINTY-LINE, a simple yet effective framework for removing generation length effects from uncertainty estimates. Through extensive analysis across tasks (translation, summarization, and mathematical reasoning), we demonstrated that uncertainty scores are often confounded by output length, and that correcting for this bias improves the reliability of uncertainty-based rejection. Our method is lightweight, model-agnostic, and requires minimal supervision only when known quality-length correlation is present. While certain assumptions limit its applicability in more complex settings, UNCERTAINTY-LINE offers a strong foundation for more interpretable and trustworthy uncertainty estimation in text generation.

Future work could explore addressing length bias directly during model training, rather than correcting it in a post-hoc manner.

7 Limitations

While UNCERTAINTY-LINE offers a simple and effective correction for length-induced bias, several important considerations remain.

We assume a linear relationship between uncertainty scores and output length, as well as quality scores and length. While this simplifies both implementation and interpretation, it may not fully capture the complexity of interactions between length and uncertainty. However, as demonstrated in Section D.1, linear approximation is a reasonable and effective first-order correction. Nonetheless, in tasks such as multi-step reasoning, where uncertainty may follow phase-specific patterns, a linear fit may be insufficient.

Our method requires a small number of quality-labeled examples to estimate the quality-length relationship. However, this only applies when there is a known or observed correlation between output length and quality. Moreover, in Appendix D.2 we demonstrate that the quality trend can be estimated using as little as 500 generations. In tasks where quality is largely length-independent, our method can be applied without any quality annotations.

This leads to another consideration - we assume prior knowledge of quality-length relationship. In tasks where this relationship is unclear or poorly understood, effectiveness may be reduced.

Ethical Considerations

UNCERTAINTY-LINE improves uncertainty quantification by removing spurious correlations between output length and estimated uncertainty. This helps to prevent misleading high- or low-uncertainty signals that often affect longer generations. However, it does not prevent the generation of incorrect or harmful content, and low uncertainty scores do not guarantee factuality. Moreover, it does not address factors such as prompt phrasing or out-of-domain data.

Reliable uncertainty estimates are crucial for enabling selective generation, abstention, or human-in-the-loop review, especially in tasks where correctness cannot be easily verified. UNCERTAINTY-LINE improves robustness to length-related bias, but it should be used as part of a broader reliability strategy, especially in critical applications.

Acknowledgments

We would like to thank Evgenii Tsymbalov for the valuable advice and feedback during this work.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *Preprint*, arXiv:2001.09977.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Conference on Machine Translation*, pages 1–61, Florence, Italy.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA.

Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, and Yi Zhang. 2020. Calibration, entropy rates, and memory in language models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1089–1099, Vienna, Austria. PMLR.

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*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5050–5063, Bangkok, Thailand.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-Polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Yarin Gal. 2016. *Uncertainty in Deep Learning*. Ph.D. thesis, University of Cambridge.

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.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Neha Gupta, Hari Narasimhan, Ankit Singh Rawat, Wittawat Jitkrittum, Aditya Menon, and Sanjiv Kumar. 2024. Language model cascades: Token-level uncertainty and beyond. In *Proceedings of the Twelfth International Conference on Learning Representations*, Vienna, Austria.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *ArXiv*, abs/2306.04459.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA.

Nikita Kotelevskii, Vladimir Kondratyev, Martin Takáč, Eric Moulines, and Maxim Panov. 2025. From risk to uncertainty: Generating predictive uncertainty measures via bayesian estimation. In *Proceedings of the Thirteenth International Conference on Learning Representations*, Singapore.

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

Rui Li, Jing Long, Muge Qi, Heming Xia, Lei Sha, Peiyi Wang, and Zhifang Sui. 2025. Towards harmonized uncertainty estimation for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 22938–22953, Vienna, Austria.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2024. Can LLMs learn uncertainty on their own? Expressing uncertainty effectively in a self-training manner. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21635–21645, Miami, Florida, USA.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *Proceedings of the Ninth International Conference on Learning Representations*, Vienna, Austria.

Andrey Malinin, Anton Ragni, Kate Knill, and Mark Gales. 2017. Incorporating uncertainty into deep learning for spoken language assessment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 45–50, Vancouver, Canada.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. EuroLLM: Multilingual language models for Europe. *Procedia Computer Science*, 255:53–62.

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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-evaluation improves selective generation in large language models. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 49–64. PMLR.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 176 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. 2025. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 743–759, Vienna, Austria.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, and 1 others. 2025. Benchmarking uncertainty quantification methods for large language models with LM-Polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the Twelfth International Conference on Learning Representations*, Vienna, Austria.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 11328–11348, Toronto, Canada.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2023. Calibrating sequence likelihood improves conditional language generation. In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali, Rwanda.

A Length Bias Analysis

A.1 Response Quality vs Generation Length

In this section we report the detailed relation between performance metrics for various tasks and length of the generated output. Figures 6, 7 and 8 show average normalized values of performance metrics for NMT tasks at each generated sequence length, as well as a linear OLS fit to this dependency. Figure 9 contains similar charts for the QA and ATS tasks.

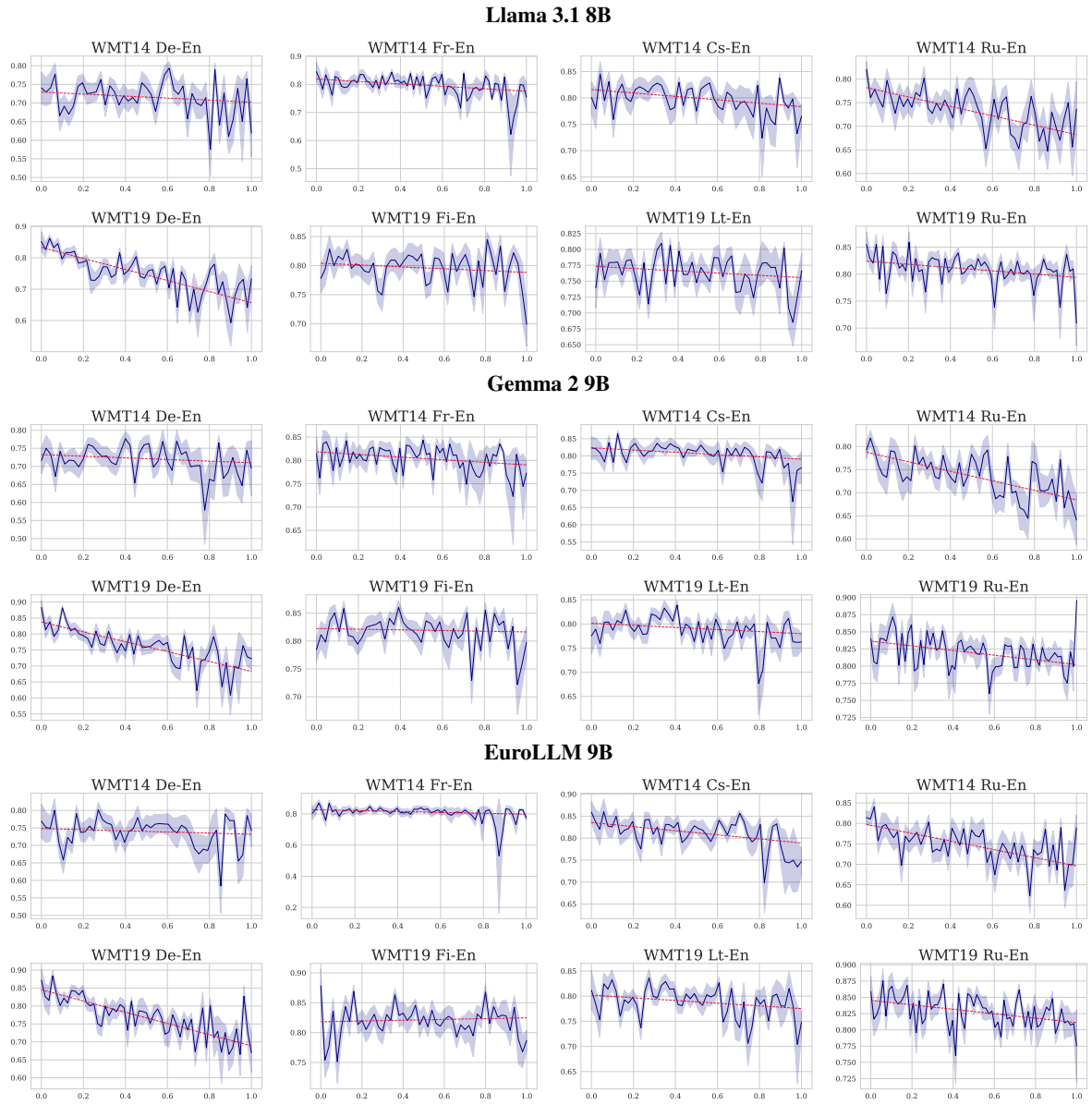

Figure 6: Comet score trends with respect to normalized generated sequence length across four machine translation datasets. Each subplot shows a linear regression fit over binned Comet scores.

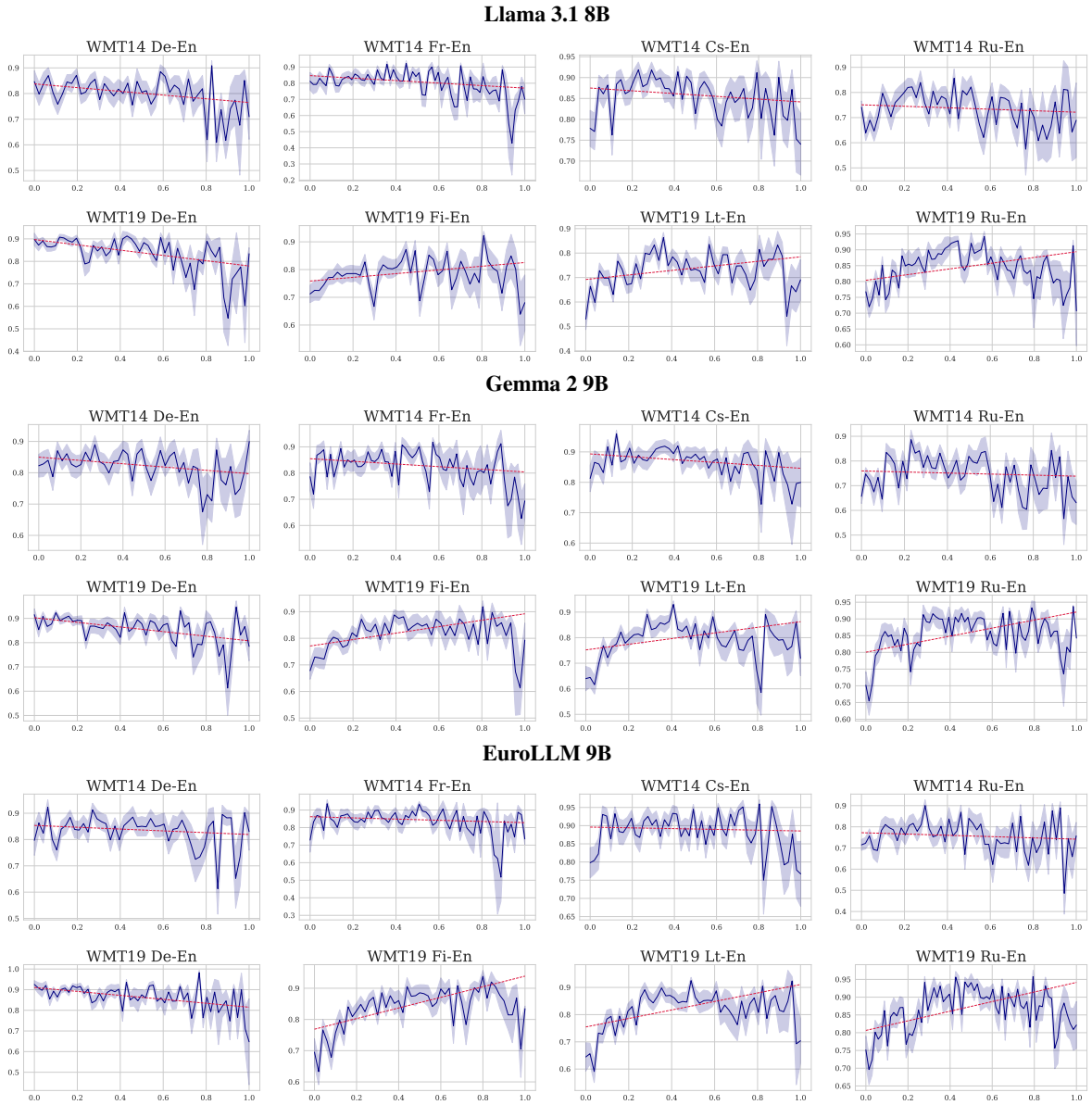

Figure 7: XComet-XXL score trends with respect to normalized generated sequence length across four machine translation datasets. Each subplot shows a linear regression fit over binned XComet-XXL scores.

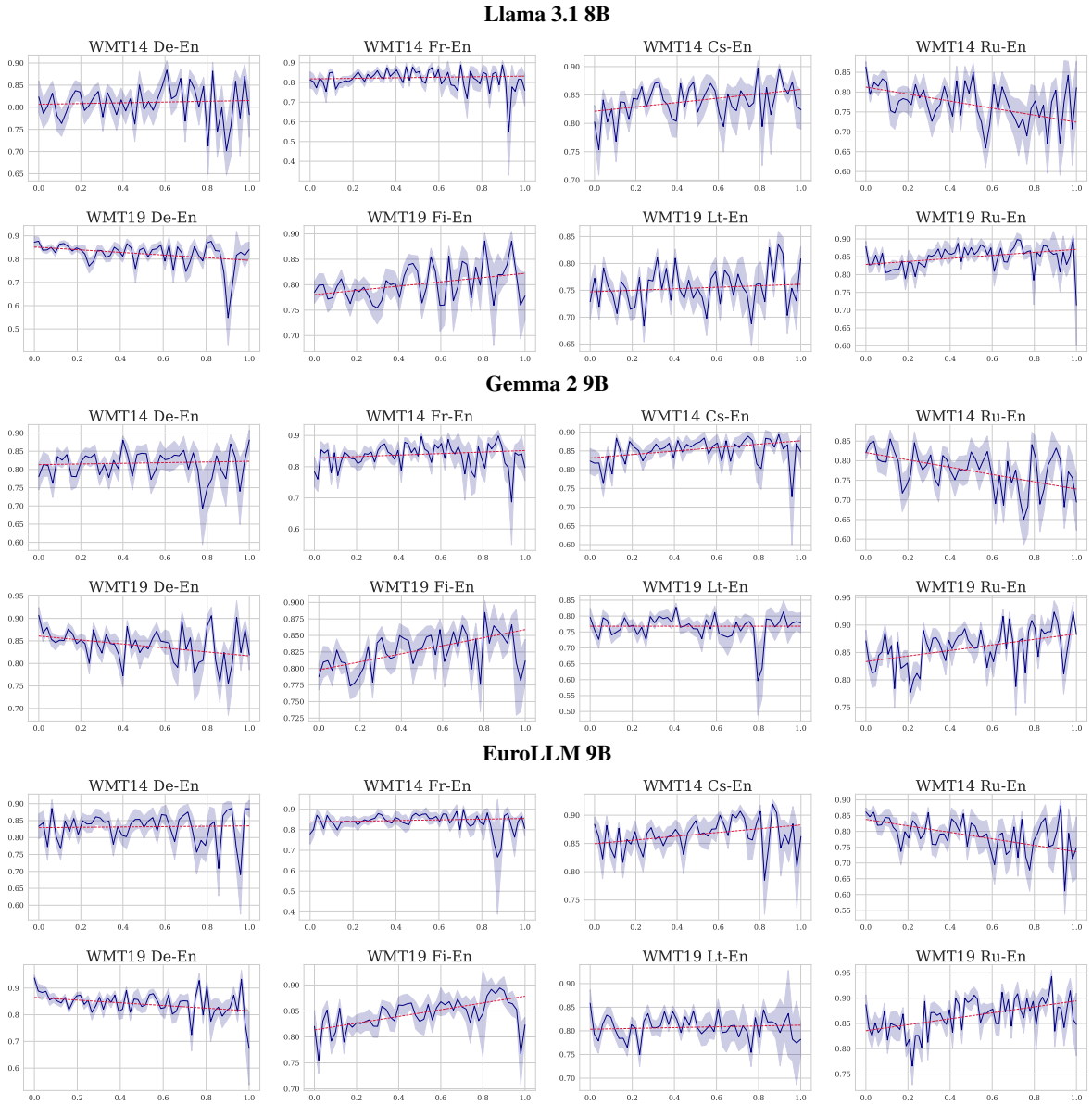

Figure 8: MetricX-XXL score trends with respect to normalized generated sequence length across four machine translation datasets. Each subplot shows a linear regression fit over binned MetricX-XXL scores.

Dataset	COMET		MetricX		XCOMET	
	slope	p-val	slope	p-val	slope	p-val
Llama 3.1 8B						
WMT14 Cs-En	-0.032	0.019	0.039	0.008	-0.033	0.092
WMT14 De-En	-0.027	0.176	0.009	0.651	-0.073	0.004
WMT14 Fr-En	-0.043	0.002	0.014	0.404	-0.077	0.002
WMT14 Ru-En	-0.100	0.000	-0.088	0.000	-0.029	0.272
WMT19 De-En	-0.176	0.000	-0.057	0.000	-0.118	0.000
WMT19 Fi-En	-0.016	0.288	0.042	0.012	0.067	0.005
WMT19 Lt-En	-0.018	0.223	0.014	0.382	0.093	0.000
WMT19 Ru-En	-0.030	0.011	0.042	0.001	0.090	0.000
Gemma 2 9B						
WMT14 Cs-En	-0.033	0.018	0.046	0.001	-0.047	0.018
WMT14 De-En	-0.023	0.251	0.009	0.636	-0.053	0.027
WMT14 Fr-En	-0.028	0.048	0.024	0.139	-0.051	0.035
WMT14 Ru-En	-0.103	0.000	-0.093	0.000	-0.021	0.404
WMT19 De-En	-0.154	0.000	-0.044	0.000	-0.095	0.000
WMT19 Fi-En	-0.006	0.647	0.061	0.000	0.121	0.000
WMT19 Lt-En	-0.022	0.108	-0.000	0.998	0.110	0.000
WMT19 Ru-En	-0.034	0.004	0.051	0.000	0.120	0.000
EuroLLM 9B						
WMT14 Cs-En	-0.047	0.001	0.033	0.010	-0.011	0.572
WMT14 De-En	-0.016	0.413	0.005	0.779	-0.035	0.157
WMT14 Fr-En	-0.029	0.042	0.017	0.294	-0.034	0.152
WMT14 Ru-En	-0.101	0.000	-0.102	0.000	-0.031	0.250
WMT19 De-En	-0.156	0.000	-0.050	0.000	-0.096	0.000
WMT19 Fi-En	0.007	0.648	0.065	0.000	0.170	0.000
WMT19 Lt-En	-0.027	0.053	0.009	0.523	0.156	0.000
WMT19 Ru-En	-0.034	0.003	0.059	0.000	0.135	0.000

Table 4: Regression slopes and p-values measuring the correlation between output length and three machine translation quality metrics (Comet, MetricX-XXL, Xcomet-XXL) across different translation datasets and models.

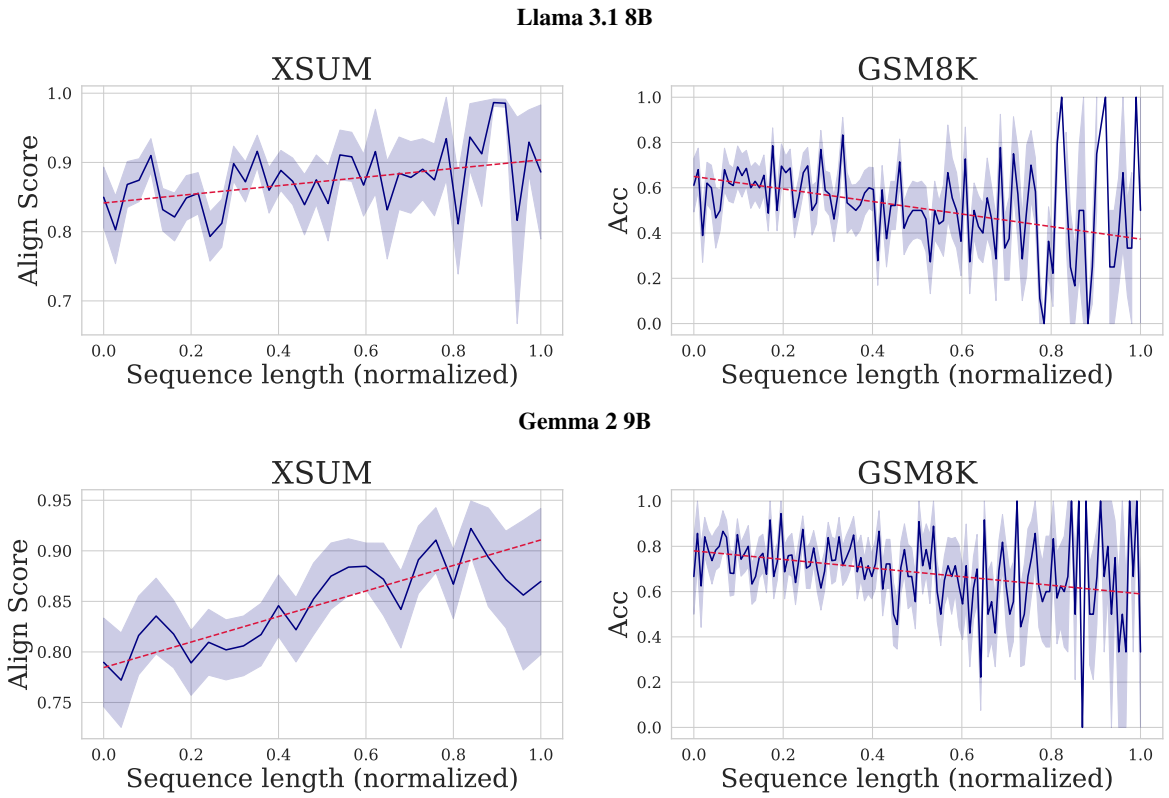

Figure 9: Align Score and Accuracy trends with respect to normalized generated sequence length across four machine translation datasets. Each subplot shows a linear regression fit over binned Align Score and Accuracy scores.

Dataset	AlignScore slope	AlignScore p-val
Llama 3.1 8B		
XSum	0.062	0.042
Gemma 2 9B		
XSum	0.126	0.000

Table 5: Regression slopes and p-values measuring the correlation between output length and summarization quality metric (Align Score).

Dataset	Accuracy slope	Accuracy p-val
Llama 3.1 8B		
GSM8k	-0.277	0.000
Gemma 2 9B		
GSM8k	-0.190	0.000

Table 6: Regression slopes and p-values measuring the correlation between output length and QA quality metric (Accuracy).

A.2 UQ Values vs Generation Length

Figures 10, 11 and 12 depict the length bias of various UQ methods under consideration. Specifics of the charts are the same as in A.1.

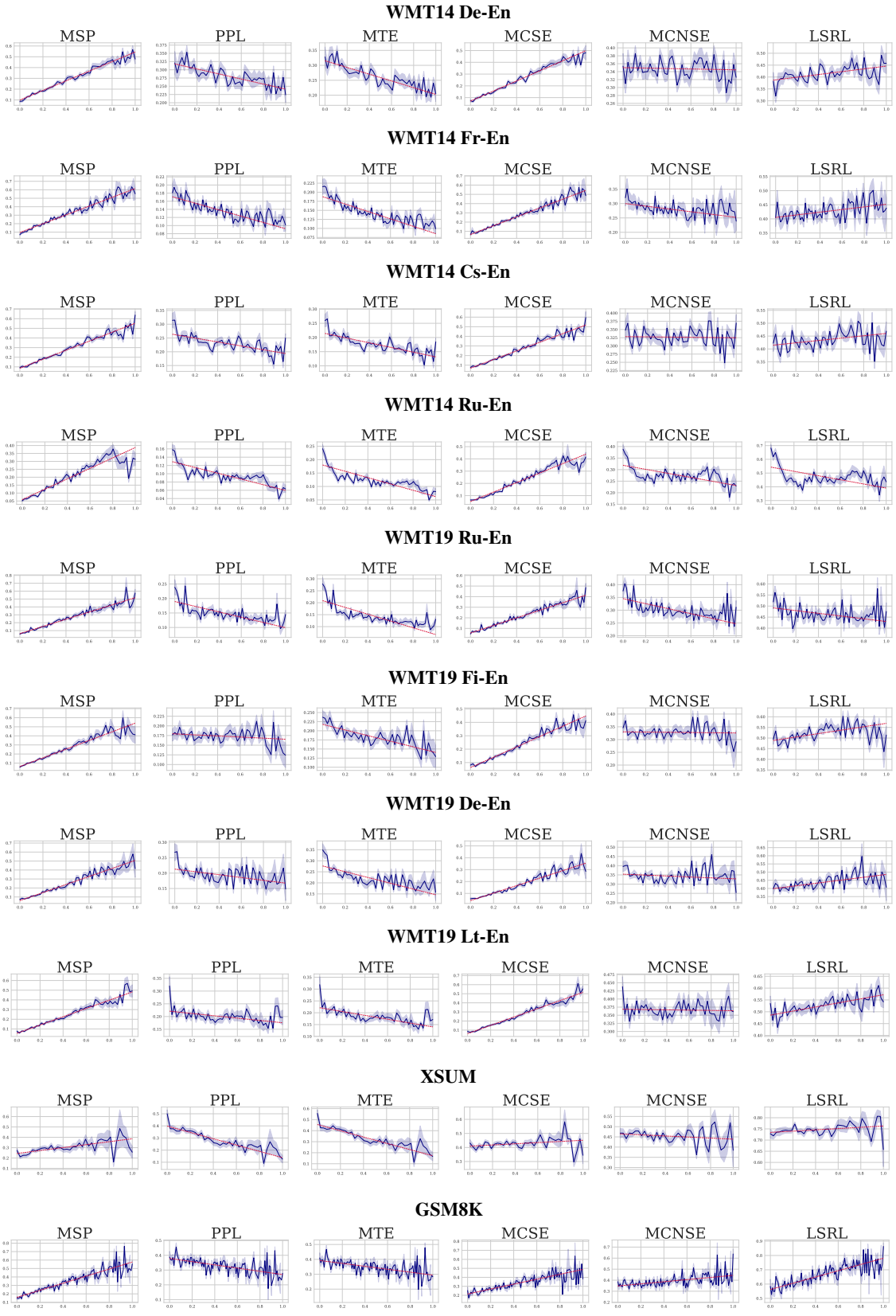

Figure 10: Uncertainty metric trends for model **LLAMA** across all datasets.

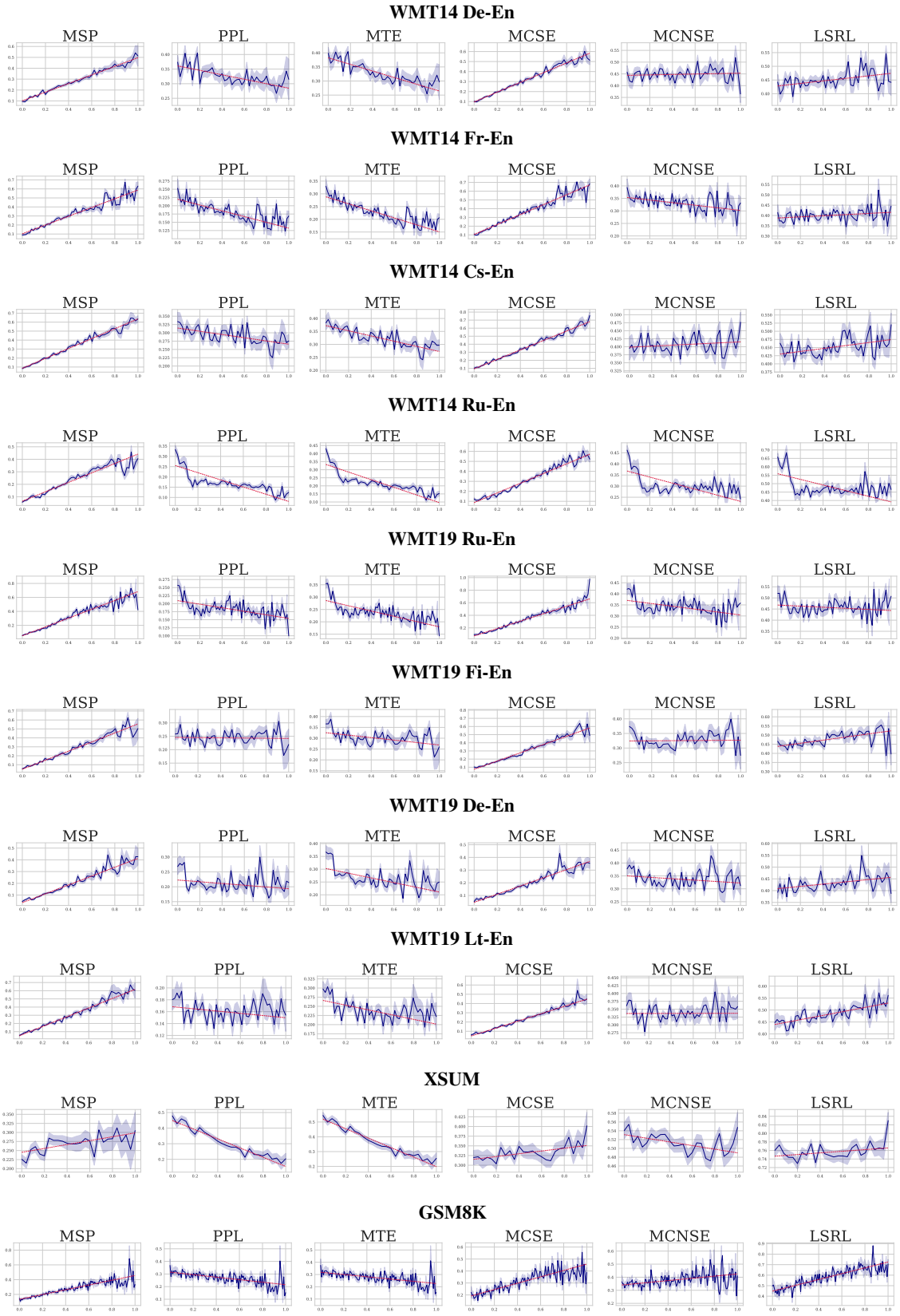

Figure 11: Uncertainty metric trends for model **GEMMA** across all datasets.

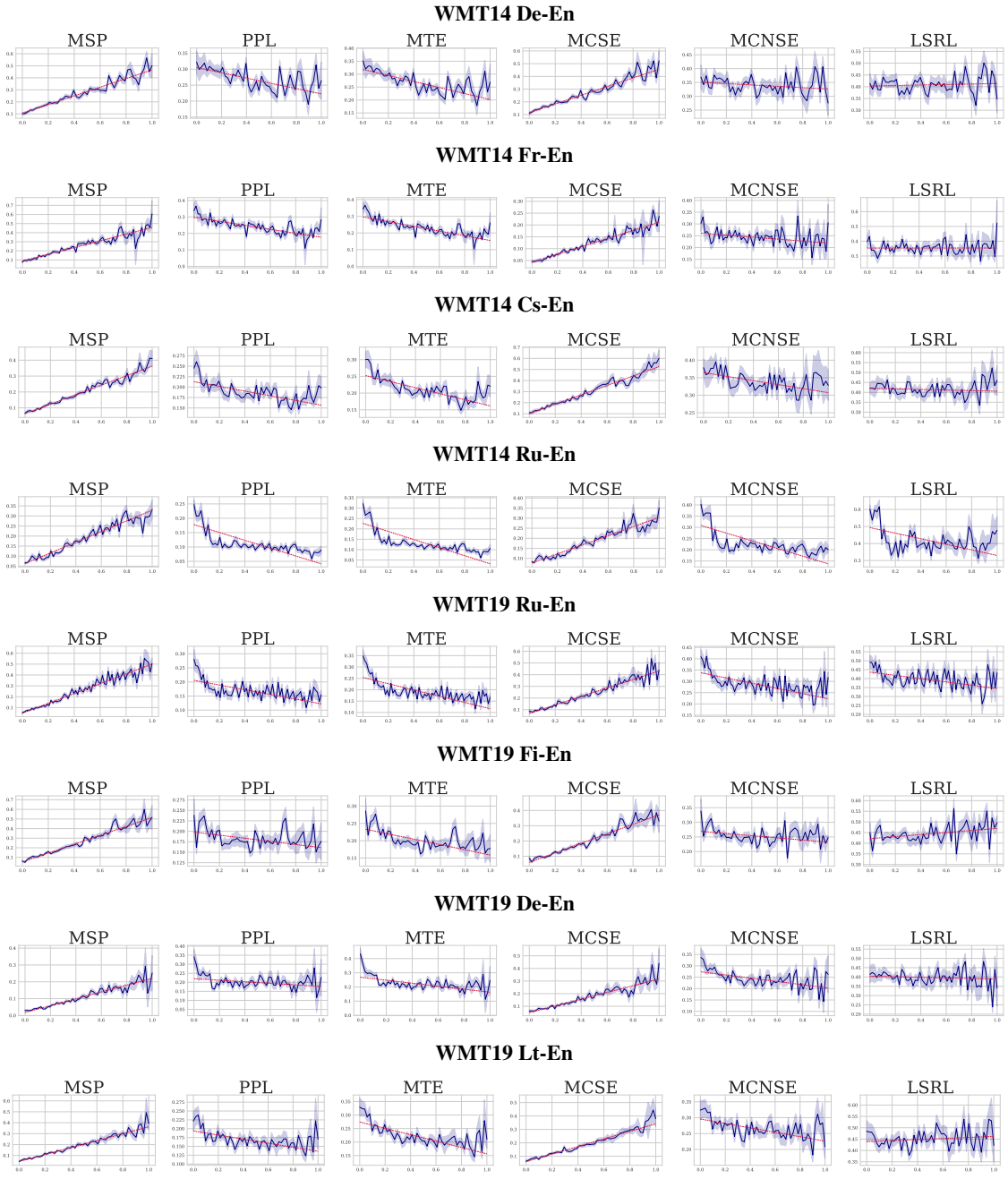

Figure 12: Uncertainty metric trends for model **EUROLLM** across all datasets.

Dataset	MSP		PPL		MTE		MCSE		MCNSE		LSRL		TokenSAR	
	slope	p-val	slope	p-val	slope	p-val	slope	p-val	slope	p-val	slope	p-val	slope	p-val
Llama 3.1 8B														
WMT14 CS-EN	0.47	0.00	-0.07	0.00	-0.08	0.00	0.45	0.00	-0.00	0.76	0.05	0.00	-0.10	0.00
WMT14 DE-EN	0.45	0.00	-0.08	0.00	-0.11	0.00	0.43	0.00	-0.01	0.67	0.06	0.00	-0.08	0.00
WMT14 FR-EN	0.52	0.00	-0.08	0.00	-0.10	0.00	0.48	0.00	-0.05	0.00	0.05	0.00	-0.06	0.00
WMT14 RU-EN	0.33	0.00	-0.07	0.00	-0.12	0.00	0.39	0.00	-0.09	0.00	-0.15	0.00	-0.07	0.00
WMT19 DE-EN	0.45	0.00	-0.05	0.00	-0.13	0.00	0.31	0.00	-0.03	0.06	0.09	0.00	-0.07	0.00
WMT19 FI-EN	0.48	0.00	-0.01	0.15	-0.08	0.00	0.39	0.00	-0.00	0.74	0.08	0.00	-0.03	0.00
WMT19 LT-EN	0.43	0.00	-0.05	0.00	-0.08	0.00	0.45	0.00	-0.00	0.73	0.09	0.00	-0.06	0.00
WMT19 RU-EN	0.46	0.00	-0.09	0.00	-0.14	0.00	0.36	0.00	-0.10	0.00	-0.06	0.00	-0.09	0.00
Gemma 2 9B														
WMT14 CS-EN	0.55	0.00	-0.05	0.00	-0.10	0.00	0.59	0.00	0.02	0.23	0.04	0.00	-0.08	0.00
WMT14 DE-EN	0.40	0.00	-0.08	0.00	-0.12	0.00	0.49	0.00	0.01	0.53	0.05	0.00	-0.09	0.00
WMT14 FR-EN	0.49	0.00	-0.09	0.00	-0.14	0.00	0.57	0.00	-0.05	0.00	0.03	0.04	-0.09	0.00
WMT14 RU-EN	0.38	0.00	-0.17	0.00	-0.24	0.00	0.49	0.00	-0.14	0.00	-0.17	0.00	-0.18	0.00
WMT19 DE-EN	0.38	0.00	-0.03	0.01	-0.09	0.00	0.32	0.00	-0.03	0.03	0.05	0.00	-0.06	0.00
WMT19 FI-EN	0.51	0.00	-0.00	0.79	-0.06	0.00	0.51	0.00	0.00	0.84	0.09	0.00	-0.03	0.01
WMT19 LT-EN	0.57	0.00	-0.02	0.03	-0.06	0.00	0.40	0.00	0.00	0.96	0.09	0.00	-0.03	0.00
WMT19 RU-EN	0.63	0.00	-0.06	0.00	-0.11	0.00	0.59	0.00	-0.07	0.00	-0.02	0.08	-0.07	0.00
EuroLLM 9B														
WMT14 CS-EN	0.30	0.00	-0.06	0.00	-0.09	0.00	0.42	0.00	-0.06	0.00	-0.01	0.43	-0.06	0.00
WMT14 DE-EN	0.36	0.00	-0.08	0.00	-0.12	0.00	0.34	0.00	-0.02	0.11	0.01	0.55	-0.09	0.00
WMT14 FR-EN	0.37	0.00	-0.12	0.00	-0.14	0.00	0.17	0.00	-0.04	0.00	0.00	0.99	-0.12	0.00
WMT14 RU-EN	0.27	0.00	-0.14	0.00	-0.19	0.00	0.23	0.00	-0.17	0.00	-0.16	0.00	-0.16	0.00
WMT19 DE-EN	0.20	0.00	-0.04	0.00	-0.11	0.00	0.27	0.00	-0.07	0.00	-0.02	0.43	-0.05	0.00
WMT19 FI-EN	0.46	0.00	-0.04	0.00	-0.07	0.00	0.31	0.00	-0.04	0.00	0.05	0.01	-0.05	0.00
WMT19 LT-EN	0.32	0.00	-0.06	0.00	-0.12	0.00	0.28	0.00	-0.07	0.00	0.02	0.23	-0.06	0.00
WMT19 RU-EN	0.45	0.00	-0.08	0.00	-0.14	0.00	0.36	0.00	-0.12	0.00	-0.09	0.00	-0.09	0.00

Table 7: Regression slopes and p-values measuring the correlation between output length and various uncertainty metrics on machine translation datasets.

Dataset	MSP		PPL		MTE		MCSE		MCNSE		LSRL		TokenSAR	
	slope	p-val	slope	p-val	slope	p-val	slope	p-val	slope	p-val	slope	p-val	slope	p-val
Llama 3.1 8B														
XSUM	0.142	0.000	-0.261	0.000	-0.283	0.000	0.042	0.008	-0.026	0.096	0.029	0.025	-0.264	0.0
Gemma 2 9B														
XSUM	0.054	0.001	-0.295	0.000	-0.320	0.000	0.036	0.007	-0.041	0.001	0.019	0.067	-0.307	0.0

Table 8: Regression slopes and p-values measuring the correlation between output length and various uncertainty metrics on XSUM dataset.

Dataset	MSP		PPL		MTE		MCSE		MCNSE		LSRL		TokenSAR	
	slope	p-val	slope	p-val	slope	p-val	slope	p-val	slope	p-val	slope	p-val	slope	p-val
Llama 3.1 8B														
GSM8K	0.418	0.000	-0.109	0.000	-0.098	0.000	0.273	0.000	0.096	0.000	0.219	0.000	-0.109	0.0
Gemma 2 9B														
GSM8K	0.322	0.000	-0.100	0.000	-0.092	0.000	0.266	0.000	0.085	0.000	0.277	0.000	-0.1	0.0

Table 9: Regression slopes and p-values measuring the correlation between output length and various uncertainty metrics on GSM8k dataset.

B Detailed Description of Uncertainty Quantification Methods

Here, we provide details of the UQ methods used in the experiments omitted from the main part of the paper.

Maximum Sequence Probability (MSP) is one of the simplest and most direct methods for estimating uncertainty. It measures the negative log-likelihood of the most likely output sequence given a specific input. Under the assumption that the model is most confident in its most probable output, lower values indicate higher confidence:

$$U_{\text{MSP}}(\mathbf{y} \mid \mathbf{x}) = -\log P(\mathbf{y} \mid \mathbf{x}). \tag{6}$$

Perplexity (PPL) is a widely used metric for evaluating uncertainty in autoregressive models (Fomicheva et al., 2020). It computes the negative average log-likelihood per token, making it explicitly length-normalized. Lower perplexity indicates higher model confidence:

$$U_{\text{PPL}}(\mathbf{y} \mid \mathbf{x}) = -\frac{1}{L} \log P(\mathbf{y} \mid \mathbf{x}), \tag{7}$$

where L is the length of the output sequence $\mathbf{y}$.

Mean Token Entropy (MTE) captures the average uncertainty at the token level. It measures how peaked or flat the model's predicted distribution is at each decoding step:

$$U_{\text{MTE}}(\mathbf{y} \mid \mathbf{x}) = \frac{1}{L} \sum_{l=1}^{L} \mathcal{H}(y_l \mid \mathbf{y}_{<l}, \mathbf{x}), \tag{8}$$

where $\mathcal{H}(y_l \mid \mathbf{y}_{<l}, \mathbf{x}) = -\sum_v P(y_l = v \mid \mathbf{y}_{<l}, \mathbf{x}) \log P(y_l = v \mid \mathbf{y}_{<l}, \mathbf{x})$ is the entropy of the token distribution at position l.

Monte Carlo Sequence Entropy (MCSE) estimates sequence-level uncertainty via sampling. We draw M sequences $\mathbf{y}^{(i)} i = 1^M$ from the model's output distribution and compute their average negative log-likelihood:

$$U_{\text{MCSE}}(\mathbf{x}) = -\frac{1}{M} \sum_{i=1}^{M} \log P(\mathbf{y}^{(i)} \mid \mathbf{x}). \tag{9}$$

Monte Carlo Normalized Sequence Entropy (MCNSE) is a length-normalized variant of MCSE. For each sampled sequence $\mathbf{y}^{(i)}$, we normalize the log-likelihood by its length $L^{(i)}$:

$$U_{\text{MCNSE}}(\mathbf{x}) = -\frac{1}{M} \sum_{i=1}^{M} \frac{1}{L^{(i)}} \log P(\mathbf{y}^{(i)} \mid \mathbf{x}). \tag{10}$$

Lexical Similarity with ROUGE-L (LSRL) measures the average pairwise lexical similarity between all sampled sequences. Unlike the previous methods, which rely on model probabilities, LSRL captures diversity among generated hypotheses by comparing their surface forms:

$$U_{\text{LSRL}}(\mathbf{x}) = 1 - \frac{2}{M(M-1)} \sum_{i<j} \text{ROUGE-L}(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}). \tag{11}$$

TokenSAR computes relevance-weighted average of the negative log probabilities of generated tokens:

$$U_{\text{TokenSAR}}(\mathbf{x}) = -\sum_{l=1}^{L} \tilde{R}_T(y_l, \mathbf{y}, \mathbf{x}) \log P(y_l \mid \mathbf{y}_{<l}, \mathbf{x}), \tag{12}$$

where the normalized relevance weight for each token y_l is given by $\tilde{R}_T(y_k, \mathbf{y}, \mathbf{x}) = fracR_T(y_k, \mathbf{y}, \mathbf{x})\sum_{l=1}^{L} R_T(y_l, \mathbf{y}, \mathbf{x})$. and $R_T(\cdot)$ denotes the token relevance function, derived from a sentence similarity function $g(\cdot, \cdot)$ as $R_T(y_k, \mathbf{y}, \mathbf{x}) = 1 - g(\mathbf{x} \cup \mathbf{y}, \mathbf{x} \cup \mathbf{y} \setminus y_k)$.

C Detailed Experimental Results

Tables 10, 12 and 11 contain PRR scores for all UQ methods, along with their LINE counterparts for NMT datasets. Table 13 contains the same data for summarization and mathematical reasoning.

	WMT14				**WMT19**			
	Cs-En	De-En	Ru-En	Fr-En	De-En	Fi-En	Lt-En	Ru-En
				Llama 3.1 8B				
MSP	0.42	0.39	0.45	0.35	0.46	0.19	0.29	0.43
MSP-LINE	0.47	0.49	0.48	0.40	0.51	0.47	0.47	0.41
PPL	0.42	0.46	0.37	0.31	0.41	0.52	0.47	0.32
PPL-LINE	0.52	0.51	0.53	0.41	0.46	0.52	0.49	0.46
MTE	0.44	0.48	0.41	0.37	0.42	0.54	0.52	0.33
MTE-LINE	**0.58**	**0.56**	**0.59**	**0.48**	**0.55**	**0.56**	**0.56**	**0.53**
MCSE	0.36	0.32	0.35	0.30	0.36	0.08	0.20	0.36
MCSE-LINE	0.38	0.36	0.33	0.28	0.38	0.32	0.36	0.32
MCNSE	0.48	0.44	0.40	0.36	0.43	0.46	0.48	0.35
MCNSE-LINE	0.49	0.44	0.47	0.39	0.45	0.46	0.48	0.44
LSRL	0.45	0.44	0.38	0.35	0.46	0.37	0.42	0.35
LSRL-LINE	0.41	0.41	0.44	0.32	0.40	0.37	0.40	0.38
TokenSAR	0.44	0.45	0.37	0.35	0.40	0.52	0.46	0.32
TokenSAR-LINE	0.51	0.52	0.52	0.41	0.47	0.53	0.49	0.46
				Gemma 2 9B				
MSP	0.40	0.37	0.43	0.29	0.49	0.18	0.35	0.40
MSP-LINE	0.48	0.50	0.47	0.38	**0.53**	0.42	0.36	**0.41**
PPL	0.44	0.48	0.38	0.36	0.44	0.46	0.30	0.31
PPL-LINE	0.46	0.51	0.50	0.40	0.47	0.46	0.32	0.35
MTE	0.44	0.49	0.38	0.37	0.44	0.49	0.30	0.30
MTE-LINE	**0.49**	**0.54**	**0.53**	**0.44**	0.51	**0.49**	**0.36**	0.40
MCSE	0.32	0.31	0.35	0.28	0.41	0.09	0.29	0.36
MCSE-LINE	0.39	0.43	0.38	0.35	0.47	0.31	0.29	0.39
MCNSE	0.44	0.50	0.42	0.37	0.47	0.41	0.35	0.37
MCNSE-LINE	0.44	0.50	0.48	0.39	0.49	0.41	0.35	0.41
LSRL	0.40	0.47	0.40	0.33	0.43	0.40	0.34	0.34
LSRL-LINE	0.38	0.46	0.45	0.32	0.41	0.40	0.28	0.35
TokenSAR	0.41	0.46	0.36	0.37	0.42	0.45	0.29	0.28
TokenSAR-LINE	0.45	0.50	0.49	0.41	0.46	0.45	0.35	0.34
				EuroLLM 9B				
MSP	0.29	0.33	0.42	0.24	0.40	0.16	0.28	0.43
MSP-LINE	0.37	0.46	0.50	0.34	0.51	0.39	0.34	0.44
PPL	0.51	0.50	0.43	0.44	0.52	0.48	0.36	0.32
PPL-LINE	0.53	0.53	0.54	0.47	0.54	0.49	0.40	0.39
MTE	0.52	0.52	0.46	0.47	0.51	0.51	0.37	0.34
MTE-LINE	**0.57**	**0.55**	**0.56**	**0.52**	**0.58**	**0.52**	**0.45**	**0.45**
MCSE	0.35	0.36	0.42	0.28	0.41	0.21	0.34	0.42
MCSE-LINE	0.46	0.47	0.46	0.40	0.49	0.40	0.37	0.39
MCNSE	0.23	0.36	0.28	0.22	0.34	0.36	0.28	0.24
MCNSE-LINE	0.22	0.36	0.33	0.22	0.34	0.35	0.28	0.29
LSRL	0.32	0.38	0.31	0.29	0.40	0.35	0.32	0.26
LSRL-LINE	0.32	0.37	0.36	0.29	0.41	0.35	0.31	0.29
TokenSAR	0.42	0.47	0.42	0.40	0.44	0.46	0.31	0.34
TokenSAR-LINE	0.42	0.49	0.53	0.41	0.45	0.46	0.35	0.41

Table 10: Detailed PRR scores for all methods and their LINE counterparts. Metric: Comet WMT22.

	WMT14				WMT19			
	Cs-En	De-En	Ru-En	Fr-En	De-En	Fi-En	Lt-En	Ru-En
					Llama 3.1 8B			
MSP	0.21	0.22	0.31	0.19	0.23	0.08	0.12	0.26
MSP-LINE	0.40	0.42	0.44	0.31	0.40	0.41	0.38	0.35
PPL	0.43	0.45	0.41	0.32	0.39	0.48	0.43	0.31
PPL-LINE	0.48	0.47	0.49	0.36	0.40	0.48	0.44	0.40
MTE	0.47	0.48	0.46	0.39	0.43	**0.52**	0.49	0.36
MTE-LINE	**0.54**	**0.51**	**0.54**	**0.43**	**0.47**	0.51	**0.49**	**0.45**
MCSE	0.16	0.14	0.20	0.16	0.14	-0.00	0.07	0.21
MCSE-LINE	0.29	0.29	0.30	0.25	0.32	0.29	0.29	0.27
MCNSE	0.42	0.38	0.40	0.32	0.38	0.39	0.44	0.32
MCNSE-LINE	0.42	0.38	0.43	0.34	0.39	0.39	0.44	0.36
LSRL	0.39	0.35	0.37	0.30	0.36	0.32	0.42	0.31
LSRL-LINE	0.38	0.35	0.38	0.29	0.35	0.33	0.40	0.33
TokenSAR	0.46	0.45	0.41	0.34	0.39	0.49	0.44	0.32
TokenSAR-LINE	0.49	0.47	0.48	0.36	0.41	0.49	0.44	0.40
					Gemma 2 9B			
MSP	0.19	0.22	0.29	0.13	0.28	0.06	0.24	0.27
MSP-LINE	0.39	0.45	0.41	0.29	0.45	0.35	0.35	0.39
PPL	0.42	0.47	0.41	0.33	0.42	0.41	0.33	0.34
PPL-LINE	0.43	0.48	0.43	0.33	0.43	0.41	0.34	0.37
MTE	0.45	0.47	0.42	0.36	0.44	**0.45**	0.34	0.35
MTE-LINE	**0.46**	**0.49**	**0.46**	**0.37**	**0.47**	0.45	**0.37**	**0.41**
MCSE	0.11	0.15	0.21	0.12	0.20	-0.03	0.17	0.23
MCSE-LINE	0.31	0.36	0.33	0.28	0.40	0.27	0.25	0.35
MCNSE	0.38	0.43	0.40	0.33	0.43	0.36	0.32	0.38
MCNSE-LINE	0.38	0.43	0.41	0.33	0.44	0.36	0.32	0.40
LSRL	0.35	0.38	0.36	0.26	0.36	0.34	0.34	0.34
LSRL-LINE	0.35	0.38	0.37	0.26	0.36	0.36	0.31	0.34
TokenSAR	0.42	0.46	0.39	0.33	0.42	0.41	0.31	0.32
TokenSAR-LINE	0.43	0.47	0.41	0.34	0.44	0.41	0.35	0.37
					EuroLLM 9B			
MSP	0.13	0.23	0.30	0.11	0.23	0.04	0.15	0.28
MSP-LINE	0.32	0.44	0.45	0.27	0.43	0.34	0.31	0.38
PPL	0.51	0.52	0.45	0.41	0.48	0.44	0.40	0.33
PPL-LINE	0.52	0.53	0.46	0.42	0.48	0.44	0.42	0.37
MTE	0.54	0.54	**0.48**	0.46	0.50	**0.49**	0.42	0.36
MTE-LINE	**0.55**	**0.55**	0.47	**0.46**	**0.51**	0.47	**0.47**	**0.42**
MCSE	0.20	0.24	0.28	0.16	0.25	0.10	0.25	0.27
MCSE-LINE	0.42	0.42	0.39	0.36	0.40	0.36	0.36	0.34
MCNSE	0.23	0.33	0.29	0.22	0.28	0.32	0.26	0.21
MCNSE-LINE	0.21	0.33	0.28	0.21	0.26	0.30	0.25	0.23
LSRL	0.29	0.34	0.30	0.27	0.31	0.30	0.31	0.20
LSRL-LINE	0.29	0.34	0.30	0.27	0.30	0.32	0.31	0.21
TokenSAR	0.44	0.49	0.45	0.39	0.43	0.43	0.35	0.35
TokenSAR-LINE	0.41	0.49	0.44	0.37	0.43	0.42	0.36	0.39

Table 11: Detailed PRR scores for all methods and their LINE counterparts. Metric: MetricX XXL.

	WMT14				WMT19			
	Cs-En	De-En	Ru-En	Fr-En	De-En	Fi-En	Lt-En	Ru-En
Llama 3.1 8B								
MSP	0.25	0.35	0.41	0.33	0.31	0.04	0.15	0.37
MSP-LINE	0.33	0.38	0.41	0.33	0.37	0.39	0.38	0.34
PPL	0.36	0.35	0.30	0.24	0.33	0.49	0.49	0.27
PPL-LINE	0.42	0.43	0.47	0.35	0.37	0.49	0.48	0.42
MTE	0.40	0.37	0.33	0.30	0.34	**0.51**	**0.53**	0.32
MTE-LINE	**0.48**	**0.47**	**0.53**	**0.42**	**0.44**	0.49	0.52	**0.51**
MCSE	0.20	0.29	0.33	0.29	0.24	-0.04	0.06	0.31
MCSE-LINE	0.26	0.27	0.28	0.24	0.29	0.27	0.28	0.28
MCNSE	0.36	0.34	0.32	0.27	0.33	0.38	0.43	0.32
MCNSE-LINE	0.36	0.35	0.39	0.31	0.35	0.38	0.42	0.40
LSRL	0.34	0.35	0.29	0.27	0.30	0.30	0.32	0.30
LSRL-LINE	0.33	0.31	0.38	0.24	0.25	0.33	0.34	0.33
TokenSAR	0.38	0.34	0.31	0.27	0.32	0.50	0.48	0.27
TokenSAR-LINE	0.43	0.43	0.49	0.34	0.37	0.49	0.48	0.43
Gemma 2 9B								
MSP	0.20	0.35	0.39	0.27	0.34	0.00	0.15	0.35
MSP-LINE	0.29	0.38	0.38	0.29	**0.38**	0.29	0.22	0.33
PPL	0.32	0.34	0.29	0.25	0.33	0.37	0.28	0.27
PPL-LINE	0.33	0.37	0.45	0.30	0.35	0.37	0.28	0.33
MTE	0.35	0.33	0.29	0.25	0.32	**0.42**	0.29	0.27
MTE-LINE	**0.37**	**0.38**	**0.48**	**0.34**	0.37	0.40	0.31	**0.37**
MCSE	0.15	0.30	0.34	0.26	0.27	-0.07	0.09	0.32
MCSE-LINE	0.23	0.32	0.34	0.26	0.33	0.20	0.12	0.32
MCNSE	0.28	0.34	0.33	0.25	0.32	0.29	0.19	0.33
MCNSE-LINE	0.28	0.33	0.40	0.27	0.33	0.29	0.19	0.36
LSRL	0.26	0.30	0.30	0.20	0.28	0.29	0.19	0.27
LSRL-LINE	0.25	0.28	0.38	0.19	0.26	0.32	0.16	0.29
TokenSAR	0.32	0.31	0.28	0.23	0.31	0.40	0.30	0.26
TokenSAR-LINE	0.34	0.36	0.45	0.30	0.35	0.39	**0.32**	0.33
EuroLLM 9B								
MSP	0.13	0.31	0.38	0.21	0.27	-0.03	0.06	0.36
MSP-LINE	0.24	0.38	0.42	0.25	0.38	0.26	0.19	0.37
PPL	0.39	0.40	0.38	0.33	0.43	0.40	0.33	0.31
PPL-LINE	0.41	0.43	0.50	0.37	0.44	0.38	0.34	0.39
MTE	0.43	0.42	0.39	0.35	0.43	**0.45**	**0.39**	0.32
MTE-LINE	**0.46**	**0.46**	**0.51**	**0.42**	**0.47**	0.42	0.38	**0.44**
MCSE	0.19	0.31	0.35	0.26	0.30	-0.01	0.08	0.34
MCSE-LINE	0.31	0.36	0.38	0.33	0.38	0.28	0.20	0.32
MCNSE	0.19	0.28	0.23	0.17	0.26	0.30	0.22	0.21
MCNSE-LINE	0.18	0.28	0.29	0.17	0.25	0.28	0.19	0.26
LSRL	0.23	0.28	0.22	0.23	0.28	0.24	0.18	0.18
LSRL-LINE	0.23	0.28	0.26	0.23	0.28	0.26	0.18	0.22
TokenSAR	0.34	0.38	0.37	0.30	0.37	0.40	0.30	0.34
TokenSAR-LINE	0.34	0.41	0.49	0.32	0.38	0.39	0.30	0.41

Table 12: Detailed PRR scores for all methods and their LINE counterparts. Metric: XComet XXL.

	XSum	GSM8k
	Llama 3.1 8B	
MSP	0.33	0.32
MSP-LINE	0.36	0.33
PPL	0.37	0.30
PPL-LINE	0.37	0.38
MTE	0.36	0.34
MTE-LINE	0.35	**0.40**
MCSE	0.03	0.35
MCSE-LINE	0.04	0.35
MCNSE	0.02	0.34
MCNSE-LINE	0.03	0.36
LSRL	0.09	0.36
LSRL-LINE	0.10	0.36
TokenSAR	**0.37**	0.30
TokenSAR-LINE	0.37	0.38
	Gemma 2 9B	
MSP	0.35	0.30
MSP-LINE	**0.38**	0.30
PPL	0.35	0.25
PPL-LINE	0.37	0.36
MTE	0.33	0.29
MTE-LINE	0.36	0.40
MCSE	0.00	0.39
MCSE-LINE	0.03	**0.40**
MCNSE	0.02	0.36
MCNSE-LINE	0.03	0.37
LSRL	0.04	0.39
LSRL-LINE	0.09	0.39
TokenSAR	0.32	0.24
TokenSAR-LINE	0.33	0.36

Table 13: Detailed PRR scores for all methods and their LINE counterparts. Metrics: AlignScore (XSum) and Accuracy (GSM8k).

D Ablation

D.1 Polynomial Detrending

Tables 14, 15, 16 contain comparison in PRR scores between first, second and third degree LINE correction to the considered base UQ methods.

	WMT14				**WMT19**			
	Cs-En	De-En	Ru-En	Fr-En	De-En	Fi-En	Lt-En	Ru-En
				Llama 3.1 8B				
MSP	0.42	0.39	0.45	0.35	0.46	0.19	0.29	0.43
MSP-LINE$_1$	0.47	0.49	0.48	0.40	0.51	0.47	0.47	0.41
MSP-LINE$_2$	0.50	0.52	0.52	0.41	0.51	0.50	0.47	0.47
MSP-LINE$_3$	0.53	0.53	0.51	0.34	0.53	0.48	0.47	0.48
PPL	0.42	0.46	0.37	0.31	0.41	0.52	0.47	0.32
PPL-LINE$_1$	0.52	0.51	0.53	0.41	0.46	0.52	0.49	0.46
PPL-LINE$_2$	0.42	0.47	0.42	0.32	0.44	0.53	0.48	0.39
PPL-LINE$_3$	0.54	0.52	0.51	0.42	0.48	0.54	0.49	0.45
MTE	0.44	0.48	0.41	0.37	0.42	0.54	0.52	0.33
MTE-LINE$_1$	0.58	0.56	**0.59**	**0.48**	**0.55**	**0.56**	**0.56**	**0.53**
MTE-LINE$_2$	0.46	0.51	0.48	0.38	0.51	0.53	0.53	0.45
MTE-LINE$_3$	**0.59**	**0.57**	0.57	**0.48**	**0.55**	0.55	0.55	0.51
MCSE	0.36	0.32	0.35	0.30	0.36	0.08	0.20	0.36
MCSE-LINE$_1$	0.38	0.36	0.33	0.28	0.38	0.32	0.36	0.32
MCSE-LINE$_2$	0.42	0.40	0.36	0.28	0.39	0.32	0.35	0.32
MCSE-LINE$_3$	0.48	0.40	0.39	0.28	0.42	0.31	0.36	0.38
MCNSE	0.48	0.44	0.40	0.36	0.43	0.46	0.48	0.35
MCNSE-LINE$_1$	0.49	0.44	0.47	0.39	0.45	0.46	0.48	0.44
MCNSE-LINE$_2$	0.42	0.46	0.37	0.30	0.42	0.45	0.47	0.37
MCNSE-LINE$_3$	0.53	0.46	0.46	0.40	0.47	0.47	0.49	0.43
LSRL	0.45	0.44	0.38	0.35	0.46	0.37	0.42	0.35
LSRL-LINE$_1$	0.41	0.41	0.44	0.32	0.40	0.37	0.40	0.38
LSRL-LINE$_2$	0.45	0.44	0.31	0.29	0.42	0.39	0.41	0.31
LSRL-LINE$_3$	0.46	0.38	0.43	0.35	0.44	0.39	0.41	0.37
TokenSAR	0.44	0.45	0.37	0.35	0.40	0.52	0.46	0.32
TokenSAR-LINE$_1$	0.51	0.52	0.52	0.41	0.47	0.53	0.49	0.46
TokenSAR-LINE$_2$	0.39	0.47	0.41	0.31	0.44	0.54	0.47	0.38
TokenSAR-LINE$_3$	0.52	0.47	0.51	0.41	0.48	0.54	0.48	0.45

Table 14: PRR scores with linear and polynomial detrending – Comet WMT22.

	WMT14				WMT19			
	Cs-En	De-En	Ru-En	Fr-En	De-En	Fi-En	Lt-En	Ru-En
				Llama 3.1 8B				
MSP	0.25	0.35	0.41	0.33	0.31	0.04	0.15	0.37
MSP-LINE$_1$	0.33	0.38	0.41	0.33	0.37	0.39	0.38	0.34
MSP-LINE$_2$	0.36	0.43	0.47	0.35	0.37	0.42	0.37	0.39
MSP-LINE$_3$	0.37	0.43	0.46	0.29	0.39	0.39	0.38	0.41
PPL	0.36	0.35	0.30	0.24	0.33	0.49	0.49	0.27
PPL-LINE$_1$	0.42	0.43	0.47	0.35	0.37	0.49	0.48	0.42
PPL-LINE$_2$	0.34	0.37	0.35	0.26	0.34	0.50	0.45	0.32
PPL-LINE$_3$	0.43	0.44	0.45	0.35	0.38	0.49	0.46	0.39
MTE	0.40	0.37	0.33	0.30	0.34	**0.51**	**0.53**	0.32
MTE-LINE$_1$	**0.48**	0.47	**0.53**	**0.42**	**0.44**	0.49	0.52	**0.51**
MTE-LINE$_2$	0.38	0.40	0.38	0.33	0.39	0.46	0.48	0.39
MTE-LINE$_3$	0.46	**0.48**	0.48	0.41	0.43	0.46	0.49	0.45
MCSE	0.20	0.29	0.33	0.29	0.24	-0.04	0.06	0.31
MCSE-LINE$_1$	0.26	0.27	0.28	0.24	0.29	0.27	0.28	0.28
MCSE-LINE$_2$	0.30	0.33	0.31	0.24	0.28	0.28	0.26	0.29
MCSE-LINE$_3$	0.32	0.33	0.34	0.24	0.30	0.24	0.27	0.33
MCNSE	0.36	0.34	0.32	0.27	0.33	0.38	0.43	0.32
MCNSE-LINE$_1$	0.36	0.35	0.39	0.31	0.35	0.38	0.42	0.40
MCNSE-LINE$_2$	0.31	0.39	0.27	0.23	0.31	0.38	0.41	0.31
MCNSE-LINE$_3$	0.39	0.39	0.36	0.32	0.36	0.38	0.42	0.38
LSRL	0.34	0.35	0.29	0.27	0.30	0.30	0.32	0.30
LSRL-LINE$_1$	0.33	0.31	0.38	0.24	0.25	0.33	0.34	0.33
LSRL-LINE$_2$	0.35	0.35	0.20	0.21	0.27	0.35	0.35	0.25
LSRL-LINE$_3$	0.35	0.27	0.32	0.27	0.28	0.34	0.35	0.32
TokenSAR	0.38	0.34	0.31	0.27	0.32	0.50	0.48	0.27
TokenSAR-LINE$_1$	0.43	0.43	0.49	0.34	0.37	0.49	0.48	0.43
TokenSAR-LINE$_2$	0.33	0.36	0.36	0.25	0.33	0.50	0.44	0.33
TokenSAR-LINE$_3$	0.41	0.36	0.46	0.34	0.38	0.49	0.45	0.39

Table 15: PRR scores with linear and polynomial detrending – XComet XXL.

	WMT14				WMT19			
	Cs-En	De-En	Ru-En	Fr-En	De-En	Fi-En	Lt-En	Ru-En
				Llama 3.1 8B				
MSP	0.21	0.22	0.31	0.19	0.23	0.08	0.12	0.26
MSP-LINE$_1$	0.40	0.42	0.44	0.31	0.40	0.41	0.38	0.35
MSP-LINE$_2$	0.42	0.44	0.46	0.32	0.40	0.44	0.38	0.39
MSP-LINE$_3$	0.43	0.44	0.46	0.29	0.41	0.41	0.38	0.39
PPL	0.43	0.45	0.41	0.32	0.39	0.48	0.43	0.31
PPL-LINE$_1$	0.48	0.47	0.49	0.36	0.40	0.48	0.44	0.40
PPL-LINE$_2$	0.40	0.46	0.41	0.29	0.40	0.49	0.42	0.35
PPL-LINE$_3$	0.49	0.48	0.48	0.35	0.41	0.48	0.45	0.39
MTE	0.47	0.48	0.46	0.39	0.43	**0.52**	0.49	0.36
MTE-LINE$_1$	**0.54**	0.51	**0.54**	**0.43**	**0.47**	0.51	0.49	**0.45**
MTE-LINE$_2$	0.45	0.50	0.47	0.36	0.46	0.49	0.47	0.40
MTE-LINE$_3$	0.53	**0.52**	**0.54**	0.42	**0.47**	0.49	**0.50**	**0.45**
MCSE	0.16	0.14	0.20	0.16	0.14	-0.00	0.07	0.21
MCSE-LINE$_1$	0.29	0.29	0.30	0.25	0.32	0.29	0.29	0.27
MCSE-LINE$_2$	0.33	0.31	0.33	0.25	0.31	0.30	0.28	0.28
MCSE-LINE$_3$	0.36	0.30	0.36	0.25	0.32	0.28	0.29	0.31
MCNSE	0.42	0.38	0.40	0.32	0.38	0.39	0.44	0.32
MCNSE-LINE$_1$	0.42	0.38	0.43	0.34	0.39	0.39	0.44	0.36
MCNSE-LINE$_2$	0.38	0.39	0.35	0.28	0.38	0.39	0.43	0.31
MCNSE-LINE$_3$	0.45	0.39	0.43	0.34	0.39	0.40	0.46	0.35
LSRL	0.39	0.35	0.37	0.30	0.36	0.32	0.42	0.31
LSRL-LINE$_1$	0.38	0.35	0.38	0.29	0.35	0.33	0.40	0.33
LSRL-LINE$_2$	0.40	0.36	0.28	0.27	0.35	0.35	0.42	0.28
LSRL-LINE$_3$	0.40	0.34	0.38	0.31	0.35	0.34	0.42	0.32
TokenSAR	0.46	0.45	0.41	0.34	0.39	0.49	0.44	0.32
TokenSAR-LINE$_1$	0.49	0.47	0.48	0.36	0.41	0.49	0.44	0.40
TokenSAR-LINE$_2$	0.40	0.45	0.41	0.29	0.40	0.50	0.42	0.35
TokenSAR-LINE$_3$	0.48	0.45	0.48	0.35	0.41	0.49	0.46	0.39

Table 16: PRR scores with linear and polynomial detrending – MetricX XXL.

D.2 Reducing number of quality labels

Obtaining quality labels can be expensive in certain setups. To address this issue, we estimate the quality (metric)-vs–length regression (equation (3)) using a *small, length-balanced* subset of the training data, rather than the full sample. The goal is to recover an effective quality trend with far fewer labels.

We first remove length outliers by keeping the 5th–95th percentiles and rescale lengths to $[0, 1]$. The length axis is then partitioned into n adaptive bins via K-means (narrower in dense regions, wider in sparse ones). From each bin we sample about S/n items without replacement, where S is the target labeled size; bins with fewer items contribute all their points, and the shortfall is redistributed across the others. Finally, we fit a metric–vs–length regression on this subset and apply it to remove the length trend at test time.

	GSM8K	**XSum**
	Llama 3.1 8B	
MSP	0.32	0.33
MSP-LINE (500 sample)	0.33	0.36
MSP-LINE (Full sample)	0.33	0.36
PPL	0.30	**0.37**
PPL-LINE (500 sample)	0.38	<u>0.37</u>
PPL-LINE (Full sample)	0.38	<u>0.37</u>
MTE	0.34	0.36
MTE-LINE (500 sample)	<u>0.39</u>	0.35
MTE-LINE (Full sample)	**0.40**	0.35
MCSE	0.35	0.03
MCSE-LINE (500 sample)	0.35	0.04
MCSE-LINE (Full sample)	0.35	0.04
MCNSE	0.34	0.02
MCNSE-LINE (500 sample)	0.36	0.03
MCNSE-LINE (Full sample)	0.36	0.03
LSRL	0.36	0.09
LSRL-LINE (500 sample)	0.36	0.10
LSRL-LINE (Full sample)	0.36	0.10
TokenSAR	0.30	0.37
TokenSAR-LINE (500 sample)	0.38	0.36
TokenSAR-LINE (Full sample)	0.38	0.37
	Gemma 2 9B	
MSP	0.30	0.35
MSP-LINE (500 sample)	0.30	**0.38**
MSP-LINE (Full sample)	0.30	**0.38**
PPL	0.25	0.35
PPL-LINE (500 sample)	0.36	0.37
PPL-LINE (Full sample)	0.36	<u>0.37</u>
MTE	0.29	0.33
MTE-LINE (500 sample)	**0.40**	0.36
MTE-LINE (Full sample)	**0.40**	0.36
MCSE	0.39	0.00
MCSE-LINE (500 sample)	<u>0.40</u>	0.03
MCSE-LINE (Full sample)	**0.40**	0.03
MCNSE	0.36	0.02
MCNSE-LINE (500 sample)	0.37	0.03
MCNSE-LINE (Full sample)	0.37	0.03
LSRL	0.39	0.04
LSRL-LINE (500 sample)	0.39	0.09
LSRL-LINE (Full sample)	0.39	0.09
TokenSAR	0.24	0.32
TokenSAR-LINE (500 sample)	0.35	0.33
TokenSAR-LINE (Full sample)	0.36	0.33

Table 17: PRR scores on GSM8K and XSum using 500 samples for quality trend fitting.