VISUALIZING UNCERTAINTY IN TRANSLATION TASKS: AN EVALUATION OF LLM PERFORMANCE AND CONFIDENCE METRICS

Jin Hyun Park
Dept. of Computer Science and Engineering
Texas A&M University
College Station, TX 77843-3112
jinhyun.park@tamu.edu

Utsawb Laminchhane
Dept. of Computer Science and Engineering
Texas A&M University
College Station, TX 77843-3112
utsawblamichhane@tamu.edu

Umer Farooq
Dept. of Multidisciplinary Engineering
Texas A&M University
College Station, TX 77843-3112
umerfarooq@tamu.edu

Uma Sivakumar
Dept. of Computer Science and Engineering
Texas A&M University
College Station, TX 77843-3112
umas0697@tamu.edu

Arpan Kumar
Dept. of Computer Science and Engineering
Texas A&M University
College Station, TX 77843-3112
arpsku17@tamu.edu

ABSTRACT

Large language models (LLMs) are increasingly utilized for machine translation, yet their predictions often exhibit uncertainties that hinder interpretability and user trust. Effectively visualizing these uncertainties can enhance the usability of LLM outputs, particularly in contexts where translation accuracy is critical. This paper addresses two primary objectives: (1) providing users with token-level insights into model confidence and (2) developing a web-based visualization tool to quantify and represent translation uncertainties. To achieve these goals, we utilized the T5 model with the WMT19 dataset for translation tasks and evaluated translation quality using established metrics such as BLEU, METEOR, and ROUGE. We introduced three novel uncertainty quantification (UQ) metrics: (1) the geometric mean of token probabilities, (2) the arithmetic mean of token probabilities, and (3) the arithmetic mean of the kurtosis of token distributions. These metrics provide a simple yet effective framework for evaluating translation performance. Our analysis revealed a linear relationship between the traditional evaluation metrics and our UQ metrics, demonstrating the validity of our approach. Additionally, we developed an interactive web-based visualization that uses a color gradient to represent token confidence. This tool offers users a clear and intuitive understanding of translation quality while providing valuable insights into model performance. Overall, we show that our UQ metrics and visualization are both robust and interpretable, offering practical tools for evaluating and accessing machine translation systems.

1 Introduction

The rapid advancement of large language models (LLMs) has significantly improved machine translation, providing tools for handling various languages and contexts [1, 2]. Despite these advancements, LLM-generated translations often need more clarity, particularly when translating bigger or more complex sentences [3]. These uncertainties pose challenges for users who rely on translation models for critical tasks, where understanding the model's confidence in

its translations is crucial [4]. Without clear indications of uncertainty, users may struggle to gauge the reliability of a translation, potentially leading to misinterpretation or misuse in high-stakes environments.

In machine translation, commonly used metrics such as BLEU [5] and ROUGE [6] provide valuable insights into translation quality. These metrics primarily assess n-gram overlap (i.e., precision), making them straightforward to interpret. Additionally, they have been shown to generally correlate with human judgment when averaged across a corpus of sentences [7]. Similarly, METEOR [8], another popular metric, is recognized for its ability to capture semantic and linguistic nuances more effectively than BLEU and ROUGE. By incorporating resources like WordNet [9] and prioritizing recall over precision, METEOR emphasizes semantic meaning. However, like the other metrics, it primarily evaluates overall alignment and does not provide token-level confidence.

Traditional translation quality metrics overlook the probabilistic uncertainty in LLM outputs. However, the probabilistic uncertainty is crucial for applications requiring explicit communication of accuracy and reliability. Quantifying and visualizing these uncertainties can significantly enhance user interpretability and decision-making, particularly in scenarios that demand precise and confident translations.

To address this gap, this study introduces a method for quantifying and visualizing uncertainties in LLM-generated translations. Uncertainty is quantified using token-level probabilities and token-level kurtosis derived from the model's internal outputs. Specifically, we define three measures to assess uncertainty: (1) the geometric mean of token probabilities, (2) the arithmetic mean of token probabilities, and (3) the arithmetic mean of the scaled-kurtosis of top-k tokens' distributions. These metrics collectively provide a comprehensive view of both overall and localized confidence within translations, enabling a deeper understanding of uncertainty. For our experiments, we used the WMT dataset with T5 [10] models of varying sizes - small, base, and large. Similar to the T5 paper, our study focuses on three language translation tasks: English to German, English to French, and English to Romanian, using the WMT dataset only. We used the same dataset and model to develop an uncertainty visualization tool. This tool is implemented as a web-based interface where output tokens are color-coded with varying gradients to represent the confidence level of each token given an input sentence.

The following research question is guiding this study:

- How can uncertainty in large language model-generated translations be effectively quantified and visualized to enhance user interpretability and decision-making across critical applications?

2 Literature Review

Predictive models, especially those based on deep learning and large language models (LLMs), are widely used across domains, ranging from healthcare [11] to natural language processing [12] and social networks [13]. A key challenge in deploying these models is managing and communicating the uncertainty inherent in their predictions.

In neural networks, *deep ensembles* estimate uncertainty by generating predictions from multiple independently trained models, with variability indicating uncertainty in regions of disagreement. Monte Carlo Dropout (MC-Dropout) estimates uncertainty by running multiple forwards passes with dropout layers enabled during inference, calculating variance in predictions. A study by Dutta et al. [14] compares these methods, finding *deep ensembles* to be more accurate, while MC-Dropout is competitive with faster inference times. Visualization techniques like Parallel Coordinates Plots (PCP) and heatmaps were used to illustrate uncertainty and model error.

In medical applications, where incorrect predictions can have serious consequences, communicating uncertainty is vital [15]. The authors emphasize the importance of quantifying and effectively conveying uncertainty in clinical machine-learning models. Furthermore, explicitly expressing uncertainty through phrases like "I'm not sure" can greatly enhance trust between clinicians and machine learning systems.

The impact of uncertainty visualization on user reliance has also been explored [16]. The paper found that users tended to trust model outputs more when uncertainty was visualized, particularly in high-stakes or difficult decision-making tasks. Visualization techniques, such as ordinal expressions of uncertainty (e.g., "low," "medium," "high"), were shown to help users make more calibrated decisions.

Similarly, a tool has been developed to visualize uncertainties and errors in multimodal AI systems [17]. The tool provides an interface for error analysis, enabling users to understand how a model processes and integrates different types of data (e.g., text and images) and where uncertainties arise. This approach is particularly useful in complex AI systems that operate across multiple domains, offering a holistic view of uncertainty across modalities.

With the rise of LLMs, there is growing concern about the generation of hallucinations - false or inaccurate information presented as credible [18]. The authors examined how such hallucinations propagate through social networks such

as Facebook and how uncertainty in these hallucinations can be measured. They also worked on the methods for quantifying the credibility of hallucinations based on the task type (e.g., question-answering, dialogue). Likewise, a specific type of uncertainty related to knowledge errors exists in LLMs [19]. They propose a method that encourages LLMs to self-reflect and express self-doubt, allowing users to better understand when the model is uncertain about its knowledge.

3 Methodology

3.1 Objective and Approach

This study aims to develop a visualization approach to effectively represent uncertainties in using LLMs for translation tasks. To accomplish this, we quantify uncertainties using custom UQ metrics derived from token-level probabilities. We also explore the relationship between these UQ metrics and established translation evaluation metrics, including BLEU, METEOR, and ROUGE. Additionally, we created a web-based application to visualize token-level uncertainties effectively. These visualizations will help users understand where and to what extent the model's outputs may deviate from the ground truth, offering a more interpretable and user-centric interface for translation tasks. We evaluated our approach using various T5 model configurations, including small, base, and large. The sizes of these models are detailed in Table 1.

Model	Vocabulary Size	Parameter Size
T5-small	32,100	60M
T5-base	32,100	223M
T5-large	32,100	738M

Table 1: Model Comparison: Vocabulary and Parameter Sizes for T5 Variants

3.2 Dataset

We use the VMT14, 15, and 16 datasets [20–22], a well-established translation task datasets containing a diverse array of multilingual sentence pairs. These datasets contain sentence pairs in English with translations into German, French, and Romanian, which serve as ground truth for our evaluation. They have been preprocessed to ensure compatibility with our metrics-based analysis.

3.3 Traditional Metrics

To assess the quality of the model's translations, we employ three widely recognized machine translation metrics:

- **BLEU**: BLEU stands for Bilingual Evaluation Understudy Score. It measures the precision of n-grams between the predicted translation and the ground truth, focusing on token choice and order accuracy [8]. The BLEU score formula is defined by the following,

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

where

1. p_n denotes the precision of n-grams, defined as the ratio of n-grams in the predicted translation that matches the ground truth to the total number of n-grams in the predicted translation.

2. w_n represents the weights assigned to each n-gram precision score, often set equally (e.g., $w_n = \frac{1}{N}$ for a uniform weight).

3. BP is the brevity penalty, calculated to adjust for length mismatches between the predicted and reference translations. It is defined as

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r \end{cases}$$

where c is the length of the predicted translation, and r is the length of the reference translation. The brevity penalty discourages excessively short translations, which might otherwise artificially inflate precision. The BLEU metric aggregates these elements to provide a score that balances accuracy across n-grams of varying lengths.

- **METEOR**: METEOR stands for Metric for Evaluation of Translation with Explicit ORdering. It emphasizes recall over precision by incorporating synonyms and stemming, yielding a refined semantic similarity measure. The METEOR score is computed by matching segments (note that *segment* differs from *token*) based on exact, stem, synonym, and paraphrase matches [23] and is represented as,

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty}) \tag{2}$$

where

1. F_{mean} is the harmonic mean of precision (P) and recall (R), calculated as

$$F_{\text{mean}} = \frac{10PR}{R + 9P}$$

where P represents the proportion of matched segments in the predicted translation to the total segments in the prediction, and R represents the proportion of matched segments in the predicted translation to the total segments in the reference.

2. Penalty discourages fragmented matches (i.e., multiple short matches over longer, more cohesive segments). The penalty is calculated as follows:

$$\text{Penalty} = \gamma \cdot \left(\frac{\text{Chunks}}{\text{Matches}} \right)^{\beta}$$

where Chunks is the number of contiguous matched segments, Matches is the total number of matched segments, and γ and β are empirically set constants. The penalty reduces the score when the matches are dispersed, reflecting a preference for translations that maintain cohesive segments.

- **ROUGE**: ROUGE is a recall-focused metric commonly used in summarization tasks, which evaluates the overlap of n-grams or longest common subsequences between the predicted and reference translations. We utilize three ROUGE variants in this study:

1. ROUGE-1: Calculates the overlap of uni-grams (single words) between the predicted and reference translations. It is defined as

$$\text{ROUGE-1} = \frac{\sum_{\text{uni-grams} \in \text{Reference}} \min\left(\text{Count}_{\text{match}}(\text{uni-gram})\right)}{\sum_{\text{uni-grams} \in \text{Reference}} \text{Count}(\text{uni-gram})}$$

2. ROUGE-2: Measures the overlap of bi-grams (pairs of consecutive words) between the predicted and reference translations. This metric is more sensitive to word order and context. It is given by

$$\text{ROUGE-2} = \frac{\sum_{\text{bi-grams} \in \text{Reference}} \min\left(\text{Count}_{\text{match}}(\text{bi-gram})\right)}{\sum_{\text{bi-grams} \in \text{Reference}} \text{Count}(\text{bi-gram})}$$

3. ROUGE-L: Based on the Longest Common Subsequence (LCS), ROUGE-L evaluates fluency and sentence structure by assessing the longest sequence of tokens that appear in the same order in both the prediction and reference. ROUGE-L is calculated as

$$\text{ROUGE-L} = \frac{\text{LCS}(\text{Reference}, \text{Hypothesis})}{\text{Length of Reference}}$$

where $\text{LCS}(\text{Reference}, \text{Hypothesis})$ represents the length of the longest common subsequence between the reference and the hypothesis (or predicted translation). This variant captures structural similarity, making it suitable for assessing overall fluency and coherence.

Each of these metrics provides a distinct perspective on translation quality. BLEU focuses on precision, METEOR emphasizes semantic and recall-focused similarity, and ROUGE is oriented toward structural and contextual overlap, particularly suited for evaluating summarization and coherence. These metrics are applied to translations generated by the LLM to quantify performance relative to the ground truth.

3.4 Uncertainty Quantification (UQ) Metrics

In addition to traditional evaluation metrics, we incorporate Uncertainty Quantification (UQ) metrics to capture the model's confidence at the token level. For each translation generated by the LLMs - T5 small, base, and large - we store token probability matrices representing the model's likelihood for each token within the translated sentence. These enable a probabilistic understanding of each translation output.

To quantify uncertainty, we calculate the UQ metrics using three key measures based on probability distributions:

1. Geometric Mean of Token Probabilities: This measure calculates the geometric mean of the probabilities across tokens in a translation, providing a multiplicative aggregation of token-level confidences. This metric is calculated as:

$$\text{GT} = \left(\prod_{i=1}^{L} p(\text{token}_i) \right)^{\frac{1}{L}} \tag{3}$$

   where L is the total number of tokens in the translated sentence, and $p(\text{token}_i)$ represents the probability of the i-th token. The geometric mean helps capture an overall confidence score emphasizing lower probabilities. Hence reflecting conservative confidence in the translation.

2. Arithmetic Mean of Token Probabilities: The arithmetic mean provides an additive aggregation of token probabilities. It offers an average confidence measure across all tokens. This metric is given by:

$$\text{AT} = \frac{1}{L} \sum_{i=1}^{L} p(\text{token}_i) \tag{4}$$

   This equation averages out the probability of each token in the sequence, giving an overall score that treats each token's confidence equally. It offers a balanced view of token-level certainty in the translation.

3. Arithmetic Mean of scaled kurtosis of token probability: To incorporate higher-order statistical moments, we compute the kurtosis over the top *k* most probable token probabilities for each token. Specifically, after obtaining the probability distribution for a token, we sort the probabilities in descending order and select the top 1,000 values. We then calculate the kurtosis for each token, resulting in an array of kurtosis values corresponding to the number of tokens in the sentence. To standardize these values, we apply min-max scaling to the kurtosis of each token within the sentence, a step that defines the term *scaled* kurtosis. Finally, we average the scaled kurtosis values to derive a single representative value for the sentence. Kurtosis in this context captures the *peakedness* and tail weight of the probability distribution for each token.

$$\text{AK} = \mathbb{E}_{i \in \text{top-}k} \left[\text{scaled-kurt}(p(\text{token}_i)) \right] \tag{5}$$

$$= \frac{1}{k} \sum_{i=1}^{k} \text{scaled-kurt}(p(\text{token}_i)) \tag{6}$$

   This measure captures confidence by emphasizing tokens with more extreme probability distributions and provides a refined view of translation certainty. We used this metric because we empirically observed that kurtosis and probability exhibit similar distributions within a sentence.

   A regression analysis between token probabilities and scaled kurtosis yielded a high R^2, indicating that kurtosis is a strong statistical measure of uncertainty (see Fig. 1). This trend was consistent across all sentences analyzed. The observed linear relationship between kurtosis and token probabilities supports the use of kurtosis as a central metric for uncertainty quantification.

These three measures, taken together, enable a better understanding of uncertainty in model predictions during translation. By visualizing these metrics alongside quality scores, we can contextualize BLEU, METEOR, and ROUGE results and offer users an intuitive representation of translation reliability.

3.5 Visualizatation

To enhance the interpretability of translation tasks and provide insights into model confidence, we developed an interactive live demo application. Built with a web-based front end and a Python-based Flask backend. Users can select from different T5 model variants (small or base) and language pairs (e.g., English to German, French, or Romanian) to observe how the model handles various translation challenges. When a sentence is submitted for translation, the server processes the input, performs the translation using the selected model, and computes token-level uncertainty metrics. These metrics are visualized on the client side using a gradient-based color scheme, where high-confidence tokens are shown in cooler colors and low-confidence tokens in warmer shades. The gradient-based color scheme is determined by token probabilities.

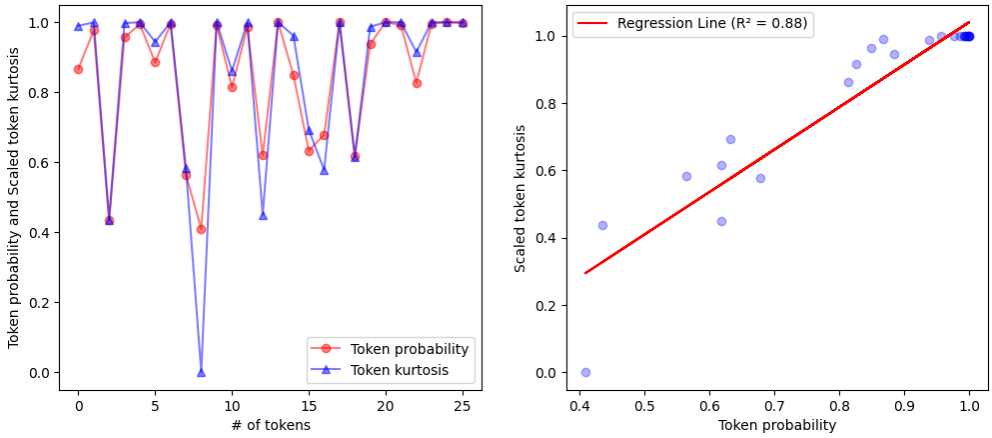


Figure 1: Token Probabilities and Scaled Kurtosis Analysis for English-to-German (EN-DE) task, showing the linear relationship and R^2 value of 0.88.

4 Results

In this study, we evaluated the performance of T5 small, base, and large on translation tasks from English to three target languages. The evaluation utilized traditional metrics and UQ metrics derived from token probabilities. We also added the BERT F1 score to test semantic similarity. We also provide visualization results for a web-based program. The results are summarized as follows:

4.1 Traditional vs Uncertainty Quantification Metrics

Fig 2 illustrates each metric's average values and standard deviations across the three translation tasks. These figures demonstrate the superior performance and reduced uncertainty associated with the T5-base, which strikes a balance between size and translation quality. Across all UQ metrics (Confidence-A-K, Confidence-A-T, and Confidence-G-T), the T5 base model either outperforms or performs on par with both T5 small and T5 large. This is a notable finding, as traditional metrics typically show that larger models deliver better performance (e.g., in EN-DE and EN-RO). This suggests that our proposed uncertainty quantification offers unique insights: increasing the number of parameters does not necessarily result in higher confidence when selecting tokens.

One limitation of UQ metrics is their tendency to show small standard deviations, with average values typically falling within the 0.7 to 0.9 range. The A-K metric behaves somewhat differently, as its values have been normalized and averaged (See Fig. 2, EN-RO). However, the saturation of UQ metric values at higher levels makes interpreting the results challenging. To address this, a similar normalization and averaging approach, as used in the A-K method, could be applied to A-T and G-T metrics for improved interpretability.

4.2 Visualizations

Fig. 3 and 4 illustrate the translation quality results for uncertainty evaluation using the same sentence. In both samples, The T5 base model demonstrates superior performance compared to the T5 small across both A-T and G-T UQ metrics. Additionally, the gradient-based color scheme visually represents token confidence. The complete code for this work is available at `https://github.com/7201krap/CSCE679_Data-Visualization/tree/main/live_demo`

We observe that T5 small and T5 base produce different outputs or probabilities when given the same input. In Fig. 3, the outputs are identical, but their probabilities differ, which is represented by variations in color. In Fig. 4, the outputs differ, with T5-base producing sentences more likely to be in German and showing higher UQ values (i.e., high confidence).

- (T5-small) Translated German: Was ist lhr Haupt?
- (T5-small) Translated English using Google Translate: What is your head?
- (T5-base) Translated German: Was ist de in Studien schwer punkt?
- (T5-base) Translated English using Google Translate: What is the focus of studies?

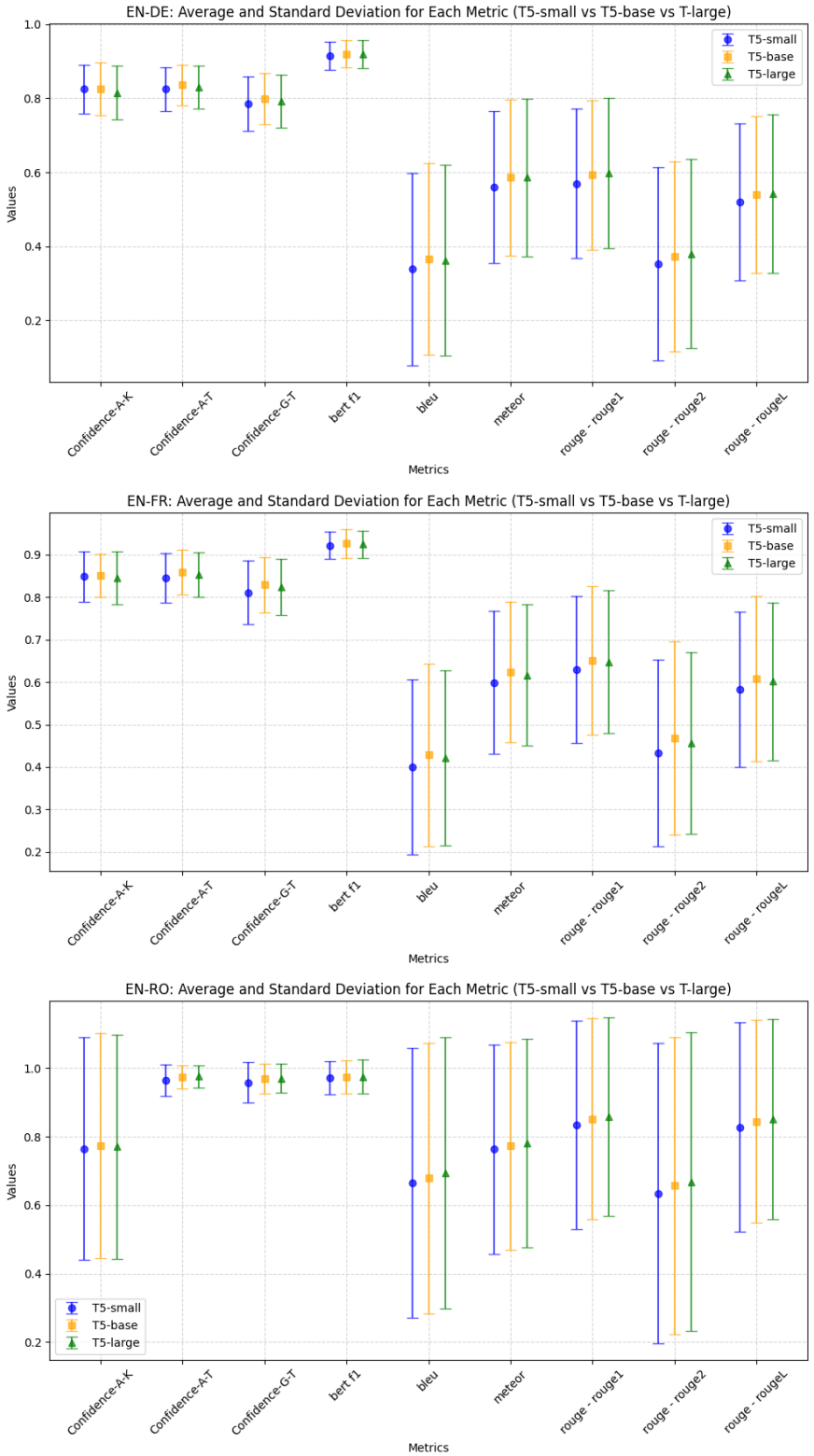


Figure 2: Translation Performance for English-to-German (Top), French (Middle), and Romanian (Bottom) tasks, showing traditional and UQ metrics for T5 models.

Figure 3: Example 1. Live Demo Application with translation UQ colored

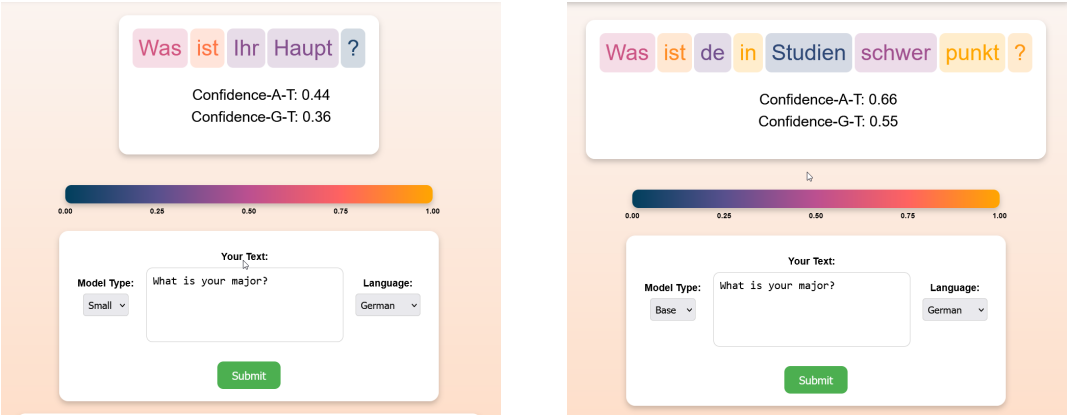
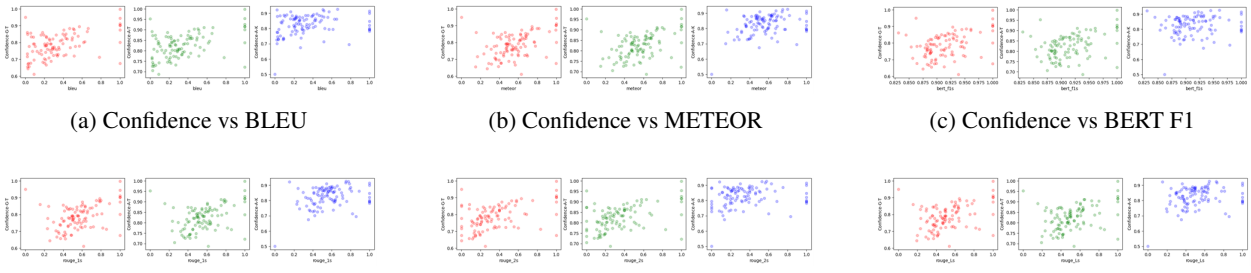


Figure 4: Example 2. Live Demo Application with translation UQ colored

5 Discussion

The findings of this study show the complex interplay between model size, translation quality, and uncertainty quantification (UQ) metrics, offering valuable insights into the performance of T5 variants across diverse evaluation criteria. While some may critique our UQ metrics as being overly simplistic and insufficient for capturing semantic meaning compared to traditional metrics, we demonstrate that our proposed UQ metrics align closely with established traditional measures, reinforcing their validity.

As illustrated in Fig. 5, a linear relationship is observed between our defined UQ metrics and traditional metrics. While the alignment is not perfect, this relationship supports the validity of our proposed metrics and analysis.



(a) Confidence vs BLEU (b) Confidence vs METEOR (c) Confidence vs BERT F1

(d) Confidence vs ROUGE-1 (e) Confidence vs ROUGE-2 (f) Confidence vs ROUGE-LCS

Figure 5: Scatterplots of Confidence Metrics vs. Traditional Metrics (BLEU, METEOR, ROUGE, BERT F1, Kurtosis, and Combined Metrics) highlighting alignment and deviations in model performance.

6 Conclusion

In this work, we proposed three novel uncertainty metrics based on token probability and kurtosis (leveraging the top-k most probable token candidates). Additionally, we developed an interactive web-based program to visualize token probabilities and confidence levels. Using T5 small, base, and large models, we conducted experiments on translation tasks across three language pairs: English-to-German, English-to-French, and English-to-Romanian.

Our analysis demonstrated that the proposed UQ metrics offer valuable new insights compared to traditional translation quality evaluation metrics. We also observed a linear relationship between our UQ metrics and their traditional counterparts, further validating their relevance. By visualizing token probabilities interactively, we enhanced user understanding and provided a more accessible and practical interface for analyzing LLM outputs. This approach offers a promising framework for more transparent and interpretable language model applications.

7 Future Works

The findings emphasize the importance of balancing model architecture and size to meet specific application needs. Integrating UQ metrics into standard evaluation pipelines can enhance the assessment of translation systems, promoting greater transparency and trust in LLM-generated translations.

This study also highlights the value of combining traditional metrics with advanced UQ techniques and visualizations to create translation systems that are both accurate and interpretable. Future research could expand this approach to other language pairs or explore alternative UQ metrics, with a particular focus on low-resource languages. Such efforts could reveal how uncertainty quantification and visualization adapt to diverse linguistic structures and varying data availability.

References

[1] Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. Transllama: Llm-based simultaneous translation system. *arXiv preprint arXiv:2402.04636*, 2024.

[2] Hui Huang, Shuangzhi Wu, Xinnian Liang, Bing Wang, Yanrui Shi, Peihao Wu, Muyun Yang, and Tiejun Zhao. Towards making the most of llm for translation quality estimation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 375–386. Springer, 2023.

[3] David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. The fine-tuning paradox: Boosting translation quality without sacrificing llm abilities. *arXiv preprint arXiv:2405.20089*, 2024.

[4] Yuvraj Virk, Premkumar Devanbu, and Toufique Ahmed. Enhancing trust in llm-generated code summaries with calibrated confidence scores. *arXiv preprint arXiv:2404.19318*, 2024.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[6] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[7] Amirkeivan Mohtashami, Mauro Verzetti, and Paul K Rubenstein. Learning translation quality evaluation on low resource languages from large language models. *arXiv preprint arXiv:2302.03491*, 2023.

[8] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[9] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[10] Wikimedia Foundation. ACL 2019 Fourth Conference on Machine Translation (WMT19), Shared Task: Machine Translation of News. Online, 2019. URL `http://www.statmt.org/wmt19/translation-task.html`. Accessed: [Insert Date of Access].

[11] Hamdan O Alanazi, Abdul Hanan Abdullah, and Kashif Naseer Qureshi. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. *Journal of medical systems*, 41:1–10, 2017.

[12] Ben J Marafino, Miran Park, Jason M Davies, Robert Thombley, Harold S Luft, David C Sing, Dhruv S Kazi, Colette DeJong, W John Boscardin, Mitzi L Dean, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. *JAMA network open*, 1(8):e185097–e185097, 2018.

[13] Adrien Guille and Hakim Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 1145–1152, 2012.

[14] Soumya Dutta, Faheem Nizar, Ahmad Amaan, and Ayan Acharya. Visual analysis of prediction uncertainty in neural networks for deep image synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[15] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

[16] Jieqiong Zhao, Yixuan Wang, Michelle V Mancenido, Erin K Chiou, and Ross Maciejewski. Evaluating the impact of uncertainty visualization on model reliance. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[17] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multiviz: Towards visualizing and understanding multimodal models. *arXiv preprint arXiv:2207.00056*, 2022.

[18] Guozhi Hao, Jun Wu, Qianqian Pan, and Rosario Morello. Quantifying the uncertainty of llm hallucination spreading in complex adaptive social networks. *Scientific reports*, 14(1):16375, 2024.

[19] Wenyuan Zhang, Jiawei Sheng, Shuaiyi Nie, Zefeng Zhang, Xinghua Zhang, Yongquan He, and Tingwen Liu. Revealing the challenge of detecting character knowledge errors in llm role-playing. *arXiv preprint arXiv:2409.11726*, 2024.

[20] 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. Findings of the 2014 workshop on statistical machine translation. In Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors, *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3302. URL `https://aclanthology.org/W14-3302`.

[21] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 workshop on statistical machine translation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3001. URL `https://aclanthology.org/W15-3001`.

[22] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*, pages 131–198. Association for Computational Linguistics, 2016.

[23] Max Grusky. Rogue scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934, 2023.