The Benefits of Being Uncertain: Perplexity as a Signal for Naturalness in Multilingual Machine Translation

Timothy Pistotti[1,2] **Michael Witbrock**[2]
Padriac Amato Tahua O'Leary[2] **Jason Brown**[1]

[1]School of Cultures, Languages and Linguistics
[2]School of Computer Science
University of Auckland
{timothy.pistotti, m.witbrock, padriac.oleary, jason.brown}@auckland.ac.nz

Abstract

Model-internal uncertainty metrics like perplexity potentially offer low-cost signals for Machine Translation Quality Estimation (TQE). This paper analyses perplexity in the "No Language Left Behind" (NLLB) multilingual model. We quantify a significant model-human perplexity gap, where the model is consistently more confident in its own, often literal, machine-generated translation than in diverse, high-quality human versions. We then demonstrate that the utility of perplexity as a TQE signal is highly context-dependent, being strongest for low-resource pairs. Finally, we present an illustrative case study where a flawed translation is refined by providing potentially useful information in a targeted prompt, simulating a knowledge-based repair. We show that as the translation's quality and naturalness improve (a +0.15 COMET score increase), its perplexity also increases, challenging the simple assumption that lower perplexity indicates higher quality and motivating a more nuanced view of uncertainty as signalling a text's departure from rigid translationese.

1 Introduction

Translation Quality Estimation (TQE) is critical for machine translation (MT) system deployment, and is particularly challenging for low-resource languages. Whereas reference-based evaluation metrics directly compare gold-standard human-generated translations and model-generated translations, TQE aims to assess the quality of translations without such references. This paper employs a glass-box approach to investigate a two-stage goal: first, whether a model's internal uncertainty, measured by perplexity (PPL), can serve as a lightweight signal to detect likely errors, and second, whether those errors can then be repaired using knowledge-guided prompting.

One challenge to using perplexity for TQE is that autoregressive models prefer their own output distributions. Our first contribution is to provide an empirical quantification of this phenomenon in a massively multilingual setting. We measure this model-human perplexity gap within the NLLB-200-3.3B model (Costa-Jussà et al., 2022), confirming that the model is systematically less perplexed by its own text than by professional human translations. Our findings show that perplexity often measures conformity to translationese rather than actual translation quality. This finding motivates a more nuanced question that we explore: beyond simple quality, can perplexity serve as a signal for a translation's naturalness? This paper investigates the possibility that as a translation moves from literal translationese towards more fluid, human-like language, its perplexity, as judged by the original model, might paradoxically increase.

Finally, we conclude with a case study demonstrating the *detect-and-repair* workflow that motivates our research. We show how a lexical error, once identified, can be repaired using an instruction-tuned model guided by external knowledge, providing initial evidence for our naturalness hypothesis.

2 Related Work

Our work builds on research into model uncertainty for TQE, model artifacts like translationese, and LLM-based refinement. Using model-internal probabilities for reference-free TQE is a well-established approach (Fomicheva et al., 2020). However, optimizing solely for low perplexity can make models less human-like, as their surprisal patterns diverge from human reading patterns on key syntactic structures (Kuribayashi et al., 2021). Our work quantifies how this sensitivity is modulated by language resource levels. We also connect this to the challenge of translationese—the distinct statistical characteristics of translated text, a known issue in MT evaluation (Zhang and Toral, 2019), and

Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 61–65

contribute by measuring the model's preference for this form of language across a multilingual setting.

3 Method

Our primary analysis centres on the NLLB-200-3.3B model. For our refinement case study, we use the Llama 3 7B Instruct model (Dubey et al., 2024), since the NLLB models are not instruction-tuned and are therefore unsuited for the open-ended editing task required by our prompt. All translation data is from the 'devtest' split of the FLORES-200 dataset (Costa-Jussà et al., 2022). We analyse three language pairs with English: Spanish (high-resource), Japanese (medium-resource), and te reo Māori (low-resource). Our metrics include PPL, SacreBLEU (Post, 2018), and COMET (Rei et al., 2020). For Japanese evaluation, we use the default 'ja-mecab' tokenizer, which relies on the MeCab morphological analyser (Kudo, 2005). The specific COMET model is 'Unbabel/wmt22-comet-da' (Rei et al., 2022), the top-performing model from the WMT22 shared task (Freitag et al., 2022).

Defining Naturalness We define *naturalness* as a translation's fluency and resemblance to human writing. In our case study, we operationalize this by assessing two of its key components: an increase in overall translation quality, measured by the COMET score, and the correction of a clear lexical error.

4 Analysing the Model-Human Perplexity Gap

Our first analysis involved generating translations for all language pairs and scoring the perplexity of both the MT and human reference for the same source sentence.

4.1 Quantifying the Gap

As is often observed in models trained with maximum likelihood estimation, the model consistently assigns lower perplexity to its own outputs than to diverse human references. Table 1 provides a precise quantification of this effect. In every case, the median perplexity of the model's own output is substantially lower than that of the human reference.

This gap is visualized in Figure 1. For ENG→MRI, the model is over 3.3 times more perplexed by the human translation than its own.

Table 1: Median perplexity of machine-generated text vs. human reference text using the NLLB-3.3B model. The 'Gap' column shows the calculated difference.

Direction	PPL (Machine)	PPL (Human)	Gap
SPA → ENG	1.44	2.67	1.23
JPN → ENG	1.73	3.17	1.44
MRI → ENG	1.60	3.63	2.03
ENG → SPA	1.45	3.29	1.84
ENG → JPN	2.77	10.67	7.90
ENG → MRI	2.15	7.18	5.03

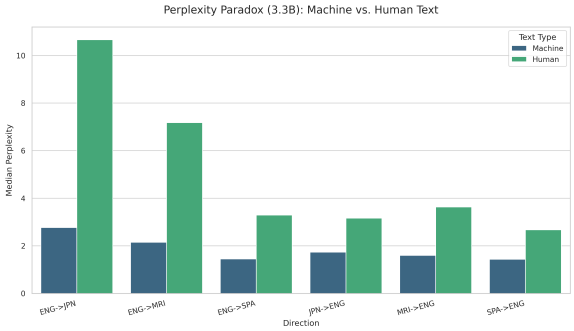

Figure 1: This chart visualizes the data from Table 1, showing the consistent gap between the median perplexity for machine-generated text and human-written text.

4.2 The Conditional Utility of Perplexity

Next, we investigated if perplexity, despite this gap, correlates with quality. Table 2 shows a clear trend: the strength of the negative correlation increases as the source language becomes lower-resourced. For te reo Māori, PPL becomes a stronger signal, with a correlation of -0.639 with COMET.

Table 2: Correlation between perplexity and quality metrics for the 3.3B model.

Direction	r (PPL vs. BLEU)	r (PPL vs. COMET)
SPA → ENG	−0.210	−0.299
JPN → ENG	−0.434	−0.446
MRI → ENG	−0.565	−0.639

This relationship is visualized in Figure 2. Translation quality drops as perplexity increases, providing support for using an adaptive perplexity filter to identify likely errors in lower-resource settings.

4.3 Dissecting the Uncertainty Signal

To better understand what drives sentence-level perplexity, we analysed features of tokens with high levels of surprisal (the model's predicted probability of a token given the preceding context). Table 3 shows that for the low-resource MRI→ENG direction, not only does perplexity have a strong negative correlation with quality, but the variance

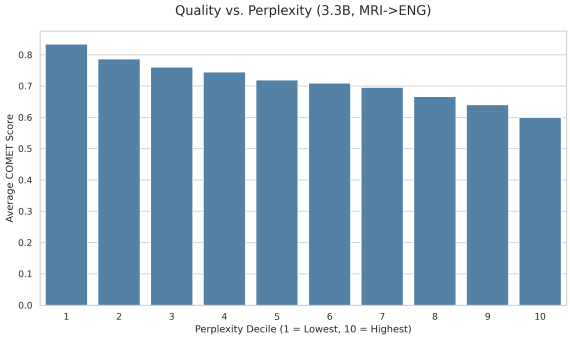

Figure 2: Average COMET score for MRI→ENG sentences binned by perplexity decile. The visible trend indicates a negative correlation between perplexity and translation quality.

of the token surprisals does as well. This suggests that translations with *spiky* or uneven uncertainty are also more likely to be of lower quality.

Table 3: Correlation of key surprisal features with COMET score for MRI→ENG (3.3B Model).

Feature	Pearson's r	p-value
perplexity_mt	-0.639	$< .0001$
variance_of_surprisal_mt	-0.173	$< .0001$

A deeper look at the most surprising tokens in the generated te reo Māori translations provides a direct explanation for this spikiness and the overall brittleness of perplexity as a metric. Table 4 shows that for ENG→MRI, the model's uncertainty is often caused by subword fragments from its tokenization of the target sentence, not complex semantic concepts. This suggests that high perplexity scores can be an artifact of statistically rare tokenization, where a single awkward subword inflates the uncertainty for an otherwise acceptable sentence.

Table 4: Top 5 most surprising tokens for ENG→MRI machine translations. The underscore (_) indicates a subword fragment.

Token	Mean Surprisal (bits)
`_ume`	24.81
`_oke`	24.38
`_nu`	24.29
`_ony`	23.74
`_ik`	23.71

4.4 Consistency Across Model Scales

To confirm that our findings are a general trait of the NLLB architecture and not specific to one model size, we also performed our core analyses on the smaller, distilled NLLB-600M model. Table 5 shows that our main conclusions are robust and consistent across both the primary and distilled models.

Table 5: Comparison of key metrics across model sizes. The '(r)' columns show the Pearson correlation between perplexity and COMET score for the given direction. The 'Gap' column shows the difference between the median perplexity of human vs. machine text for ENG→MRI.

Model	MRI→ENG (r)	SPA→ENG (r) (r)	ENG→MRI GAP
600M	-0.625	-0.448	7.73
3.3B	-0.639	-0.299	5.03

The table highlights two consistent behaviours. First, both models showed a much stronger negative correlation between perplexity and quality for the low-resource pair (MRI→ENG) than for the high-resource pair (SPA→ENG). Second, both models exhibit a large Model-Human Perplexity Gap for the low-resource ENG→MRI direction. This confirms that the conditional utility of perplexity and the preference for translationese are fundamental traits of the NLLB architecture, not just a quirk of a single model.

5 Case Study: Refining for Naturalness

Our broader motivation for studying uncertainty is a two-step process: first, using a signal like perplexity to detect potential errors, and second, using that signal to trigger a knowledge-based repair. This case study serves as a proof-of-concept for the second step. To explore our naturalness hypothesis, we selected a representative *expected failure* from the NLLB-3.3B model. We define expected failures as sentences in the top decile of perplexity (low model confidence) and the bottom decile of COMET score (low quality). The chosen translation contained a clear lexical error, mistranslating "spirits" as the Spanish word for ghosts instead of alcohol. We then constructed a prompt for Llama 3 containing the source, the flawed translation, and hand-selected contextual information to attempt a repair.

The results, shown in Table 6, demonstrate a clear success. The simulated Retrieval Augmented Generation (RAG)-based edit corrected the primary semantic error, leading to a significant improvement in quality.

The refinement process not only improved the translation's quality (a +0.15 jump in COMET score) but also moved its perplexity score (1.43) closer to the median perplexity of human text for

Table 6: A single expected failure sentence ('Illegal spirits can contain various dangerous impurities...') before and after RAG-based refinement. The original NLLB translation contains a lexical error, mistranslating 'spirits' as the Spanish word for ghosts (*espíritus*). The refined output is significantly higher quality and also has a slightly higher perplexity.

Version	Translation Text	PPL	BLEU	COMET
NLLB-3.3B (Original)	*Los espíritus ilegales* pueden contener varias impurezas peligrosas...	1.30	7.51	0.7627
Llama 3 (RAG)	*Las bebidas alcohólicas ilegales* pueden contener varias impurezas peligrosas...	**1.43**	**10.32**	**0.9156**

this direction (3.29). We note that the refined text was generated by Llama 3, and the perplexity was measured by NLLB. While this model mismatch is a confounding variable, the result illustrates that a higher-quality translation is not always one with the lowest possible perplexity according to the original model, and points to a potential for RAG-based or other knowledge-based MT refinement in lower-resource settings.

6 Discussion

Our experiments expose two key findings: a significant and consistent model-human perplexity gap exists, and the utility of perplexity as a quality signal is conditional on accounting for the resource level of the language pair.

The quantification of the perplexity gap is a central finding. The NLLB-3.3B model is systematically more confident in its own, often literal, outputs than the arguably more nuanced professional human translations. This indicates, not entirely surprisingly, that the model's internal sense of surprise is calibrated to its own output distribution—to its version of translationese—rather than to natural human language. We also observe an interaction between model scale and resource level. For the high-resource SPA→ENG pair, the correlation between perplexity and quality weakens with the larger model (from r=-0.448 to r=-0.299), hinting that once fluency reaches a ceiling, uncertainty becomes a less informative signal.

Token-level surprisal analysis further explains why perplexity can be a brittle quality indicator. In our study, high perplexity often stemmed from tokenization artifacts, such as rare subword fragments, rather than genuine semantic or syntactic difficulty. This suggests that raw perplexity scores may reflect superficial statistical anomalies more than deep meaning errors. Accounting for this may improve use of PPL for translation quality improvement.

Finally, our case study illustrates that perplexity might signal naturalness rather than correctness. This reframes its potential role: instead of simply minimizing perplexity, future work might aim to align a translation's perplexity profile with that of high-quality human-generated text, or with text whose correctness has been improved or certified by other means.

7 Limitations and Future Work

Our analysis focussed on the NLLB 3.3B model. Results may differ for the largest NLLB variants, which could exhibit different perplexity distributions. Additionally, our evaluation relies exclusively on FLORES-200, which consists of multidirectional translations of encyclopedic text. Testing across other domains will be essential to assess generalisability.

Our simulated RAG refinement was a proof of concept with a hand-selected context. A next step would be to develop and evaluate a full RAG system with automated retrieval strategies to determine if the observed improvements hold. More broadly, our results suggest that perplexity may serve as a signal for naturalness; scaling up the case study to correlate perplexity with human fluency judgements across diverse language pairs and domains could validate this.

8 Conclusion

We quantified the model–human perplexity gap in a large multilingual model, showing that perplexity often measures conformity to translationese rather than semantic quality. We found that its utility as a quality signal is strongest in low-resource settings.

Our case study demonstrates that targeted refinements can improve a translation while increasing perplexity, challenging the view of perplexity solely as a metric to minimize. A richer interpretation treating it as a useable signal for naturalness could open new directions for improving translation quality.

9 Ethical Considerations

Our ethical position is that everyone should have equitable access to language technology in their own language. In the context of low-resource languages, access to quality MT models is greatly lacking.

Our research primarily utilizes the publicly available FLORES-200 dataset, which was created in collaboration with native speakers for the express purpose of advancing multilingual NLP research and is considered a standard benchmark in the field.

The large language models used in this study, NLLB and Llama 3, are known to contain biases from their training data. While our work focuses on improving translation for low-resource languages like te reo Māori—a step towards more equitable technology—the underlying models may still generate outputs that reflect societal biases or perform inequitably across different demographic groups.

Furthermore, TQE has a dual-use potential. While our goal is to use uncertainty signals to improve translation quality and naturalness, automated TQE systems could also be used to justify the deployment of imperfect MT systems in sensitive contexts (e.g., medical, legal) without adequate human oversight. We advocate for the use of TQE as a tool to assist human translators, not to replace them, especially in high-stakes applications for low-resource communities.

Future work in this sensitive domain has to go beyond the inadequate consent-compensate-cooperate model of ethical behaviour. Development of language technology need not only be built with the consent of the target language communities, but should be shared with these language users at its inception to ensure alignment with their cultural values and use cases.

References

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/*.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT 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.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 415–425.