Unsupervised Word-level Quality Estimation for Machine Translation Through the Lens of Annotators (Dis)agreement

Gabriele Sarti[1] **Vilém Zouhar**[2] **Malvina Nissim**[1] **Arianna Bisazza**[1]

[1]CLCG, University of Groningen [2]ETH Zurich

{g.sarti, a.bisazza}@rug.nl

Abstract

Word-level quality estimation (WQE) aims to automatically identify fine-grained error spans in machine-translated outputs and has found many uses, including assisting translators during post-editing. Modern WQE techniques are often expensive, involving prompting of large language models or ad-hoc training on large amounts of human-labeled data. In this work, we investigate efficient alternatives exploiting recent advances in language model interpretability and uncertainty quantification to identify translation errors from the inner workings of translation models. In our evaluation spanning 14 metrics across 12 translation directions, we quantify the impact of human label variation on metric performance by using multiple sets of human labels. Our results highlight the untapped potential of unsupervised metrics, the shortcomings of supervised methods when faced with label uncertainty, and the brittleness of single-annotator evaluation practices.

1 Introduction

Word-level error spans are widely used in machine translation (MT) evaluation to obtain robust and fine-grained estimates of translation quality (Lommel et al., 2014; Freitag et al., 2021a,b; Kocmi et al., 2024b). Due to the cost of manual annotation, word-level quality estimation (WQE) was proposed for assisting in annotating error spans over MT outputs (Zouhar et al., 2025). Modern WQE approaches generally rely on costly inference with large language models (LLMs) or ad-hoc training with large amounts of human-annotated texts (Fernandes et al., 2023; Kocmi and Federmann, 2023; Guerreiro et al., 2024), making them impractical for less resourced settings (Zouhar et al., 2024). To improve the efficiency of MT quality assessment, several works explored the use of signals derived from the internals of neural MT systems

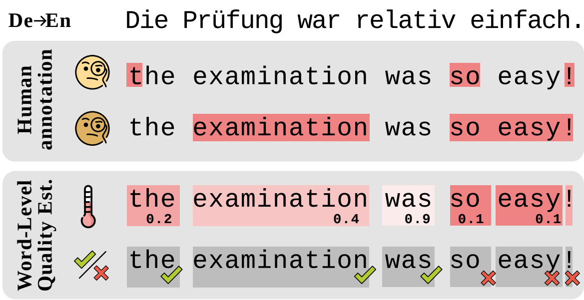

Figure 1: Example of German→English translation with two sets of human word-level error span annotations and two examples of continuous and binary WQE metrics.

(Fomicheva et al., 2020b, 2021; Leiter et al., 2024), for identifying problems in MT outputs, such as hallucinations (Guerreiro et al., 2023a,b; Dale et al., 2023a,b; Himmi et al., 2024). However, previous works focus on sentence-level metrics for overall translation quality, and do not evaluate performance on multiple label sets due to high annotation costs (Fomicheva et al., 2022; Zerva et al., 2024).[1]

In this work, we conduct a more comprehensive evaluation spanning 10 unsupervised metrics derived from models' inner representations and predictive distributions to identify word-level translation errors. We test three open-source multilingual MT models and LLMs of different sizes across 12 translation directions, including typologically diverse languages and challenging textual domains. Importantly, we focus on texts with *multiple* human annotations to measure the impact of individual annotator preferences on metric performance, setting a "human-level" baseline for the WQE task.

We address the following research questions: **i)** How accurate are unsupervised WQE metrics in detecting MT errors compared to trained metrics and human annotators? **ii)** Are popular supervised WQE metrics well-calibrated? **iii)** Are the relative performances of WQE metrics affected by the variability in human error annotations?

[†]Materials: gsarti/labl/examples/unsup_wqe.

[1]Other relevant works are discussed in Appendix A

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 18321–18338
November 4-9, 2025 ©2025 Association for Computational Linguistics

	DivEMT	WMT24	QE4PE
Languages	EN→AR,IT, NL,TR,UK,VI	EN→JA,ZH, HI,CS,RU CS→UK	EN→IT,NL
Errors type	Post-edit	Annotation	Post-edit
Label sets	1	1	6
Domains	Wiki	Multiple	Social, Biomed
MT Model	mBART-50	Aya23	NLLB
# Segments	2580	5124	3888

Table 1: Summary of tested datasets. Error spans are obtained from explicit error annotations or post-edited spans. Additional details are available in Appendix B.

We conclude with recommendations for improving the evaluation and usage of future WQE systems.

2 Data

We use datasets containing error annotations or post-edits on the outputs of open-source models to extract unsupervised WQE metrics from real model outputs, avoiding potential confounders. We select the following datasets, summarized in Table 1:

DivEMT (Sarti et al., 2022) contains a single set of post-edits over translations produced by mBART-50 (Tang et al., 2021) for a subset of Wiki texts from the FLORES dataset (Goyal et al., 2022) spanning six typologically diverse target languages (EN→AR,IT,NL,TR,UK,VI). We use it to conduct cross-lingual comparisons over a fixed set of examples.

WMT24 (Kocmi et al., 2024a) contains error spans on the outputs of the Aya23-35B LLM (Aryabumi et al., 2024) produced for the WMT24 General Translation Shared Task spanning multiple domains across six directions (EN→JA,ZH,HI,CS,RU and CS→UK). It was selected to extend our evaluation to a state-of-the-art LLM, given the popularity of such systems in MT (Kocmi et al., 2023).

QE4PE (Sarti et al., 2025) contains multiple human professional post-edits over translations produced by the NLLB 3.3B model (Costa-jussà et al., 2024) for EN→IT and EN→NL on challenging textual domains (social posts and biomedical abstracts). This dataset is used to conduct our evaluation across multiple annotation sets.

3 Evaluated Metrics

The following metrics were evaluated using the Inseq library (Sarti et al., 2023, 2024b). Appendix C provides additional details on tested metrics.

Predictive Distribution Metrics. We use the **Surprisal** of the predicted token t^*, as negative log-probablity $-\log p(t_i^*|t_{<i})$, and the **Entropy** H of the output distribution P_N over vocabulary V, $-\sum_{i=1}^{|V|} p(t_i|t_{<i}) \log_2 p(t_i|t_{<i})$, as simple metrics to quantify pointwise and full prediction uncertainty (Fomicheva et al., 2020b). For surprisal, we also compute its expectation ($\mathbf{MCD_{AVG}}$) and variance ($\mathbf{MCD_{VAR}}$) with $n = 10$ steps of Monte Carlo Dropout (MCD, Gal and Ghahramani, 2016) to obtain a robust estimate and a measure of epistemic uncertainty in predictions, respectively.[2]

Vocabulary Projections. We use the LogitLens (LL, nostalgebraist, 2020) to extract probability distributions $P_0, \dots, P_{N-1}$ over V from intermediate activations at every layer $l_0, \dots, l_{N-1}$ of the decoder. We use the surprisal for the final prediction at every layer (**LL-Surprisal**) to assess the presence of layers with high sensitivity to wrong predictions. Then, we compute the KL divergence between every layer distribution and the final distribution P_N, e.g. $\text{KL}(P_{N-1} \| P_N)$, to highlight trends in the shift in predictive probability produced by the application of remaining layers (**LL KL-Div**). Finally, we adapt the approach of Baldock et al. (2021) and use the number of the first layer for which the final prediction corresponds to the top logit as a metric of model confidence, l s.t. $\arg\max P_l = t^*$ and $\arg\max P_i \neq t^* \; \forall i < l$ (**LL Pred. Depth**).

Context mixing. We use the entropy of the distribution of attention weights[3] over previous context as a simple measure of information locality during inference (Ferrando et al., 2022; Mohebbi et al., 2023). Following Fomicheva et al. (2020a), we experiment with using the mean and the maximum entropy across all attention heads of all layers as separate metrics ($\textbf{Attn. Entropy}_{\text{VAR/MAX}}$). Finally, we evaluate the Between Layer OOD method proposed by Jelenić et al. (2024), which employs gradients to estimate layer transformation smoothness

[2] Epistemic uncertainty reflects models' lack of knowledge rather than data ambiguity. MCD is tested only on encoder-decoder models since Aya layers do not include dropout.

[3] For the encoder-decoder model, self-attention and cross-attention weights are concatenated and renormalized.

	Method	DivEMT		WMT24		QE4PE	
		AP	F1*	AP	F1*	AP	F1*
	Random	.34	.50	.05	.09	.17	.27
UNSUPERVISED	Surprisal	.43	.53	.08	.13	.23	.32
	Out. Entropy	.46	.51	<u>.10</u>	<u>.16</u>	.23	.31
	Surprisal MCD $_{\text{AVG}}$	.43	.53	-	-	.24	.33
	Surprisal MCD $_{\text{VAR}}$	<u>.47</u>	<u>.54</u>	-	-	<u>.26</u>	<u>.34</u>
	LL Surprisal $_{\text{BEST}}$	.42	.53	.09	.15	.23	.32
	LL KL-Div $_{\text{BEST}}$	.43	.51	.07	.12	.20	.29
	LL Pred. Depth	.39	.51	.06	.12	.20	.29
	Att. Entropy $_{\text{AVG}}$	.37	.50	.05	.09	.18	.28
	Att. Entropy $_{\text{MAX}}$	.34	.50	.05	.09	.16	.28
	BLOOD $_{\text{BEST}}$	.34	.50	-	-	.17	.28
SUPER.	XCOMET-XL	.42	.45	.09	.19	.23	.34
	XCOMET-XL $_{\text{CONF}}$	.54	**.55**	.15	.23	.32	**.37**
	XCOMET-XXL	.43	.41	.09	.20	.22	.31
	XCOMET-XXL $_{\text{CONF}}$	**.56**	**.55**	**.16**	**.24**	**.33**	**.37**
HUM.	Hum. Editors $_{\text{MIN}}$	-	-	-	-	.24	.34
	Hum. Editors $_{\text{AVG}}$	-	-	-	-	.28	.41
	Hum. Editors $_{\text{MAX}}$	-	-	-	-	.32	.47

Table 2: Average Precision (AP) and Optimal F1 (F1*) for metrics across tested datasets. Results are averaged across all languages and annotators, with <u>best unsupervised</u> and **overall best** results highlighted.

for OOD detection (**BLOOD**).

Supervised baselines. We also test the state-of-the-art supervised WQE model XCOMET (Guerreiro et al., 2024) in its XL (3.5B) and XXL (10.7B) sizes, using them as binary metrics. Contrary to the continuous metrics from the previous section, binary labels from XCOMET cannot be easily calibrated to match subjective annotation propensity. Hence, we propose to adapt the XCOMET metric to use the sum of probability for all error types as a token-level continuous confidence metric, $s(t^*) = p(\text{MINOR}) + p(\text{MAJOR}) + p(\text{CRITICAL})$, which we dub **XCOMET$_{\text{CONF}}$**.

Human Editors. For QE4PE, we report the min/mean/max agreement between each annotator's edited spans and those of the other five editors as a less subjective "human-level" quality measure.

4 Experiments

How Accurate are Unsupervised WQE Metrics? Table 2 reports the average metrics performance across all translation directions across the tested datasets.[4] We report Average Precision (AP) as it provides a threshold-independent measure of ranking quality across the full score range. Such a metric enables us to compare continuous metrics with

[4]Full breakdown available in the Appendix (Tables 5 to 8).

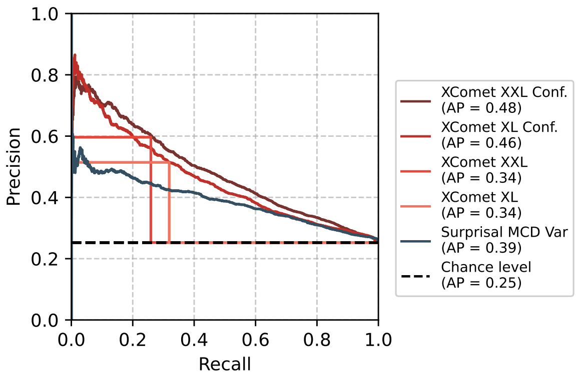

Figure 2: Precision-Recall tradeoff for binary and confidence-weighted XCOMET variants and the Surprisal MCD$_{\text{VAR}}$ metric for DivEMT EN→IT.

different scales and provides an expectation for precision when the annotator's annotation propensity is unknown beforehand. We use the best F1 score (F1*), i.e. the F1 score for best threshold calibrated to maximize the precision-recall tradeoff, to simulate a realistic evaluation setup with calibration where continuous metric scores are binarized into positive/negative labels matching human annotation.[5] Our results show that, despite high variability in error span prevalence across different models, languages and annotators, metric rankings remain generally consistent, suggesting the presence of **robust relations between various signals sourced from models' inner workings and translation errors**. Among unsupervised metrics, we find those based on the output distribution to be most effective at identifying error spans, in line with previous segment-level QE results (Fomicheva et al., 2020b). Notably, the Surprisal MCD$_{\text{VAR}}$ shows strong performances in line with the default XCOMET models. For the multi-label QE4PE dataset, we find that the best supervised metrics score on par with the average consensus of human annotators (Hum. Editors$_{\text{AVG}}$). In contrast, unsupervised metrics generally obtain lower performances.

Confidence Weighting Enables XCOMET Calibration. From Table 2 results, default XCOMET metrics underperform compared to the best unsupervised techniques, a surprising result given their ad-hoc tuning. On the contrary, our XCOMET$_{\text{CONF}}$ method consistently reaches better results across all tested sets. Figure 2 shows the precision-recall tradeoff for these metrics on the EN→IT subset of

[5]AP for the random baseline corresponds to the proportion of tokens marked as errors, which varies greatly across datasets and annotators.

Source $_{EN}$	So why is it that people jump through extra hoops to install Google Maps?
MT $_{IT}$ (NLLB)	Quindi perché le persone devono fare un salto in più per installare Google Maps?
Annotator $t1$	Quindi perché le persone devono fare un [salto] passaggio in più per installare Google Maps?
Annotator $t2$	Quindi perché le persone [devono fare un salto in più] fanno i salti mortali per installare Google Maps?
Annotator $t3$	Quindi perché le persone [devono fare un salto in più] effettuano dei passaggi ulteriori e superflui per installare Google Maps?
Annotator $t4$	[Quindi] Allora perché le persone [devono fare] fanno un [salto] passaggio in più per installare Google Maps?
Annotator $t5$	[Quindi perché le persone devono fare un salto in più] E allora mi chiedo: perché gli utenti iPhone si affannano tanto per installare Google Maps?
Annotator $t6$	Quindi perché le persone [devono fare un salto in più] fanno di tutto per installare Google Maps?
Edit Counts (Fig. 3)	Quindi (2) perché le persone (1) devono fare (5) un (4) salto (6) in più (4) per installare Google Maps?
XCOMET-XL	Quindi perché le persone devono fare (minor) un salto in più (minor) per installare Google Maps?
XCOMET-XXL	Quindi perché (minor) le persone devono fare un salto in più (major) per installare Google Maps?

XCOMET-XL $_{CONF}$	Quindi	perché	le	persone	devono	fare	un	salto	in	più	per	install	are	Google	Maps	?
	.41	.36	.51	.50	.69	.73	.51	.81	.74	.76	.39	.47	.53	.26	.36	.24
XCOMET-XXL $_{CONF}$	Quindi	perché	le	persone	devono	fare	un	salto	in	più	per	install	are	Google	Maps	?
	.51	.83	.20	.20	.42	.84	.90	.95	.86	.78	.03	.00	.01	.00	00	.00
Surprisal MCD $_{VAR}$	Quindi	perché	le	persone	devono	fare	un	sal	to	in	più	per	installare	Google	Maps	?
	.05	.01	.04	.00	.41	.09	.04	.59	.00	.12	.00	.00	.00	.00	.00	.00

Table 3: Annotated example from the EN→IT portion of the QE4PE dataset. **Top:** Annotator edits with highlighted final text and replaced text on top, with count-based aggregation showing inter-annotator agreement. **Bottom:** Word-level annotations for best-performing metrics discussed in the study.

the DIVEMT dataset.[6] In their default form commonly used for evaluation via the `unbabel-comet` library, XCOMET metrics consistently outperform Surprisal MCD$_{VAR}$ in terms of precision (51-60%, compared to 34% optimal precision for MCD$_{VAR}$), but identify only 32-26% of tokens annotated as errors, resulting in lower AP. The low recall of these metrics may be problematic in WQE applications, where omitting an error could result in oversights by human post-editors, who trust the comprehensiveness of WQE predictions. On the contrary, confidence-weighted XCOMET$_{CONF}$ models show strong performances across the whole recall range, resulting in consistent improvements in both F1* and AP Table 2. Concretely, these results confirm that default XCOMET performance does not reflect the full capacity of the metric, and **operating with granular confidence scores can be beneficial when calibration is possible**. This said, for cases with a larger proportion of translated words labeled as errors, such as the DivEMT dataset, we remark that the F1* performance of XCOMET$_{CONF}$ metrics is very close to that of human annotators (e.g., Translator 6 for QE4PE results of Table 6) and unsupervised metrics (e.g., all DivEMT languages in Table 7). While this can be attributed in part to a higher number of subjective choices when more errors are identified, these results suggest that supervised metrics might still underperform on problematic texts, despite our proposed confidence-weighting procedure.

Metrics Performance for Multiple Annotations. While our evaluation so far employed human error span annotations as binary labels, we set out to assess how more granular labeling schemes impact the performance of these metrics. Given L sets of binary labels (up to 6 per language for QE4PE), we assign a score $s \in \{1, \dots, L\}$ to every MT token using the number of annotators that marked it as an error, resulting in edit counts reflecting human agreement rate. Table 3 provides an example of six human annotations with proposed edits, and labels derived from best-performing metrics. Figure 3 presents the correlation of various metrics when the number of annotators available is increased, with median values and confidence bounds are obtained from edit counts across all combinations of L label sets.[7] The increasing trend for correlations across all reported metrics indicates

[6]Results for all datasets in the Appendix (Figures 4 to 7).

[7]x=1 corresponds to binary labels from previous sections.

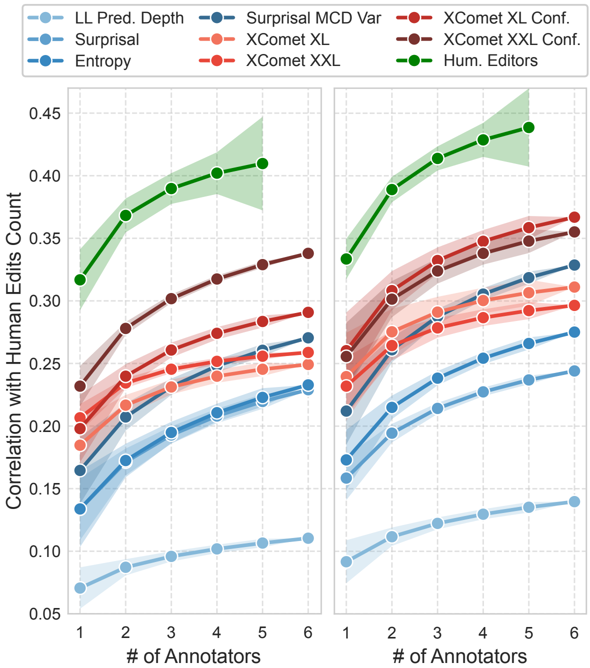

Figure 3: Spearman correlation between WQE metric scores and human edit counts across multiple annotation sets for QE4PE EN→IT (left) and EN→NL (right).

that these methods reflect well the *aleatoric uncertainty* in error span labels, i.e. the disagreement between various annotators. In particular, the Surprisal MCD_{VAR} metric sees a steeper correlation increase than other well-performing metrics, surpassing default XCOMET supervised approaches for higher correlation bins. This suggests the epistemic uncertainty derived from noisy model predictions might be a promising way to anticipate the aleatoric uncertainty across human annotators for WQE. We observe that 95% confidence intervals for high-scoring metrics are largely overlapping when a single set of labels is used, indicating that **rankings of metric performance are subject to change depending on subjective choices of the annotator**. While this poses a problem when attempting a robust evaluation of WQE metrics, we remark that including multiple annotations largely mitigates this issue. As a result, we recommend explicitly accounting for human label variation by including multiple error annotations in future WQE evaluations to ensure generalizable findings.

5 Conclusion

We conducted a comprehensive evaluation of supervised and unsupervised WQE metrics across multiple languages and annotation sets. Our results show that **i)** While unsupervised metrics generally lag behind state-of-the-art supervised systems, some uncertainty quantification methods based on the predictive distribution show promising correlation with human label variation; **ii)** Popular supervised WQE metrics have generally low levels of recall, and can benefit from confidence weighting to when calibration is possible; and **iii)** Individual annotator preferences are key confounders in WQE evaluations and can be mitigated by making use of multiple annotation sets. We offer the following practical recommendations for evaluating WQE systems:

- Use agreement between multiple human annotations to control the effect of subjective preferences and rank WQE metrics robustly.
- Employ an in-distribution calibration set of error spans before testing to ensure fair metric comparisons, and favor evaluations accounting for precision-recall tradeoffs to ensure their usability across various confidence levels.
- Previous work showed the effectiveness of visualization reflecting prediction confidence (Vasconcelos et al., 2025), such as highlights for various error severity levels (Sarti et al., 2025). Consider using continuous WQE metrics in real-world applications such as WQE-augmented post-editing to convey fine-grained confidence variations.

Acknowledgements

Gabriele Sarti and Arianna Bisazza acknowledge the support of the Dutch Research Council (NWO) for the project InDeep (NWA.1292.19.399). Arianna Bisazza is further supported by the NWO Talent Programme (VI.Vidi.221C.009). We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high-performance computing cluster.

Limitations

Our findings are accompanied by several limitations. Firstly, our choice of tested datasets was limited by the availability of annotated outputs generated by open-source MT models. While several other datasets matching these criteria exist (Fomicheva et al., 2022; Yang et al., 2023; Dale et al., 2023b), we restricted our assessment to a sufficient subset to ensure diversity across languages and tested models to support our findings. To facilitate comparison with other datasets, our evaluation for WMT24 treats available error spans as binary labels and does not directly account for error severity in human-annotated spans. Our choice of unsupervised metrics was primarily driven by previous work on uncertainty quantification in MT, and ease of implementation for popular methods in mechanistic interpretability literature (Ferrando et al., 2024). However, our choices in the latter category were limited, as most methods are now developed and tested specifically for decoder-only transformer models. Finally, despite their strong performance, we found unsupervised methods based on MCD to require substantial computational resources, and as such we could not evaluate them on Aya23 35B. While our primary focus was to establish baseline performances across various popular methods, future work should leverage the latest insights from more advanced techniques, such as those requiring the tuning of vocabulary projections (Belrose et al., 2023; Yom Din et al., 2024) or the identification of "confidence neurons" to modulate predictive entropy (Stolfo et al., 2024).

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications.

Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. 2021. Deep learning through the lens of example difficulty. In *Advances in Neural Information Processing Systems*, volume 34, pages 10876–10889. Curran Associates, Inc.

Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, Luke Ong, Philip Torr, Kwok-Yan Lam, Robert Trager, David Krueger, Sören Mindermann, José Hernandez-Orallo, Mor Geva, and Yarin Gal. 2025. Open problems in machine unlearning for ai safety.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653. Association for Computational Linguistics.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. A close look into the calibration of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1343–1367. Association for Computational Linguistics.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. 2024. Contextcite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, volume 37, pages 95764–95807. Curran Associates, Inc.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50. Association for Computational Linguistics.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault, and Marta Costa-jussà. 2023b. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits.

Transformer Circuits Thread. Https://transformer-circuits.pub/2021/framework/index.html.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513. Association for Computational Linguistics.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974. European Language Resources Association.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020a. BERGAMOT-LATTE submissions for the WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1010–1017. Association for Computational Linguistics.

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

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023a. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784. Association for Computational Linguistics.

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.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR 2017)*.

Anas Himmi, Guillaume Staerman, Marine Picot, Pierre Colombo, and Nuno M Guerreiro. 2024. Enhanced hallucination detection in neural machine translation through simple detector aggregation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18573–18583. Association for Computational Linguistics.

Fran Jelenić, Josip Jukić, Martin Tutek, Mate Puljiz, and Jan Snajder. 2024. Out-of-distribution detection by leveraging between-layer transformation smoothness. In *The Twelfth International Conference on Learning Representations*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *J. Mach. Learn. Res.*, 25(1).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172. European Association for Machine Translation.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400. Association for Computational Linguistics.

Marius Mosbach, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva. 2024. From insights to actions: The impact of interpretability and analysis research on NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3105. Association for Computational Linguistics.

nostalgebraist. 2020. Interpreting GPT: the logit lens. *AI Alignment Forum*.

Barbara Plank. 2022. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511. Association for Computational Linguistics.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. 2023. Conformal nucleus sampling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 27–34. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Raphael Rubino, Atsushi Fujita, and Benjamin Marie. 2021. Error identification for machine translation with metric embedding and attention. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 146–156. Association for Computational Linguistics.

Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7795–7816. Association for Computational Linguistics.

Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2024a. Quantifying the plausibility of context reliance in neural machine translation. In *The Twelfth International Conference on Learning Representations*.

Gabriele Sarti, Nils Feldhus, Jirui Qi, Malvina Nissim, and Arianna Bisazza. 2024b. Democratizing advanced attribution analyses of generative language models with the inseq toolkit. In *xAI-2024 Late-breaking Work, Demos and Doctoral Consortium Joint Proceedings*, pages 289–296, Valletta, Malta. CEUR.org.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435. Association for Computational Linguistics.

Gabriele Sarti, Vilém Zouhar, Grzegorz Chrupała, Ana Guerberof-Arenas, Malvina Nissim, and Arianna Bisazza. 2025. QE4PE: Word-level quality estimation for human post-editing.

Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. 2024. Confidence regulation neurons in language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 125019–125049. Curran Associates, Inc.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association* *for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466. Association for Computational Linguistics.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735. Association for Computational Linguistics.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939. Association for Computational Linguistics.

Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2025. Generation probabilities are not enough: Uncertainty highlighting in ai code completions. *ACM Trans. Comput.-Hum. Interact.*, 32(1).

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269. Association for Computational Linguistics.

Zhen Yang, Fandong Meng, Yuanmeng Yan, and Jie Zhou. 2023. Rethinking the word-level quality estimation for machine translation from human judgement. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2012–2025. Association for Computational Linguistics.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2024. Jump to conclusions: Shortcutting transformers with linear transformations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9615–9625. ELRA and ICCL.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99. Association for Computational Linguistics.

Chrysoula Zerva and André F. T. Martins. 2024. Conformalizing machine translation evaluation. *Transactions of the Association for Computational Linguistics*, 12:1460–1478.

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

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500. Association for Computational Linguistics.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

A Additional Background

In this section, we provide additional background information regarding the topics of our work.

Unsupervised Quality Estimation for Machine Translation. The use of unsupervised signals from MT models for the task of MT quality estimation was introduced by Fomicheva et al. (2020b). Their evaluation revealed that high-performing unsupervised methods could rival state-of-the-art supervised QE models in predicting sentence-level translation quality. Since then, several evaluation campaigns assessed the quality of QE methods (Specia et al., 2021; Zerva et al., 2022; Blain et al., 2023; Zerva et al., 2024), including a shared task dedicated to explainable QE metrics (Fomicheva et al., 2021). However, such evaluations have typically focused on segment-level evaluation quality, with word-level error spans being generally obtained by attributing the predictions of supervised segment-level metrics (Rubino et al., 2021; Rei et al., 2023). By contrast, recent work on LLMs evaluates various metrics to detect errors from the generator model, without the need for additional systems, both at the sentence level (Fadeeva et al., 2023) and at the token level (Fadeeva et al., 2024). Our work follows the latter approach by testing unsupervised metrics extracted from an MT model during generation, akin to out-of-distribution detection in signal processing research (Hendrycks and Gimpel, 2017).

Actionable Insights from Interpretability. Advances in interpretability research have elucidated multiple mechanisms underlying decision-making, knowledge representation, and biases in LMs (Ferrando et al., 2024). However, a better understanding of the model's inner workings often did not translate to tangible gains in model design and other practical applications, which remain rarely explored (Mosbach et al., 2024). Some examples in this direction include using targeted machine unlearning methods for safety-critical scenarios (Barez et al., 2025), or the use of attribution for trustworthy context citations in LM generations (Cohen-Wang et al., 2024; Sarti et al., 2024a; Qi et al., 2024). In this work, signals extracted from model internals are employed to detect errors in models' generated outputs.

Uncertainty Estimation for Language Models The estimation of uncertainty in language models has garnered increasing attention (Baan et al., 2023), particularly in the context of generation tasks for which the set of plausible responses is large (Giulianelli et al., 2023). Predictive uncertainty is typically decomposed into its *aleatoric* and *epistemic* components, representing respectively the irreducible variability in the modeled phenomena, and the improvable confidence in model predictions (Kiureghian and Ditlevsen, 2009). Popular methods for uncertainty estimation involve the calibration of predictive probabilities to reflect aleatoric uncertainty (Jiang et al., 2020; Ulmer et al., 2022; Zhao et al., 2023; Chen et al., 2023), and conformal sets prediction (Zerva and Martins, 2024; Ravfogel et al., 2023). In this work, we leverage uncertainty signals from the predictive distribution of MT models and their internal processing to efficiently predict the resulting generation quality at a fine-grained, token-level scale.

Human Label Variation. Human label variation is a type of uncertainty that arises from the inherent variability in human judgments (Plank et al., 2014; Plank, 2022), which can be hard to disentangle from actual annotation mistakes (Snow et al., 2008; Weber-Genzel et al., 2024). The use of multiple references was recently recommended to ensure a sound evaluation of generative LMs reflecting human-plausible levels of variability (Giulianelli et al., 2023), contrary to standard practices that employ a single set of "gold" labels. In our analysis on QE4PE data containing multiple edits, we adopt a perspectivist approach[8] to ensure a robust assessment of WQE metrics by accounting for annotators' disagreement (Uma et al., 2021).

B Details about Models and Datasets

B.1 MT Models

mBART-50 1-to-many. The original multilingual BART (mBART-25) model by Liu et al. (2020) is an encoder-decoder Transformer model pretrained on monolingual documents in 25 languages with the BART denoising objective for sequence-to-sequence learning (Lewis et al., 2020). Tang et al. (2021) extended mBART-25 by including 25 additional languages during pre-training and performing multilingual translation fine-tuning across 50 languages. In this work, we employ the *one-to-many* version of the model specialized in out-of-English translation that was employed by Sarti

[8] pdai.info

et al. (2022) to produce part of the translations post-edited by DivEMT annotators.[9] The model is a standard Transformer with 12 layers of encoder and 12 layers of decoder, with model dimension of 1024 and 16 attention heads (~680M parameters).

NLLB 3.3B (No Language Left Behind) is a collection of multilingual MT models covering up to 202 languages, including low-resource directions (Costa-jussà et al., 2024). The largest NLLB model available is a mixture-of-experts model with 54.4B parameters, which comes with high computational cost. In this work we employ the largest available dense variant of the model (~3.3B parameters), which was used by Sarti et al. (2025) for collecting the QE4PE post-editing dataset.[10] The model is an encoder-decoder Transformer with 24 layers for each module, a model dimension of 2048 and 16 attention heads per layer.

Aya23 35B is a large language model introduced by Aryabumi et al. (2024) to improve the multilingual capabilities of the original Aya model (Üstün et al., 2024) on a selected set of 23 languages. The model was included in the WMT24 evaluation of Kocmi et al. (2024a), yielding the best translation performance among the tested open-source models. The model is a decoder-only Transformer model with 40 layers, a model dimension of 8196 and 64 attention heads per layer.

B.2 Datasets

DivEMT was created by (Sarti et al., 2022) to evaluate the impact of language typology on MT quality, and how that would influence the productivity of human post-editors working with those systems. The dataset includes out-of-English machine translations for Wiki data produced by Google Translate and mBART-50 1-to-many, with edits made by professional translators in six languages. In this work, we evaluate unsupervised metrics on the mBART-50 1-to-many model, converting the human post-edits into token-level labels.

WMT24 employed in this study is taken from the General Machine Translation Shared Task at WMT 2024 (Kocmi et al., 2024a). It contains evaluation of several machine translation systems across English→{Czech, Hindi, Japanese, Chinese, Russian} (634 segments) and Czech→Ukrainian (1954 segments). The human evaluation was conducted using the Error Span Annotation protocol (ESA, Kocmi et al., 2024b), which has human annotators highlighting erroneous spans in the translation and marking them as either MINOR or MAJOR errors. This dataset covers the *news*, *social*, and *speech* (with automatic speech recognition) domains. We adopt the official prompting setup from the WMT24 campaign, using the Aya23 model alongside the provided prompt and three in-context translation examples per language to ensure uniformity with previous results.[11]

QE4PE The QE4PE dataset was created by Sarti et al. (2025) for measuring the effect of word-level error highlights when included in real-world human post-editing workflows. The QE4PE data provides granular behavioral metrics to evaluate the speed and quality of post-editing of 12 annotators for EN→IT and EN→NL across four error span highlighting modalities, including the unsupervised Surprisal MCD_{VAR} method and the supervised XCOMET-XXL we also test in this study. Provided that the presence of error span highlights was found to influence the editing choices of human editors, we limit our evaluation to the six human annotators per language that post-edited sentences without any highlights (3 for the *Oracle Post-edit* task to produce initial human-based highlights, and 3 for the *No Highlight* modality in the main task). This prevents us from biasing our evaluation of WQE metrics in favor of the metrics that influenced editing choices. We use the post-edited versions to synthetically create error spans, which can be used as binary labels to evaluate WQE metrics.

C Details about Tested Metrics

Monte Carlo Dropout (MCD) is a technique introduced by Gal and Ghahramani (2016) for estimating model uncertainty at inference time. MCD utilizes the dropout mechanism in neural networks (Srivastava et al., 2014), a regularization technique commonly employed during training, to produce a set of noisy predictions from a unique model at inference time, thereby approximating Bayesian inference. For a given input x, T forward passes are performed through the network. In each pass $t \in T$, a different random dropout mask Θ_t is applied, resulting in a slightly different output probabilities $p(x \mid \Theta_t)$. The set of T predictions $\{p(x \mid \Theta_1), \ldots, p(x \mid \Theta_T)\}$ can be seen as sam-

[9] facebook/mbart-large-50-one-to-many-mmt

[10] facebook/nllb-200-3.3B

[11] wmt-conference/wmt-collect-translations

ples from an approximate posterior distribution. In this work, we employ the mean of the negative log probabilities as a robust estimate of surprisal:

$$\text{Surprisal MCD}_{\text{avg}} = \hat{y}_{\text{MCD}} = \frac{1}{T}\sum_{t=1}^{T} -\log p(x|\Theta_t)$$

Moreover, we estimate predictive uncertainty by calculating the variance of predictive probabilities under the same setup:

$$\text{Surprisal MCD}_{\text{var}} = \frac{1}{T}\sum_{t=1}^{T} \left(-\log p(x|\Theta_t) - \hat{y}_{\text{MCD}}\right)$$

Vocabulary Projections. The Logit Lens (nostalgebraist, 2020) is an interpretability technique used to understand the internal workings of Transformer models, particularly how their predictions evolve layer by layer. Activations h_l produced by the model layer l are projected to vocabulary space using the model unembedding matrix, W_U, commonly used to produce output logits. For the NLLB and mBART-50 models, we apply a final layer normalization before the projection, as per the model architecture. In contrast, for the Aya model, we scale the logits by 0.0625 (the default `logit_scale` defined in the model configuration). Following the residual stream view of the Transformer model (Elhage et al., 2021), the resulting logits provide a view into the model's predictive confidence at that specific depth of processing.

Context mixing. Several works studied the mixing of contextual information across language model layers to attribute model predictions to specific input properties (Ferrando et al., 2022; Mohebbi et al., 2023; Ferrando et al., 2023 *inter alia*). In this work, we employ simple estimates of context relevance using attention weights produced during the Transformer attention operation. More specifically, for every attention head at every layer of the decoder module, we extract a score for each token in the preceding context, employing cross-attention weights to account for source-side context in encoder-decoder models.

XCOMET is a suite of MT evaluation metrics introduced by Guerreiro et al. (2024), extending the popular COMET metric (Rei et al., 2020) to combine sentence-level and word-level error span prediction for improved explainability of results. XCOMET metrics are available in 3B (XL) and 11B (XXL) sizes and support both reference-based and reference-less usage, hence enabling usage for quality estimation purposes. Concretely, XCOMET models are Transformer encoders fine-tuned from pre-trained XLMR encoders (Goyal et al., 2021) using a mix of sentence-level Direct Assessment scores and word-level MQM error spans. In this work, we focus on the word-level error span prediction capabilities of the model in a quality estimation setup, where it classifies every input token according to MQM severity levels {OK, MINOR, MAJOR, CRITICAL} using a learned linear layer.[12]

Token-level Evaluation. Error spans used as labels in our evaluation are defined at the character level, while metric scores depend on the tokenization employed by either the MT model (for unsupervised metrics) or XCOMET (for supervised metrics). To facilitate comparison, we label tokens as part of an error span if at least one character contained within them was marked as an error or edited by an annotator. Tables 3 and 4 provide examples of various segmentations for the same MT output.

Constraining generation Evaluating metrics at the word level can be challenging due to the need for perfect uniformity between model generations and annotated spans. For this reason, we extract unsupervised metrics during generation while force-decoding the annotated outputs from the MT model to ensure perfect adherence with annotated error spans. In general, such an approach could introduce a problematic confounder in the evaluation, as observed results may be the product of constraining a model towards an unnatural generation, rather than reflecting the underlying phenomena. However, in this study, we carefully ensure that the generation setup matches exactly the one of previous works where the annotated translations were produced, using the same MT model and the same inputs.[13] Hence, the constraining process is a simple insurance of conformity in light of potential discrepancies introduced by different decoding strategies, and does not affect the soundness of our method.

[12] The default XCOMET metric was used with the `unbabel-comet` library (`v2.2.6`).

[13] Generation parameters are not relevant in this setting, provided that they only alter the selection of the following output token, which we do via force-decoding.

Source $_{EN}$	So the challenges in this are already showing themselves. I'm likely going to have a VERY difficult time getting a medical clearance due to the FAA's stance on certain medications.
MT $_{IT}$ (Aya23)	Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům.
Annotator	Takže problémy (minor) s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas (major) dostat lékařské potvrzení kvůli postoji FAA k některým lékům.
XCOMET-XL	Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný (minor) čas dostat (minor) lékařské potvrzení (minor) kvůli postoji FAA k některým lékům
XCOMET-XXL	Takže problémy s tím se již projevují (minor). Pravděpodobně budu mít PŘESNĚ obtížný čas dostat (major) lékařské potvrzení kvůli postoji FAA k některým lékům.
XCOMET-XL $_{CONF}$	Takže (0.23) problémy (0.28) s (0.26) tím (0.28) se (0.17) již (0.19) projevují (0.31) . (0.17) Pravděpodobně (0.23) budu (0.40) mít (0.48) PŘESNĚ (0.79) obtížný (0.65) čas (0.76) dostat (0.64) lékařské (0.50) potvrzení (0.51) kvůli (0.19) postoji (0.34) FAA (0.27) k (0.20) některým (0.20) lékům (0.21) . (0.17)
XCOMET-XXL $_{CONF}$	Takže (0.25) problémy (0.24) s (0.26) tím (0.31) se (0.29) již (0.23) projevují (0.26) . (0.01) Pravděpodobně (0.01) budu (0.03) PŘESNĚ (0.37) obtížný (0.30) čas (0.32) dostat (0.24) lékařské (0.10) potvrzení (0.13) kvůli (0.01) postoji (0.00) FAA (0.00) k (0.00) některým (0.00) lékům (0.00) . (0.00)
Out. Entropy	Takže (0.88) problémy (1.93) s (1.88) tím (0.84) se (1.66) již (1.13) projevují (0.89) . (0.11) Pravděpodobně (0.44) budu (0.22) mít (0.09) PŘESNĚ (2.09) obtížný (3.70) čas (0.09) dostat (1.40) lékařské (1.02) potvrzení (0.64) kvůli (0.69) postoji (0.24) FAA (0.80) k (1.01) některým (0.55) lékům (0.18) . (0.11)

Table 4: Annotated example from the EN→CS portion of the WMT24 dataset. **Top:** Annotator edits with highlighted Error Span Annotation of minor and major errors. **Bottom:** Word-level annotations for best-performing metrics discussed in the study.

Method	**QE4PE$_{t1}$**		**QE4PE$_{t2}$**		**QE4PE$_{t3}$**		**QE4PE$_{t4}$**		**QE4PE$_{t5}$**		**QE4PE$_{t6}$**		**QE4PE$_{avg}$**	
	AP	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***
Random Baseline	.08	.14	.15	.26	.06	.12	.11	.19	.22	.36	.18	.30	.13	.23
Surprisal	.11	.20	.21	.31	.11	.17	.16	.25	.30	.40	.25	.35	.19	.28
Out. Entropy	.12	.18	.22	.30	.10	.16	.17	.24	.30	.39	.26	.34	.19	.27
Surprisal MCD $_{AVG}$	.12	.20	.22	.32	.11	.17	.16	.26	.30	<u>**.41**</u>	.26	<u>.36</u>	.19	.29
Surprisal MCD $_{VAR}$	<u>.13</u>	<u>.21</u>	<u>.26</u>	<u>.33</u>	<u>.12</u>	<u>.20</u>	<u>.19</u>	<u>.27</u>	<u>.31</u>	.40	<u>.29</u>	<u>.36</u>	<u>.22</u>	<u>.30</u>
LL Surprisal $_{BEST}$	.11	.19	.21	.32	.11	.16	.16	.25	.29	.40	.26	.35	.19	.28
LL KL-Div $_{BEST}$	.09	.16	.19	.28	.08	.14	.13	.21	.25	.37	.22	.31	.16	.25
LL Pred. Depth	.09	.16	.18	.28	.07	.13	.14	.21	.25	.37	.21	.31	.16	.24
Attn. Entropy $_{AVG}$	.11	.16	.17	.27	<u>.12</u>	.17	.11	.19	.23	.36	.19	.31	.15	.24
Attn. Entropy $_{MAX}$	.09	.14	.15	.26	.10	.18	.09	.19	.20	.36	.16	.30	.13	.24
BLOOD $_{BEST}$	.08	.14	.16	.26	.06	.12	.11	.19	.23	.36	.18	.30	.14	.23
XCOMET-XL	.11	.24	.22	.35	.10	.20	.16	.30	.27	.35	.23	.34	.18	.30
XCOMET-XL $_{CONF}$	**.20**	.25	.30	**.36**	.14	.21	.25	.31	**.37**	.40	.31	.36	.26	.32
XCOMET-XXL	.13	**.27**	.22	.32	.10	**.24**	.17	.31	.28	.32	.23	.31	.19	.30
XCOMET-XXL $_{CONF}$	.19	**.27**	**.31**	**.36**	**.17**	**.24**	**.26**	**.32**	**.37**	**.41**	**.33**	**.39**	**.27**	**.33**
Human Editors $_{MIN}$	.17	.33	.26	.38	.10	.21	.16	.26	.25	.36	.23	.30	.19	.31
Human Editors $_{AVG}$	.20	.38	.29	.43	.14	.30	.22	.39	.32	.38	.30	.40	.25	.39
Human Editors $_{MAX}$	.24	.43	.31	.47	.20	.41	.24	.43	.37	.50	.33	.50	.28	.46

Table 5: WQE metrics' performance for predicting error spans from the six edit sets over NLLB 3.3B translations in the EN→IT QE4PE dataset (Sarti et al., 2025). <u>Best unsupervised</u> and **overall best** metric results are highlighted.

Method	QE4PE$_{t1}$		QE4PE$_{t2}$		QE4PE$_{t3}$		QE4PE$_{t4}$		QE4PE$_{t5}$		QE4PE$_{t6}$		QE4PE$_{avg}$	
	AP	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***
Random Baseline	.07	.14	.34	.51	.22	.36	.19	.32	.13	.24	.22	.36	.20	.32
Surprisal	.12	.19	.41	.51	.30	.39	.29	.37	.21	.30	.31	.41	.27	.36
Out. Entropy	.11	.18	.41	.51	.31	.37	.29	.36	.20	.27	.31	.39	.27	.35
Surprisal MCD AVG	.12	.19	.42	.52	.31	.40	.30	.40	.21	.30	.31	.42	.28	.37
Surprisal MCD VAR	.13	.21	.45	**.53**	.36	.41	.34	.40	.24	.32	.36	.42	.31	.38
LL Surprisal BEST	.12	.19	.42	**.53**	.30	.40	.29	.38	.21	.30	.31	.41	.27	.37
LL KL-Div BEST	.09	.15	.39	.52	.28	.37	.25	.34	.17	.26	.29	.38	.25	.34
LL Pred. Depth	.09	.16	.37	.52	.26	.37	.24	.33	.17	.25	.27	.38	.23	.33
Attn. Entropy AVG	.09	.15	.37	.51	.22	.36	.20	.32	.13	.24	.23	.37	.21	.32
Attn. Entropy MAX	.09	.15	.35	.51	.22	.36	.18	.32	.12	.24	.21	.37	.19	.32
BLOOD BEST	.07	.13	.35	.51	.22	.36	.19	.32	.14	.24	.23	.36	.20	.32
XCOMET-XL	.13	.27	.39	.39	.31	.44	.28	.32	.20	.35	.31	.44	.27	.38
XCOMET-XL CONF	**.24**	**.31**	.47	**.53**	**.43**	**.45**	**.40**	**.43**	.29	**.36**	**.43**	**.46**	**.38**	**.42**
XCOMET-XXL	.13	.28	.39	.29	.30	.35	.26	.35	.19	.31	.30	.35	.26	.32
XCOMET-XXL CONF	**.24**	.30	**.48**	**.53**	**.43**	**.45**	**.40**	.42	**.31**	.35	**.43**	.45	**.38**	**.42**
Human Editors MIN	.16	.29	.43	.51	.34	.45	.33	.47	.26	.42	.36	.46	.32	.43
Human Editors AVG	.17	.33	.44	.51	.34	.45	.33	.47	.26	.42	.36	.46	.32	.43
Human Editors MAX	.19	.36	.46	.51	.36	.51	.37	.53	.32	.51	.40	.53	.35	.49

Table 6: WQE metrics' performance for predicting error spans from the six edit sets over NLLB 3.3B translations in the EN→NL QE4PE dataset (Sarti et al., 2025). Best unsupervised and **overall best** metric results are highlighted.

Method	Italian		Dutch		Arabic		Turkish		Vietnamese		Ukrainian		Average	
	AP	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***
Random Baseline	.25	.40	.28	.43	.33	.49	.34	.50	.35	.52	.48	.65	.34	.50
Surprisal	.34	.45	.36	.46	.42	.51	.43	.54	.46	.55	.55	.65	.43	.53
Out. Entropy	.37	.43	.39	.45	.45	.50	.49	.52	.48	.54	.58	.65	.46	.51
Surprisal MCD AVG	.34	.45	.37	.47	.43	.52	.44	.54	.46	.55	.56	.65	.43	.53
Surprisal MCD VAR	.39	.46	.41	.47	.47	.53	.49	.55	.48	.55	.61	**.67**	.48	.54
LL Surprisal BEST	.33	.44	.36	.45	.41	.51	.44	.54	.44	.55	.55	.66	.42	.53
LL KL-Div BEST	.34	.42	.37	.45	.41	.51	.44	.52	.44	.52	.56	.65	.43	.51
LL Pred. Depth	.30	.42	.32	.44	.39	.50	.40	.52	.39	.53	.54	.66	.39	.51
Attn. Entropy AVG	.28	.41	.30	.43	.35	.49	.37	.51	.40	.52	.50	.65	.37	.50
Attn. Entropy MAX	.25	.41	.26	.43	.34	.49	.34	.50	.35	.52	.47	.65	.34	.50
BLOOD BEST	.26	.40	.28	.43	.35	.52	.35	.50	.36	.52	.49	.65	.35	.51
XCOMET-XL	.34	.39	.37	.44	.41	.47	.44	.50	.42	.44	.56	.44	.42	.45
XCOMET-XL CONF	.46	.47	.49	**.50**	.51	.53	**.58**	**.56**	.53	.55	.68	**.67**	.54	**.55**
XCOMET-XXL	.34	.36	.35	.35	.43	.47	.45	.48	.43	.42	.57	.41	.43	.42
XCOMET-XXL CONF	**.48**	**.49**	**.50**	**.50**	**.55**	**.54**	**.58**	**.56**	**.56**	**.57**	**.70**	**.67**	**.56**	**.55**

Table 7: WQE metrics' performance for predicting error spans from multiple edit sets (one per language) over mBART-50 translations across the six topologically diverse target languages of DIVEMT (Sarti et al., 2022).

Method	En→ Ja		En→ Zh		En→ Hi		Cs→ Uk		En→ Cs		En→ Ru		Average	
	AP	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***	**AP**	**F1***
Random Baseline	.02	.03	.03	.07	.03	.07	.05	.09	.06	.11	.08	.16	.05	.09
Surprisal	.03	.07	.05	.09	.05	.09	.14	.20	.10	.16	.13	.19	.08	.13
Out. Entropy	.03	.08	.06	.11	.06	.10	**.20**	.27	.12	.18	.14	.20	.10	.16
LL Surprisal BEST	.03	.07	.05	.09	.05	.09	.14	.20	.10	.16	.13	.19	.08	.13
LL KL-Div BEST	.02	.05	.04	.07	.04	.08	.10	.17	.09	.15	.12	.19	.07	.12
LL Pred. Depth	.02	.05	.04	.08	.04	.09	.09	.18	.08	.14	.11	.18	.06	.12
Attn. Entropy AVG	.02	.03	.03	.07	.03	.07	.03	.09	.05	.11	.07	.16	.04	.09
Attn. Entropy MAX	.01	.03	.03	.07	.03	.07	.03	.09	.05	.11	.08	.16	.04	.09
XCOMET-XL	.04	.09	.05	.11	.06	.12	.13	.28	.11	.24	.16	.32	.09	.19
XCOMET-XL CONF	**.08**	.14	**.10**	.16	**.10**	**.19**	.18	**.30**	.19	.29	.24	.32	.15	.23
XCOMET-XXL	.04	.11	.06	.13	.05	.11	.13	.28	.11	.24	.16	**.33**	.09	.20
XCOMET-XXL CONF	.07	**.15**	.09	**.19**	.09	.17	.19	.29	**.22**	**.30**	**.28**	**.33**	**.16**	**.24**

Table 8: WQE metrics' performance for predicting error spans from the ESA annotations (one set per language) over Aya23-35B outputs for the WMT24 dataset (Kocmi et al., 2024a).

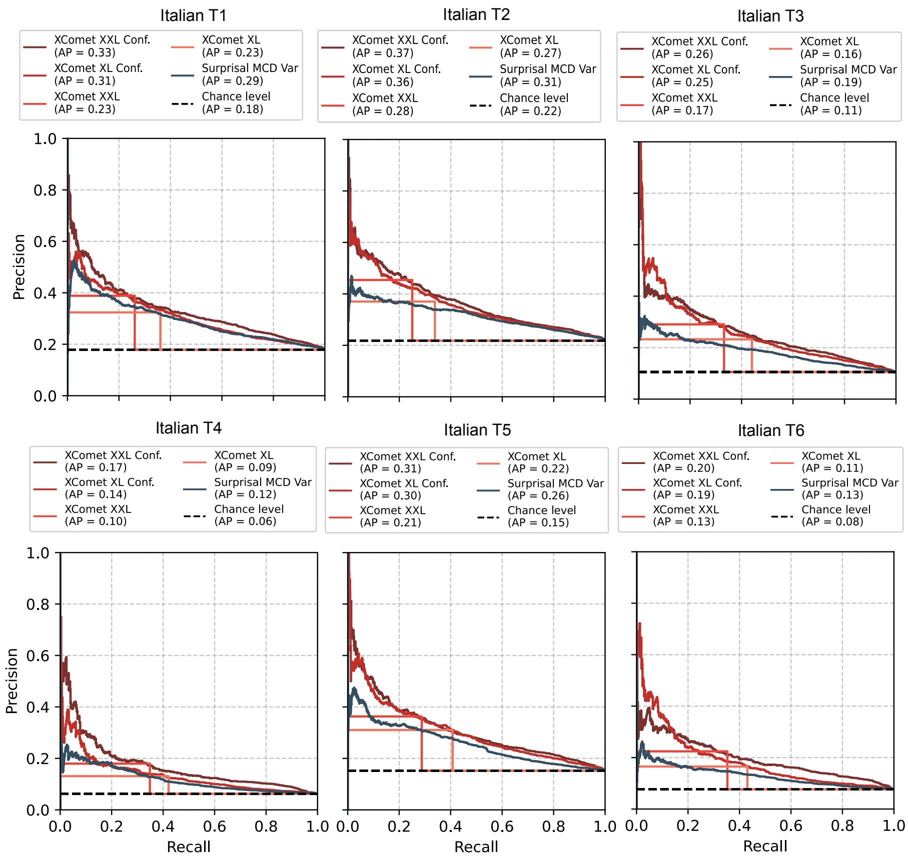

Figure 4: Precision-recall curves for XCOMET metrics and Surprisal MCD_{VAR} for all annotators of QE4PE EN→IT.

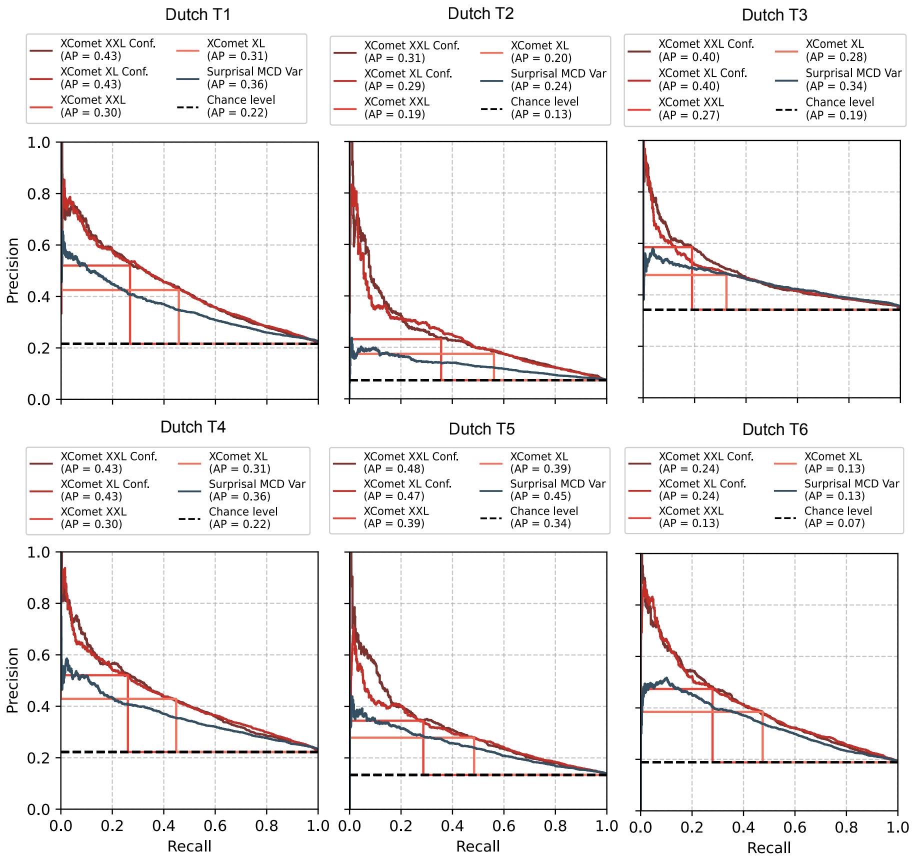

Figure 5: Precision-recall curves for XCOMET metrics and Surprisal MCD_{VAR} for all annotators of QE4PE EN→NL.

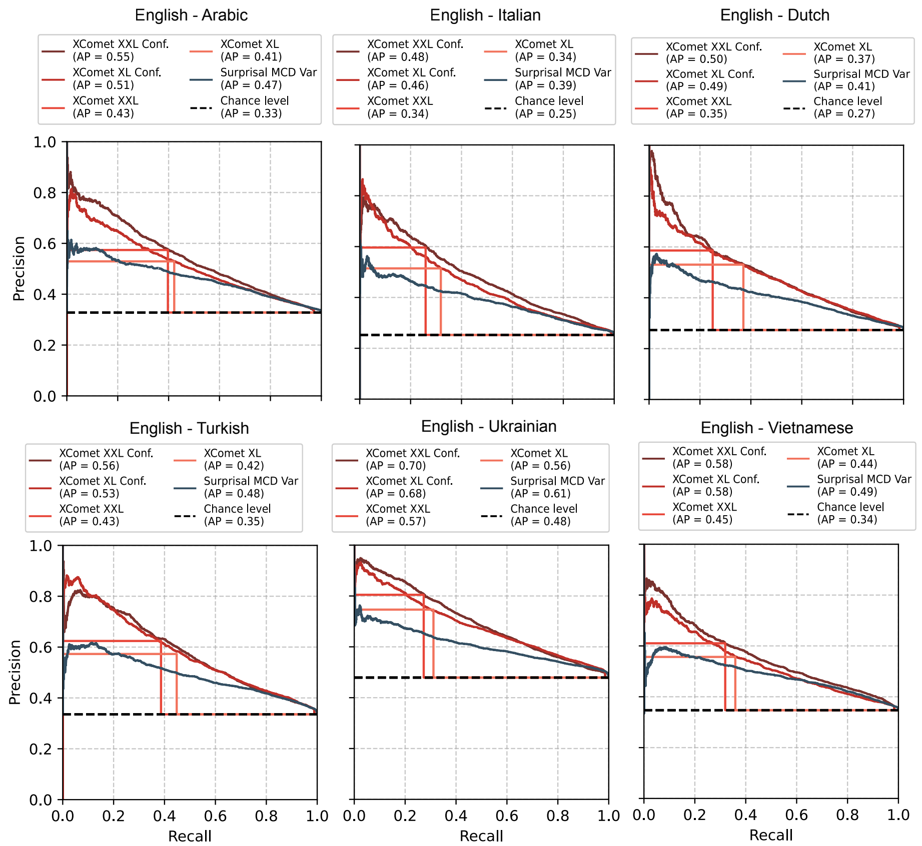

Figure 6: Precision-recall curves for XCOMET metrics and Surprisal MCD_{VAR} on all DIVEMT languages.

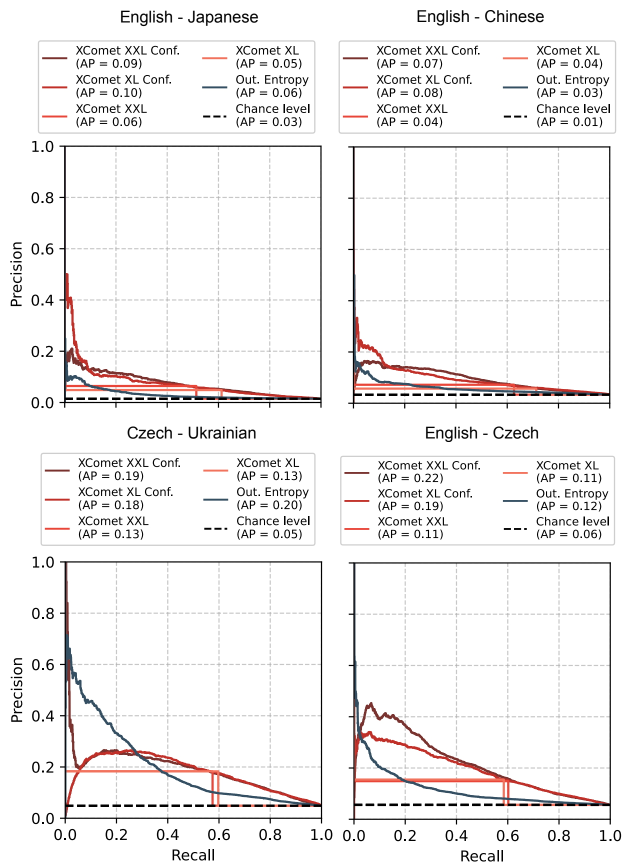

Figure 7: Precision-recall curves for XCOMET metrics and Out. Entropy on all WMT24 languages.