EXPLAINABLE CED: A Dataset for Explainable Critical Error Detection in Machine Translation

Dahyun Jung[1], **Sugyeong Eo**[1], **Chanjun Park**[2†], **Heuiseok Lim**[1†]

[1]Korea University, [2]Upstage AI
{dhaabb55, djtnrud, limhseok}@korea.ac.kr
chanjun.park@upstage.ai

Abstract

Critical error detection (CED) in machine translation is a task that aims to detect errors that significantly distort the intended meaning. However, the existing study of CED lacks explainability due to the absence of content addressing the reasons for catastrophic errors. To address this limitation, we propose EXPLAINABLE CED, a dataset that introduces the attributes of error explanation and correction regarding critical errors. Considering the advantage of reducing time costs and mitigating human annotation bias, we leverage a large language model in the data construction process. To improve the quality of the dataset and mitigate hallucination, we compare responses from the model and introduce an additional data filtering method through feedback scoring. The experiment demonstrates that the dataset appropriately reflects a consistent explanation and revision for errors, validating the reliability of the dataset.

SRC	But something more fundamental is at play: Brazil may finally be overcoming some of the deepest obstacles to its economic development, obstacles that held the country back for decades.
MT	Noch etwas Grundsätzlicheres jedoch zeichnet sich ab: Großbritannien ist möglicherweise endlich dabei, einige der größten Hürden für seine wirtschaftliche Entwicklung zu überwinden - Hürden, die das Land über Jahrzehnte hinweg zurückgeworfen haben. (But something more fundamental is emerging: Britain may finally be overcoming some of the biggest hurdles to its economic development - hurdles that have set the country back for decades.)
Error Type	NAM
Error Explanation	The translation introduces a mistranslation by replacing "Brazil" with "Großbritannien" (Great Britain).
Error Correction	Noch etwas Grundsätzlicheres jedoch zeichnet sich ab: Brasilien ist möglicherweise endlich dabei, einige der größten Hürden für seine wirtschaftliche Entwicklung zu überwinden - Hürden, die das Land über Jahrzehnte hinweg zurückgeworfen haben. (But something more fundamental is emerging: Brazil may finally be overcoming some of the biggest hurdles to its economic development - hurdles that have set the country back for decades.)

Table 1: An example of the EXPLAINABLE CED. **SRC** is the source sentence in English, and **MT** is the MT sentence in German.

1 Introduction

Critical error detection (CED) is a sub-task of quality estimation (QE) that aims to identify sentences where the intended meaning from the source text is distorted due to catastrophic errors in machine translation (MT) systems (Specia et al., 2021; Zerva et al., 2022). These distortions potentially lead to offensive interpretations or cause social, legal, or economic issues. While critical errors are infrequent and can be considered a long-tail problem, it is essential to prevent issues caused by them, thereby emphasizing the importance of CED in ensuring the quality of MT systems (Raunak et al., 2022).

However, the existing binary classification approach of the CED, which detects the presence of fatal errors, merely blocks the erroneous output from the MT system. Users, not native speakers of the target language, cannot specify the description for the error and its solutions (Sharou and Specia, 2022). These limitations highlight the necessity for a comprehensive approach to address critical errors to provide more precise guidance for non-native users (Fomicheva et al., 2021a; Hase and Bansal, 2020).

In this regard, we propose a novel EXPLAINABLE CED dataset that includes descriptions of portions significantly mistranslated from the original intention and the corrected text that aligns with the intended meaning. Each instance in EXPLAINABLE CED consists of the source sentence, target sentence, error type, error explanation, and sentence with the error corrected. To develop the dataset, we use a large language model (LLM)-based method. By leveraging the LLM, we can further reduce the time and computational resources in the data collection process. This approach al-

† Corresponding Author

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 25–35

leviates issues associated with the inconsistency between human annotators and inherent biases that are uncontrollable (Kruglanski and Ajzen, 1983; Pronin, 2007; Ntoutsi et al., 2020; Ouyang et al., 2022). Additionally, LLM not only exhibits exceptional performance across overall MT tasks (Vidal et al., 2022; Lu et al., 2023; Raunak et al., 2023) but also demonstrates more efficient capabilities in data labeling compared to humans (Chen et al., 2023; He et al., 2023). However, as LLMs still face challenges related to hallucinations (Bang et al., 2023), our focus on data generation is on mitigating hallucinations. When hallucinations are incorporated in the responses of LLMs, the responses may differ from each other and encompass potentially contradictory information (Liu et al., 2022; Wang et al., 2023; Manakul et al., 2023). To mitigate hallucinations in LLMs, we adopt a method that compares responses to various prompts and selects consistent instances to enhance coherence. Furthermore, we aim to improve the quality of the data by filtering it based on feedback scores.

In the experiment, we introduce supplementary inputs to investigate the mitigation of the hallucination in the dataset. The results reveal that each instance in the dataset is structured to retain mutually similar semantics, indicating that the dataset is constructed to reflect the hallucination mitigation strategy. We hope that this research will offer solutions to critical errors and aid in future studies aimed at improving the reliability of MT.

2 Related Works

The conference on machine translation (WMT) in 2021 introduces the CED for QE (Specia et al., 2021). Jiang et al. (2021) proposes a classifier that adds sampling to handle unbalanced data to detect critical errors and integrates existing techniques for finding errors. Rubino et al. (2021) introduces a system that uses pre-trained XLM-R as a predictor and stacked FFN layers as a binary classifier and uses commercial machine translation tools to help detect errors. Eo et al. (2022) utilizes prompt-based fine-tuning, combining demonstration and commercial machine translation systems to perform the classification task.

Explainable QE is an explainability subtask following its first edition at Eval4NLP 2021 (Fomicheva et al., 2021b). The interpretability of QE systems may be compromised due to their reliance on models with numerous parameters. To address this issue and maintain user trust, explainable QE is proposed (Fomicheva et al., 2021b). Tao et al. (2022) proposes the sentence-level QE model's predictor as a feature extractor for sentence word embeddings and utilizes the inverse value of maximum similarity between each word in the target and the source as the word translation error risk value. From a different perspective, perturbation-based QE proposes an unsupervised word-level QE approach for evaluating black-box MT systems (Dinh and Niehues, 2023). The knowledge-prompted estimator employs the chain of thought prompting method to provide enhanced interpretability for QE (Yang et al., 2023).

As evidenced in previous studies, CED primarily focuses on classifying binary labels that indicate the presence or absence of errors. A limitation of this binary classification is that it only enables the MT system to prevent the presentation of translation results. Consequently, users fail to receive translated outputs and struggle to understand and recognize the errors that occur correctly. To address this limitation, we propose a task that allows users to comprehend and accept the critical errors that arise and provides them with corrected translations.

3 EXPLAINABLE CED

In this section, we introduce a detailed description of the components constituting the EXPLAINABLE CED dataset and a methodology for constructing the dataset through a three-phase process, considering consistency and hallucination. The dataset contains three elements to explain translation errors when given a source sentence and an MT sentence containing the error (Table 1). The components are designed as follows:

Error Type refers to a categorized label reflecting the characteristics of errors. When multiple errors are present, we prioritize and address only the most severe ones. We adopt the categories defined by Specia et al. (2021) as follows:

- **Toxicity (TOX)** is associated with hate speech and aggressive language, which varies based on individual, race, gender, etc. Such errors manifest either through the introduction of toxicity in the MT sentence when the source sentence is devoid of toxicity or through the complete removal of toxicity in the translation when the source sentence contains it.

- **Safety (SAF)** can lead to potential safety risks

	Error Type						Error Explanation		Error Correction	
	TOX	SAF	NAM	SEN	NUM	ETC	Sentences	Tokens	Sentences	Tokens
En-De	2,096	520	3,793	512	1,676	76	8,673	382,941	8,673	294,239
En-Cs	861	10	533	38	4	4	1,450	66,175	1,450	30,274
En-Zh	559	11	611	47	9	6	1,243	64,565	1,243	23,545
En-Ja	444	14	263	33	7	10	771	31,172	771	12,895

Table 2: Statistics of the dataset for four language pairs. **Error Type** presents the number of instances for each category. **Error Explanation** and **Error Correction** display the number of sentences and tokens.

for readers, as it constitutes a translation error. The errors may occur when content not present in the source sentence is introduced in the translation or when content from the source sentence is omitted.

- **Named Entity (NAM)** occurs when named entities are mistranslated, omitted, or not translated in the target sentence. If it can be determined that the term is a user's name, then it is considered a named entity error. A partially translated named entity is not considered a critical error if it can be understood to refer to the same entity.
- **Sentiment (SEN)** occurs when the sentiment of a sentence is reversed. However, a sentiment error does not necessarily have to indicate a complete negation. For example, changing "possibly" to "with certainty" constitutes a sentiment error.
- **Number (NUM)** is related to numbers. Such errors manifest as either mistranslated numbers or the omission of numbers in the source sentence within the translation sentence.
- **Et Cetera (ETC)** doesn't belong to any of the five categories above, but seriously compromises the original text's meaning.

Error Explanation refers to a description in natural language that details the occurrence of errors in MT sentences. This includes explicit instances indicating which part of the sentence contains the translation error. Beyond the labeled instances, the explanation offers a profound insight into the cause and characteristics of the problem, thereby heightening the awareness of the error's severity.

Error Correction refers to the revised sentence where the translation sentence, which distorted the original meaning, is corrected. The correction aims to modify the erroneous parts in the translation sentence with the least amount of editing.

3.1 Data Collection

We use the CED dataset publicly released at WMT21 and 22 (Specia et al., 2021; Zerva et al., 2022)[1]. We structure the EXPLAINABLE CED dataset by annotating sentences with critical errors based on the pre-constructed dataset. Our dataset comprises language pairs of English-German (En-De), English-Czech (En-Cs), English-Chinese (En-Zh), and English-Japanese (En-Ja). We split the dataset into train/validation/test subsets with a ratio of 80%/10%/10%. The statistics of the dataset are presented in Table 2, and examples can be found in Appendix A.

We employ ChatGPT (OpenAI-Blog, 2022) (gpt-3.5-turbo) to construct our dataset. All instructions used in the construction of the dataset are disclosed in Appendix B. Our approach to data generation is based on the following incremental framework:

1) Selecting the Category We configure the type based on the properties of errors. To minimize inconsistencies that arise from identical requests and enhance the reliability of the dataset, we measure the agreement among multiple responses. Potential discrepancies caused by variations in prompt format are considered. We utilize three distinct instructions to extract types by feeding the model with source and target sentences. By comparing these outputs, we identify the type that garners majority agreement. For instance, if the model's responses are TOX, NUM, and NUM, we annotate with NUM, as it holds the majority consensus.

2) Generating the Description In this phase, we employ three methods to mitigate hallucinations and enhance the quality of the explanation. First, we structure the model's input by providing not only the source and MT sentences but also the type generated from the previous stage. This approach allows the generation of explanations aligned with specific error types, ensuring semantic consistency

[1]This dataset is based on the Wikipedia comment domain, which has a high percentage of TOX and NAM errors.

Input	En-De ACC	En-De F1	En-Cs ACC	En-Cs F1
SRC+MT	81.83	59.21	82.35	33.73
SRC+MT+EXP	89.47	**72.78**	**94.12**	**38.56**
SRC+MT+COR	83.51	61.84	88.24	36.14
SRC+MT+EXP+COR	**90.84**	71.22	**94.12**	**38.56**
Input	**En-Zh ACC**	**En-Zh F1**	**En-Ja ACC**	**En-Ja F1**
SRC+MT	85.71	55.98	69.23	29.12
SRC+MT+EXP	**94.81**	**64.66**	76.92	32.46
SRC+MT+COR	88.31	58.97	74.36	30.99
SRC+MT+EXP+COR	93.51	63.55	**79.49**	**33.70**

Table 3: Performance comparison of models based on input differences in error type classification experiments. The best result is in **bold**. **EXP** is error explanation and **COR** is error correction.

in the sentences produced by the model. Second, to improve the quality of the generation, we apply a self-refine approach (Madaan et al., 2023). By leveraging the model's internal feedback, we refine the generated outcomes. Third, we compare the two explanation sentences produced and purified using distinct instructions to minimize disparities in model responses. We select the sentence with the highest model preference score from those sentences that fall within the top 20% in terms of both similarity and model preference scores. The similarity score is measured using mSimCSE (Wang et al., 2022), while the model preference score is assessed using the GPT score (Liu et al., 2023).

3) Post-editing the Translation Generating error-correcting translation sentences considers the type and description generated in the previous steps. This process is handled in a similar way to step 2.

4 Experiments

We investigate the experimental results for three tasks using the EXPLAINABLE CED dataset. We validate the dataset with a focus on whether hallucination is mitigated due to low-quality data being filtered out[2]. To verify that the dataset contains consistent content, we conduct experiments incorporating each dataset element as input. This assumes that if adding each dataset component to the input yields a positive impact, it suggests that the dataset is composed of consistent responses.

[2]Appendix C shows the improvement in GPT score performance following the dataset construction process.

4.1 Error Type Classification

We conduct experiments to categorize types of errors. The model $f(y_{type} \mid x_{src}, x_{err})$ outputs a probability distribution over error types y_{type} when given a source sentence x_{src} and its corresponding translation with errors x_{err}, where y_{type} represents potential translation error types: TOX, SAF, NAM, SEN, NUM, ETC. This model enables the automatic classification of error types in the translation.

We experiment by incorporating additional components from our dataset, such as error explanation and correction, as inputs. In this context, the model is represented as $f(y_{type} \mid x_{src}, x_{err}, x_{exp})$, $f(y_{type} \mid x_{src}, x_{err}, x_{cor})$, and $f(y_{type} \mid x_{src}, x_{err}, x_{exp}, x_{cor})$, where x_{exp} denotes the error explanation sentence, while x_{cor} refers to the sentence with the error corrected.

Experiment Settings For training, we use XLM-RoBERTa (Conneau et al., 2020). The model is implemented with PyTorch[3] and Hugging Face[4]. We utilize the pre-trained language models 'xlm-roberta-large' checkpoints. We use a batch size 64, the Adam optimizer with a learning rate 2e–5, and train for ten epochs. The experiments are performed on an NVIDIA RTX A6000 environment. For evaluating the multi-label classification performance, we employ accuracy and F1 score.

Results and Discussions Table 3 is the experimental results to classify the categories of critical errors. The results demonstrate the efficacy of models considering explanations or corrections, compared to the baseline performance that only takes into account the source and MT. Across all the language pairs, performance improves when additional input is incorporated, indicating maintained alignment between the data. Notably, the inclusion of EXP resulted in an increase of 13.57 in the F1 score for En-De, 4.83 for En-Cs, and 8.68 for En-Zh, suggesting the meaningful utility of EXP in error analysis. However, for En-Ja, the combination of EXP+COR yielded the best results. This indicates that while error explanations alone can offer valuable insights, pairing them with corrections in the dataset can produce synergistic effects for specific languages.

4.2 Error Explanation Generation

The experiments involve examining the source sentence and its mistranslated version, and then ex-

[3]https://pytorch.org/

[4]https://huggingface.co/

Input	En-De		En-Cs		En-Zh		En-Ja	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
SRC+MT	5.89	27.14	1.02	15.19	2.44	14.80	**4.11**	**15.52**
SRC+MT+TYPE	**11.95**	**29.34**	4.43	16.00	3.60	16.23	2.04	11.90
SRC+MT+COR	11.94	28.84	4.53	21.79	4.73	22.18	2.69	11.60
SRC+MT+TYPE+COR	11.67	28.66	**4.67**	**22.59**	**8.05**	**23.97**	2.04	6.45

Table 4: Performance comparison of models based on input differences in error explanation generation

Input	En-De			En-Cs			En-Zh			En-Ja		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
SRC+MT	**53.08**	**70.64**	76.87	14.56	27.02	**50.42**	3.59	14.51	**54.33**	3.79	13.76	**48.12**
SRC+MT+TYPE	52.51	**70.64**	76.87	15.51	**27.41**	50.38	3.42	17.38	50.93	15.51	**27.51**	47.67
SRC+MT+EXP	52.94	70.63	**77.10**	**16.05**	26.24	49.95	**6.98**	**18.83**	49.20	**16.05**	26.67	43.98
SRC+MT+TYPE+EXP	52.56	70.61	76.97	13.57	24.33	48.49	3.42	17.71	50.93	13.57	24.33	37.66

Table 5: Performance comparison of models based on input differences in error correction generation

plaining the errors in the translation sentence. The model is trained to pinpoint the errors in the translation and describe the details of those errors in natural language. We also add experiments that include error type and correction sentences as additional input to assess the consistency of the dataset.

Experiment Settings We train using mT5 (Xue et al., 2021) and utilize 'google/mt5-base' checkpoints. We use a batch size 32, the Adam optimizer with a learning rate 1e–4, and train for 20 epochs. For evaluation, we employ metrics such as BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004).

Results and Discussions We present the results of experiments in generating error explanation sentences in Table 4. The TYPE yields the highest scores in both BLEU and ROUGE in the En-De language pair. The En-Cs and En-Zh language pairs exhibit higher performances when using both TYPE and COR. This indicates that the addition of information positively impacts the task of generating error explanations. Therefore, it can be demonstrated that the data are consistent within each language pair. For the En-Ja, including other input has a detrimental effect. This may be attributed to the limited amount of training data, suggesting that the model might not have adequately learned to incorporate supplementary information in longer natural language sentences.

4.3 Error Correction Generation

We design experiments to correct mistranslations that semantically align with the original text. We also add experiments that include error type and explanation sentence as input.

Experiment Settings This is the same as in Section 4.2, except that we consider a metric, COMET-22 (Rei et al., 2022). COMET-22 takes into account different types of human judgments.

Results and Discussions Table 5 shows the results of generating sentences with corrected critical errors in translation. The experiments show that BLEU and ROUGE exhibit different patterns compared to COMET. For the En-De, the baseline achieves higher BLEU and ROUGE, which measure word overlap, while the EXP and TYPE+EXP, which consider additional schemes, demonstrate better performance in terms of COMET that reflects human judgments. This suggests that EXP can help address the semantic aspects of translation post-editing. However, the opposite trend is observed in other languages compared to En-De. Performance improvements are significant in error type classification and explanation generation due to additional inputs, while not so in this task, underscores the greater challenge of error correction over detection.

5 Conclusion

We introduced EXPLAINABLE CED dataset to provide explainability for critical errors in MT. This dataset offered descriptions of errors across error types and fixing them. In constructing the dataset, we proposed a framework for leveraging LLM. The objective was to mitigate the hallucination by maintaining consistency in the model's responses and to enhance the quality of the generation by the self-refine. The results indicated that our dataset maintains consistency despite being generated by various model responses.

Limitations

Our dataset exhibited an imbalance in language pairs and category labels. This was primarily due to the difficulty in collecting translations containing critical errors, which occur sparsely. We constructed our dataset utilizing the maximum available data and plan to supplement our dataset with additional data containing critical errors in the future.

This study utilized the ChatGPT for constructing our dataset rather than the superior-performing GPT-4 (OpenAI, 2023). This decision was primarily driven by cost and time considerations. Deploying GPT-4 would have incurred approximately 15 times the expense of ChatGPT. As a result, we opted for ChatGPT to minimize costs, and the actual expenditure for building the dataset was around $40. While ChatGPT is efficient, it may not capture the depth and nuance that GPT-4 potentially offers. While we have employed ChatGPT in this context and have invested efforts in instruction tuning and methods to mitigate hallucination, there are inherent trade-offs.

Ethics Statement

MT systems serve as crucial means of conveying information. However, erroneous or misleading information may be propagated due to translation errors. For instance, mistranslations can potentially give culturally sensitive or offensive content and infringe on individuals' privacy by exposing personal information. Our task aims to prevent such severe consequences and enhance the reliability of MT systems. Furthermore, we employed the LLM designed to adhere to ethical guidelines and principles. In instances of significant toxicity, the data were marked as containing offensive and toxic content. Consequently, from an ethical standpoint, the model automatically filtered out potentially concerning portions of the dataset.

Acknowledgements

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03045425)

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Jiao Chen, Luyi Ma, Xiaohan Li, Nikhil Thakurdesai, Jianpeng Xu, Jason H. D. Cho, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. Knowledge graph completion models are few-shot learners: An empirical study of relation labeling in e-commerce with llms.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Tu Anh Dinh and Jan Niehues. 2023. Perturbation-based qe: An explainable, unsupervised word-level quality estimation method for blackbox machine translation. *arXiv preprint arXiv:2305.07457*.

Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim. 2022. KU X upstage's submission for the WMT22 quality estimation: Critical error detection shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 606–614, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021a. The eval4nlp shared task on explainable quality estimation: Overview and results.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021b. 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior?

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. Annollm: Making large language models to be better crowdsourced annotators.

Genze Jiang, Zhenhao Li, and Lucia Specia. 2021. ICL's submission to the WMT21 critical error detection shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 928–934, Online. Association for Computational Linguistics.

Arie W Kruglanski and Icek Ajzen. 1983. Bias and error in human judgment. *European Journal of Social Psychology*, 13(1):1–44.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356.

OpenAI. 2023. Gpt-4 technical report.

OpenAI-Blog. 2022. Chatgpt: Optimizing language models for dialogue.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Emily Pronin. 2007. Perception and misperception of bias in human judgment. *Trends in cognitive sciences*, 11(1):37–43.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. Salted: A framework for salient long-tail translation error detection. *arXiv preprint arXiv:2205.09988*.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing.

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

Raphael Rubino, Atsushi Fujita, and Benjamin Marie. 2021. NICT Kyoto submission for the WMT'21 quality estimation task: Multimetric multilingual pretraining for critical error detection. In *Proceedings of the Sixth Conference on Machine Translation*, pages 941–947, Online. Association for Computational Linguistics.

Khetam Al Sharou and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180, Ghent, Belgium. European Association for Machine Translation.

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, Online. Association for Computational Linguistics.

Shimin Tao, Su Chang, Ma Miaomiao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo, Minghan Wang, and Yinglu Li. 2022. CrossQE: HW-TSC 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 646–652, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Blanca Vidal, Albert Llorens, and Juan Alonso. 2022. Automatic post-editing of MT output using large language models. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 84–106, Orlando, USA. Association for Machine Translation in the Americas.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Yau-Shian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Hao Yang, Min Zhang, Shimin Tao, Minghan Wang, Daimeng Wei, and Yanfei Jiang. 2023. Knowledge-prompted estimator: A novel approach to explainable machine translation assessment.

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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Dataset Examples by Language Pairs

We present examples of the dataset for each language pair. Table 6 illustrates a toxicity error, where toxic words appear in En-De. Table 7 shows a translation in En-Cs that omits the entity. Table 8 displays an example of a mistranslated number in En-Zh. Table 9 presents an example of a sentiment error in En-Ja that reverses the speaker's intention.

B Prompt Examples

B.1 Data Generation Prompt

The design of appropriate prompts is important for LLM performance. We compare prompts generated by humans and LLMs to devise an effective design strategy. We create four prompts, two from each of the two categories, and generate 100 examples for each prompt. We compare the GPT scores for each example to select the most effective prompts. Through this process, we identify the optimal prompt and effectively utilize the performance of the LLM. Table 10 presents the prompts used for data generation.

B.2 Feedback and Evaluation Prompt

Table 11 is utilized for the self-refine method and evaluating GPT scores. For the self-refine method, feedback sentences are employed, while in the GPT score assessment, the total explanation score and total modified translation score are used to evaluate the explanation and correction, respectively.

C Measuring the Efficiency of the Generation Process

Figure 1 illustrates the changes in the GPT scores for both self-refine and data selection through comparison. This demonstrates that the self-refine process positively impacts both EXP and COR quality. Furthermore, we can confirm that the latter method is significantly more meaningful by comparing the performance between outputs generated from a single prompt and those selected from multiple prompts.

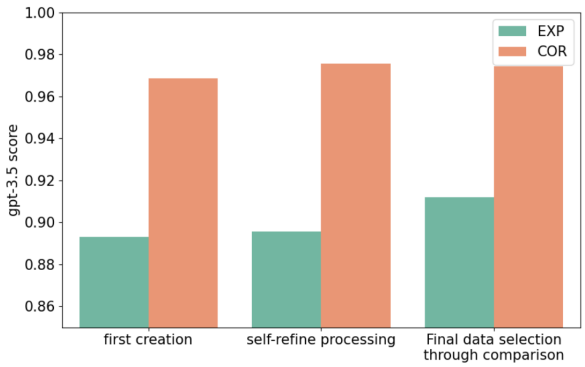

Figure 1: Variation in data quality scores during generation process

	En-De
SRC	The sooner you people just cut the crap and get on with writing wikipedia the encyclopedia and hang up your sheriff badges the better.
MT	Je früher Sie Leute einfach die Scheiße zu schneiden und mit dem Schreiben von wikipedia die Enzyklopädie und hängen Sie Ihre Sheriff Auszeichnungen , desto besser . (The sooner you people just cut the shit and start writing wikipedia's encyclopedia and hang up your sheriff awards, the better.)
Error Type	TOX
Error Explanation	The translation includes a profanity "Scheiße" that is not present in the source sentence.
Error Correction	Je früher Sie Leute einfach den Unsinn lassen und mit dem Schreiben von wikipedia die Enzyklopädie und hängen Sie Ihre Sheriff Auszeichnungen, desto besser. (The sooner you people just drop the nonsense and start writing wikipedia's encyclopedia and hang up your sheriff awards, the better.)

Table 6: An example TOX type for the English-German language pair

	En-Cs
SRC	Yep, I'm Catholic. Convert since 1998. I go to St. John's. You?
MT	Jo , jsem katolík , předělávám se od roku 1998 . (Yeah, I'm Catholic, I've been converting since 1998.)
Error Type	NAM
Error Explanation	The translation sentence contains a mistranslated named entity. The source sentence mentions "St. Johnś", which is a proper name for a church or a location. However, in the translation sentence, there is no mention of "St. Johnś" and it is not clear what the translator intended to translate it as.
Error Correction	Jo, jsem katolík. Konvertita od roku 1998. Chodím do kostela sv. Jana. A ty? (Yeah, I'm Catholic. Convert since 1998. I go to St. John's. And you?)

Table 7: An example NAM type for the English-Czech language pair

	En-Zh
SRC	and page is getting 76K hits an hour as a result.©
MT	果 , 面 每 小 收到 76 万次 点 。(As a result, the page receives 760,000 hits per hour.)
Error Type	NUM
Error Explanation	In the source sentence, it states "76K hits an hour", which means 76,000 hits per hour. However, in the translation sentence, it appears as "76 万次 点", which translates to "760,000 hits". This is an incorrect translation of "76K" and may result in a major deviation from the intended meaning.
Error Correction	果 , 面每小收到76K次点。(As a result, the page received 76K hits per hour.)

Table 8: An example NUM type for the English-Chinese language pair

	En-Ja
SRC	Yeah, I'd prefer if you stop ignoring my request and act upon it, not wait until it expires automatically.
MT	はい、私 は あなた が 私 の 要求 を 無視 し 、それ に う こと を 停止 し 、それ が 自動 的 に 終了 する まで 待つ こと を 好む でしょう 。(Yes, I would prefer that you ignore my request, stop complying with it, and wait until it is automatically terminated.)
Error Type	SEN
Error Explanation	In the source sentence, the speaker is expressing a preference for someone to stop ignoring their request and to act upon it before it expires automatically. However, in the translation sentence, the sentiment is reversed and it appears as if the speaker prefers the other person to ignore their request and wait until it expires automatically. This is a deviation in sentiment polarity that completely changes the meaning of the original sentence.
Error Correction	はい、私はあなたが私の要求を無視するのをやめて、それにし、自動的に期限が切れるのを待つのではなく、すぐにすることを好みます。(Yes, I prefer that you stop ignoring my request and address it immediately instead of waiting for it to expire automatically.)

Table 9: An example SEN type for the English-Japanese language pair

Translations with critical errors are defined as translations that deviate in meaning as compared to the source sentence in such a way that they are misleading and may carry health, safety, legal, reputation, religious, or financial implications.

{Critical Error Category}

Read the translation of the source sentence and perform the three tasks below. Please read the instructions carefully before completing the task.

- **Error Type**: Please indicate which category the critical error in the translation sentence belongs to. If you find multiple categories of errors, please indicate only the most serious one, and if it does not belong to any category, please indicate "ETC". If there is no error, mark it as "NOT" and do nothing further.

- **Error Explanation**: Please explain why the error occurred. Please describe the single most serious error from the error category, and be concise in no more than two sentences, including examples.

- **Error Correction**: Please fix the error in the translation sentence, minimizing it to the part where the critical error mentioned in the description appears.

Table 10: Prompt for generating each scheme. {Critical Error Category} is the description of error type in Section 3.

In machine translation, a critical error is an error that completely changes the meaning of the source text. Explanation is a sentence that explains why this error occurred. This helps us understand what caused the error and helps us avoid similar mistakes in the future. Correction is a sentence that corrects the erroneous translation. This is the process of fixing the translation to correctly reflect the source text's exact meaning. Based on the original and translation sentences, your work evaluates and scores the explanation and correction. Please make sure you read and understand these instructions carefully.

∗ **Explanation Scoring** ∗

- Specificity: Judge whether the explanation of translation errors is detailed and illustrated with examples. A score of 5 indicates a detailed explanation with examples, while a score of 1 indicates a less detailed explanation.

- Severity: Determine whether the explanation describes a critical error that distorts the meaning of the original text. A score of 5 indicates that the explanation describes a critical error, while a score of 1 indicates that the explanation describes an error that is not critical.

- Understandability: Score if the translation is described in a way that makes it easy to understand what is wrong with the translation. A score of 5 indicates that the explanation is easy to understand, while a score of 1 indicates that the explanation is difficult to understand.

- Brevity: The explanation should not include unnecessary information that does not help you understand the error. A score of 5 indicates that the explanation does not contain unnecessary information, while a score of 1 indicates that the explanation does contain unnecessary content.

- Focus: The explanation should focus on errors that appear in the translation, not to consider errors in the source itself. A score of 5 indicates that the explanation accounts for errors present in the translation, while a score of 1 indicates that the explanation only accounts for errors in the source sentence.

∗ **Modified Translation Scoring** ∗

- Semantic preservation: Determine whether the modified translation accurately reflects the meaning of the source text. A score of 5 indicates a translation that completely preserves the original meaning, while a score of 1 indicates a translation that significantly distorts or loses the original meaning.

- Error: Determine whether the corrected translation is free of critical errors. A score of 5 indicates no critical errors, while a score of 1 indicates many critical errors.

- Minimal editing: Indicates how much the translation had to be edited to correct the error. A score of 5 means that the original sentence required minimal editing, while a score of 1 means that the translation required significant editing.

- Naturalness: Rate the extent to which the corrected translation is a natural sentence in the target language. A score of 5 means that the sentence is very natural and fluent, while a score of 1 means that the sentence is awkward or unnatural.

- Reflectivity: Judge whether all the errors in the explanation have been corrected in the revised translation. A score of 5 indicates that all errors have been corrected, while a score of 1 indicates that many corrections have not been incorporated.

Table 11: Prompt for evaluating the generated error description and correction