TMU-HIT's Submission for the WMT24 Quality Estimation Shared Task: Is GPT-4 a Good Evaluator for Machine Translation?

Ayako Sato[†], **Kyotaro Nakajima**[†], **Hwichan Kim**[†]
Zhousi Chen[‡], **Mamoru Komachi**[‡]
[†]Tokyo Metropolitan University, ‡Hitotsubashi University
{sato-ayako, nakajima-kyotaro, kim-hwichan}@ed.tmu.ac.jp
{zhousi.chen, mamoru.komachi}@hit-u.ac.jp

Abstract

In machine translation quality estimation (QE), translation quality is evaluated automatically without the need for reference translations. This paper describes our contribution to the sentence-level subtask of Task 1 at the Ninth Machine Translation Conference (WMT24), which predicts quality scores for neural MT outputs without reference translations. We fine-tune GPT-4o mini, a large-scale language model (LLM), with limited data for QE. We report results for the direct assessment (DA) method for four language pairs: English-Gujarati (En-Gu), English-Hindi (En-Hi), English-Tamil (En-Ta), and English-Telugu (En-Te). Experiments under zero-shot, few-shot prompting, and fine-tuning settings revealed significantly low performance in the zero-shot, while fine-tuning achieved accuracy comparable to last year's best scores. Our system demonstrated the effectiveness of this approach in low-resource language QE, securing 1st place in both En-Gu and En-Hi, and 4th place in En-Ta and En-Te. The code used in our experiments is available at the following URL [1].

1 Introduction

Machine translation quality estimation evaluates translation automatically without reference translation. This practice reduces the cost of manual translation and enables efficient evaluation. The subsequent quality score flags the necessity of resorting to a more reliable translation system or revision from human post-editing. Quality estimation can be performed at various granularity levels, including word, phrase, sentence, and document.

In this paper, we describe our contribution to the QE shared task at the Ninth Machine Translation Conference (WMT24). We participate in the Task 1 of the shared task and we specifically focus on the sentence-level subtask, which involves predicting the quality score of neural MT outputs at the sentence level without access to reference translations. There are two different annotation methods for QE: Multidimensional Quality Metric (MQM) (Freitag et al., 2021) and Direct Assessment (DA) (Fomicheva et al., 2022), and we report the results of DA score prediction. Our study targets four language pairs: English-Gujarati (En-Gu), English-Hindi (En-Hi), English-Tamil (En-Ta), and English-Telugu (En-Te). The participating systems are assigned the task of predicting the quality score of each source and target sentence pair, and their performance is evaluated using Spearman's rank correlation coefficient as the primary metric, and Pearson and Kendall coefficients as supplementary metrics.

	En-Gu	En-Hi	En-Ta	En-Te
baseline	0.661	0.678	0.592	0.414
gpt4o-mean	**0.712**	**0.735**	**0.616**	0.457
gpt4o-prob	**0.712**	0.734	0.608	**0.460**

Table 1: Spearman's rank correlation coefficient between our predictions and human DA judgments of WMT24 test data. The best score obtained for each language pair is marked in bold.

We present a system for quality estimation utilizing a large language model (LLM), inspired by the success of LLMs in regression tasks (Liu et al., 2023; Enomoto et al., 2024). Specifically, we manually designed a prompt for quality estimation and employed GPT-4o mini (OpenAI, 2024) to generate assessment scores multiple times based on this prompt. We then used either the averaged score of these generated scores or their weighted sum based on the generation probability as the final score. Evaluation experiments were conducted in both zero-shot and three-shot settings. Noticeably, we fine-tuned GPT-4o mini using the training data released at WMT23 (Kocmi et al., 2023) and as-

[1] https://colab.research.google.com/drive/1p8VMnAkRfuVpbvM_revV2ZaN76sSxmiE?usp=sharing

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 529–534
November 15–16, 2024. ©2024 Association for Computational Linguistics

sessed its performance.

We first evaluated our systems using the development data released at WMT23. The results indicated that the Spearman's correlations in the zero- and few-shot settings ranged from 0.2 to 0.4, while those for the fine-tuned GPT-4o mini ranged from 0.4 to 0.7. Compared to a single generation, the estimated score derived from the average or weighted sum based on multiple output values was found to perform consistently better. Subsequently, we evaluated the system based on the fine-tuned GPT-4o mini using the test data from WMT24. Table 1 presents the results of our system and the baseline for Task 1 (Rei et al., 2022). The system achieved Spearman's correlation scores of 0.712, 0.735, 0.616, and 0.460 in the En-Gu, En-Hi, En-Ta, and En-Te language pairs, respectively, surpassing the baseline system's performance. We achieved the 1st place in En-Gu and En-Hi, and 4th place in En-Ta and En-Te.

2 Related Work

GEMBA (Kocmi and Federmann, 2023) is a translation quality metric that utilizes a large language model (LLM). It has been shown to have a high correlation with the human-rated MQM score of the WMT22 Metrics shared task. Their experiments covered three language pairs (English to German, English to Russian, and Chinese to English) of the WMT22 Metrics shared task using seven GPT variants from GPT-2 to GPT-4 models. Lu et al. (2024) investigated various prompts to improve segment-level evaluation performance. They experimented with Llama2-70B model (Touvron et al., 2023) and Mixtral-8x7b model (Jiang et al., 2024) in addition to GPT-3.5-Turbo model, and showed that the method using GPT-3.5-Turbo had the best performance. These previous studies highlight the potential of LLM evaluators as human alternatives. Our system uses the latest model, GPT-4o mini, and also estimates quality for more challenging translations for low-resource languages.

Enomoto et al. (2024) used an LLM to solve the lexical complexity prediction task. They reported a bias in the numerical values generated by an LLM in that certain values occur frequently regardless of the input. To mitigate the bias and achieve a more precise numerical output, we run the generation several times and obtain the final scores by either average or expectation weighted by generation probabilities.

3 Methodology

Our system uses LLMs to estimate translation quality scores. Following GEMBA (Kocmi and Federmann, 2023), to assess translation quality via prompting an LLM, the following arguments are required:

- source language name: {{source language}}
- target language name: {{target language}}
- source sentences: {{$src_1, ..., src_N$}}
- translated sentences: {{$hyp_1, ..., hyp_N$}}
- few-shot examples: {{examples}} (optional)

We define the instructions to be input into the LLM as follows:

> *Please analyze the given source and translated sentences and output a translation quality score on a continuous scale ranging from 0 to 100.*
> *Translation quality should be evaluated based on both fluency and adequacy.*
> *A score close to 0 indicates a low quality translation, while a score close to 100 indicates a high quality translation.*
> *Do not provide any explanations or text apart from the score.*
>
> {{examples}}
> {{source language}} *Sentence:* {{src_i}}
> {{target language}} *Sentence:* {{hyp_i}}
> *Score:*

The instruction template is designed to include a description of the task, the score range, and a description of the evaluation criteria. To restrict the output to numerical values only, it is important to state "Do not provide any explanations or text apart from the score." explicitly.

According to Kocmi and Federmann (2023), there are some numbers that are particularly prone to output, such as "95". To mitigate such bias in the output distribution, the final score is computed from the sampled generated results with reference to the G-Eval framework (Liu et al., 2023). $Score_{mean}$ is the simple average of the generated scores, while $Score_{prob}$ is the score weighted by the generation probabilities. Let $S = \{s_1, s_2, ..., s_n\}$ represent the set of scores generated by the prompt, and let $p(s_i)$ be the softmax output probability of each generated score.

		En-Gu			En-Hi			En-Ta			En-Te		
Method	Setting	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ
						Single generation							
	Zero-shot	0.205	0.501	0.337	0.379	0.434	0.291	0.277	0.514	0.514	0.271	0.280	0.208
	Three-shot	0.413	0.491	0.309	0.390	0.400	0.282	0.463	0.420	0.344	0.294	0.305	0.226
	Fine-tuned	0.599	0.659	0.453	0.510	0.639	0.367	0.618	0.704	0.453	0.283	0.263	0.205
						Multi generation							
$score_{mean}$	Zero-shot	0.453	0.505	0.328	0.389	0.449	0.277	0.514	0.529	0.373	0.274	0.275	0.193
	Three-shot	0.447	0.512	0.319	0.422	0.426	0.294	0.498	0.428	0.358	0.290	0.303	0.205
	Fine-tuned	0.680	**0.717**	0.506	0.564	0.686	0.409	0.661	**0.747**	0.487	0.392	**0.361**	0.270
$score_{prob}$	Zero-shot	0.451	0.499	0.323	0.394	0.447	0.275	0.519	0.521	0.368	0.274	0.276	0.190
	Three-shot	0.448	0.514	0.319	0.423	0.427	0.295	0.500	0.426	0.358	0.290	0.303	0.202
	Fine-tuned	**0.683**	0.715	**0.508**	**0.568**	**0.690**	**0.412**	**0.663**	0.746	**0.489**	**0.399**	0.360	**0.277**

Table 2: Spearman (ρ), Pearson (r) and Kendall (τ) correlation between the proposed approaches and human DA judgments of WMT23 dev data. The best Spearman score obtained for each language pair is marked in bold. Single generation is a prediction method that uses the output value generated only once as the estimated score, and the other two are methods that calculate the average value or the expected value based on the generation probability, based on the output values by 20 times generation.

The final scores are calculated using the following formulae:

$$score_{mean} = \frac{1}{n}\sum_{i=1}^{n} s_i \quad (1)$$

$$score_{prob} = \sum_{i=1}^{n} p(s_i) \times s_i \quad (2)$$

In this study, the experiment is conducted with $n = 20$. Among the generated outputs, non-numeric tokens and numbers outside the specified range are excluded from S.

4 Experiments

We conduct experiments to investigate two RQs. **RQ 1**: Which methods are more effective in improving performance of low-resource language QE? **RQ 2**: How effective are sampling methods in mitigating numerical output bias?

4.1 Settings

Model We use GPT-4o mini ("gpt-4o-mini-2024-07-18") (OpenAI, 2024) for our experiments. It is priced at 15 cents per million input tokens and 60 cents per million output tokens and more than 60% cheaper than GPT-3.5 Turbo.

Our system fine-tunes GPT-4o mini. The fine-tuning process is conducted using OpenAI's API.

Data For the remaining sections, we only report results on the WMT23 dev dataset. The examples used for few-shot prompting are randomly obtained from the WMT23 training dataset. WMT23 training data consists of 7,000 sentence pairs in each language and is also used for fine-tuning.

4.2 Results

Spearman, Pearson, and Kendall correlation coefficients between predicted and gold scores for each language pair are shown in Table 2.

4.2.1 Strategies for Low-Resource Languages

For **RQ 1**, we compare the performance of the three settings: zero-shot, few-shot prompting, and fine-tuning. Few-shot in $score_{mean}$ and $score_{prob}$ improved Spearman correlation coefficients slightly by 0.033 for En-Hi and 0.016 for En-Te, while En-Gu and En-Ta scores decreased by 0.006 and 0.016 respectively. In other words, the few-shot strategy is not very effective for low-resource languages. On the other hand, in the single genaration setting, En-Gu and En-Ta improved by 0.208 and 0.186, respectively, indicating that the few-shot is more effective when the generation times are limited.

Fine-tuning improves performance in almost all evaluation metrics and is an effective measure for low-resource languages. Our systems submitted to the shared task (Table 1) are also the result of the fine-tuned models.

4.2.2 Strategies for Distributional Bias

For **RQ 2**, we compare the performance of the three settings: single generation, $score_{mean}$, and $score_{prob}$. In the fine-tuned model, the difference in Spearman's rank correlation coefficient with sin-

	En-Gu	En-Hi	En-Ta	En-Te
Manually	**0.451**	**0.394**	**0.519**	**0.274**
AutoCoT	0.444	0.387	0.514	0.238

Table 3: Spearman's rank correlation coefficient between our predictions in a zero-shot setting using two different prompt generation methods and human DA judgments of WMT23 dev data.

gle generation (average of four languages) is 0.072 for $score_{mean}$ and 0.076 for $score_{prob}$. Compared to single genaration, the other two sampling methods performed better, demonstrating the effectiveness of these methods in mitigating the effects of bias during generation. The performance of $score_{mean}$ and $score_{prob}$ is almost equal, and either method can be used.

5 Discussion

5.1 Is AutoCoT necessary for G-Eval?

In Chiang and Lee (2023), they find that the auto Chain-of-Thought (CoT) used in G-Eval does not always make G-Eval more aligned with human ratings. In this section, we examine the methods used to construct the prompt for this task.

To replicate the G-Eval framework (Liu et al., 2023) procedures, it is necessary to construct an initial prompt to generate the evaluation steps using AutoCoT. Specifically, we first manually designed a prompt that contains the definition of the QE task and the desired evaluation criteria as follows:

> *You will be given a source and a translated sentence. Your task is to rate translated sentence on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.*
>
> *Evaluation Criteria:*
>
> *Translation Quality (0 - 100) - the quality of a translation based on the adequacy and fluency of the sentence.*

Then, we added a line of "*Evaluation Steps:*" to the prompt and let GPT-4 [2] generate the following evaluation steps by CoT automatically:

> *Evaluation steps:*
>
> *1. Read the source sentence and the translated sentence carefully.*
>
> *2. Evaluate the translated sentence based on its adequacy and fluency.*
>
> *- Adequacy: How much of the meaning expressed in the source text is also expressed in the target text? A score of 100 means all the meaning is transferred, and 0 means none of it is.*
>
> *- Fluency: Does the translation sound like something a native speaker would say? A score of 100 means it sounds completely native, and 0 means it doesn't sound native at all.*
>
> *3. Give the translation a score between 0 and 100, where 0 is the worst and 100 is the best.*

We compare the performance in a zero-shot setting using prompts created by AutoCoT and those created manually. As shown in Table 3, manual prompts performed better than AutoCoT for all languages. This result follows the findings of Chiang and Lee (2023), and we decided to use manually constructed prompts in our systems to get results that correlated better with human judgment.

5.2 Is it difficult for GPT-4 evaluators to evaluate Telugu text?

In our results, the performance on Telugu data was lower than other languages. This may be attributed to the linguistic complexity of Telugu, which features complex noun and verb conjugations, as well as its status as a low-resource language. Kishore and Shaik (2024) demonstrated that ChatGPT is less accurate in Telugu grammar and vocabulary compared to Gemini. Performance is expected to improve by using LLMs specialized for each language (e.g., Telugu GPT[3]) rather than relying on a single, generalized model.

6 Conclusion

Our study demonstrates the efficacy of using a LLM for sentence-level quality estimation in machine translation. By leveraging GPT-4o mini, we achieved improvements over baseline systems in predicting quality scores for various language pairs. The fine-tuned GPT-4o mini model exhibited robust performance in low-resource language QE,

[2] We used `gpt-4-0613` following Liu et al. (2023)

[3] `https://chatgpt.com/g/g-RjoqGo7g0-telugu-gpt`

with Spearman's correlation scores significantly higher than those in the zero- and few-shot settings. These findings emphasize that fine-tuning with an annotated QE dataset is crucial for enhancing performance in low-resource languages. However, in practical scenarios, creating and obtaining such datasets for low-resource languages poses significant challenges. Therefore, efforts to effectively improve performance using a small amount of data, as explored in works like (Lauscher et al., 2020; Kim and Komachi, 2023), are important directions for future research.

Acknowledgements

This work was partly supported by JST, PRESTO Grant Number JPMJPR2366, Japan.

References

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598.

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.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Hwichan Kim and Mamoru Komachi. 2023. Enhancing few-shot cross-lingual transfer with target language peculiar examples. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 747–767, Toronto, Canada. Association for Computational Linguistics.

Katikela Sreeharsha Kishore and Rahimanuddin Shaik. 2024. Evaluating telugu proficiency in large language models: A comparative analysis of ChatGPT and Gemini. *Preprint*, arXiv:2404.19369.

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, Singapore. 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.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8801–8816, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine*

Translation (WMT), pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.