Toward an LLM-as-a-judge for Patent Claim Translation based on Human Evaluation Criteria of WAT2025

Yingyi Fu[1] Haruto Azami[1] Takehito Utsuro[1] Masaaki Nagata[2]
[1]Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba
[2]NTT Communication Science Laboratories

Abstract

This study develops an LLM-as-a-judge method for patent claim translation to improve the evaluation of machine translation (MT) systems, filter the training data, and optimize models using reinforcement learning. While previous studies have proposed LLM-as-a-judge methods for MT evaluation, they are not sufficient for patent claim translation. Therefore, we adapt GEMBA-MQM for patent claim translation in multiple ways, from simply informing the LLM that the task is patent translation to incorporating the Human Evaluation Criteria of the 12th Workshop on Asian Translation (WAT2025). Surprisingly, on JA-EN translation direction, the simplest adaptation achieves the best performance, while most of the adaptations outperform the baseline and effectively evaluate translation quality.

1 Introduction

Recently, large language models (LLMs) have shown high performance in patent claim translation [1]. However, translation accuracy decreases significantly when dealing with challenging patent claims, such as those with long sentences. One of the most important problems for patent claim translation is that the correlation between existing automatic metrics, such as CometKiwi [2] and MetricX [3], and human evaluation is low [4]. Thus, it is important to develop a new evaluation metric for patent translation. A study by Kocmi and Federmann showed that GEMBA-MQM has strong potential for detecting translation errors and scoring machine translation (MT) systems [5]. However, the original GEMBA-MQM does not perform well on patent claim translation tasks.

To develop an evaluation method for patent claim translation, we propose several GEMBA-MQM-based prompting methods, ranging from simply informing the LLM that of the task to incorporating the Human Evaluation Criteria of the 12th Workshop on Asian Translation (WAT2025).

The organizers of WAT2025 provide a development set for the patent claim translation task. It consists of 19 Japanese-English (JA-EN) and 11 English-Japanese (EN-JA) patent claims with two manually-made (post-edited) reference translations, except for one claim in EN-JA translation direction, which has only one reference translation. We used the average success rate, defined as the percentage of cases where the reference translation is scored higher than the machine translation output, as a meta-evaluation metric for a set of prompts for translation evaluation. For the calculation of the average success rate, we filtered out input JA-EN patent claims whose machine translation outputs with high lexical diversity. Surprisingly, the simplest adaptation achieves the best performance on JA-EN translation direction, while most of the adaptations outperform the baseline and effectively evaluate translation quality. Moreover, most of the adaptations outperform the baseline and effectively evaluate the patent claim translation quality.

2 GEMBA-MQM

Kocmi and Federmann suggested that LLMs can be prompted to evaluate the quality of machine translation [6]. Based on their previous research, they proposed an LLM-as-a-judge approach using the GPT-4 model [7] to detect errors in MT outputs and to score them, which is known as GEMBA-MQM [5].

The error classes provided in the GEMBA-MQM prompt are based on the Multidimensional Quality Metrics (MQM) framework [8]. These error classes include several categories of translation errors, such as accuracy, fluency, and terminology. Some error categories include subcategories. For example, the accuracy error includes addi-

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Table 1 Human Evaluation Criteria of WAT2025 and GEMBA-MQM Original Prompt

Human Evaluation Criteria of WAT2025		GEMBA-MQM Original Prompt	
Top Category	Mid Category (Sub Category)	Top Category	Mid Category
accuracy	addition, omission, untranslated text, **mistranslation (numerals / symbols, article, incorrect dependency, unknown dependency, ambiguity)**	accuracy	addition, **mistranslation**, omission, untranslated text
fluency	punctuation, spelling, grammar, register, inconsistency, character encoding	fluency	fluency character encoding, grammar, inconsistency, punctuation, register, spelling
terminology	inappropriate for context, inconsistent use	terminology	inappropriate for context, inconsistent use
style	awkward	style	awkward
locale convention	address format, currency format, date format, name format, telephone format, time format	locale convention	currency, date, name, telephone, or time format
other	—	other	—
non-translation	—	non-translation	—
source error	—		

The gray part (locale convention) is omitted in the implementations.
Boldface indicates the main difference between the Human Evaluation Criteria of WAT2025 and the GEMBA-MQM original prompt.

tion, mistranslation, omission, and untranslated text. More details on the error categories are provided in Table 1. Each error is assigned a severity level and a corresponding weight, with the default values set to 25 for critical, 5 for major, and 1 for minor errors. The GEMBA-MQM score is calculated based on the error counts and their weights, and the result is represented as a negative number. To intuitively compare scores across different systems, we calculate the final score in our experiment as follows:

$$\text{final\_score} = 100 + \text{GEMBA-MQM\_score} \quad (1)$$

3 Human Evaluation Criteria of WAT2025

To adapt GEMBA-MQM for patent claim translation, we attempt to improve the GEMBA-MQM prompt based on the Human Evaluation Criteria of WAT2025 [4]. Table 1 shows the error types of the Human Evaluation Criteria of WAT2025 and the prompt used in the original GEMBA-MQM. The main difference between the two is that the Human Evaluation Criteria of WAT2025 provide sub-categories for the mistranslation error.

Table 2 COMET Scores

Translations	COMET	
	ja-en	en-ja
Qwen3-8B-cpt-sft+grpo(Merge)	**85.71**	**89.02**
Qwen3-8B-grpo-merge-ntt	83.86	85.26
Qwen3-8B-cpt	83.21	68.69
Qwen3-8B	79.97	75.60
plamo-2-translate	82.90	88.00

COMET Scores of the entire development set.
Boldface indicates the highest value in each column.

4 Dataset for Patent Claim Translation

The JA-EN development data in this study, are taken from the development set of WAT2025 Patent Claims Translation / Evaluation Tasks[1)]. The WAT2025 development set consists of 19 Japanese claims and 11 English claims. Each claim has two reference translations, except for one claim in EN-JA translation direction, which has only one reference translation. We utilize several MT systems developed internally for patent translation[1] to generate translations. We also used an open-weight LLM specialized in machine translation between Japanese and

1) https://sites.google.com/view/pat-claims-trans-2025

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Table 3 Success Rate on Filtered JA-EN Development Set (%)

Translations	Success Rate (%)						
	Prompt 0 (GPT-4)	Prompt 1 (GPT-4)	Prompt 1 (GPT-5.2)	Prompt 2 (GPT-5.2)	Prompt 3 (GPT-5.2)	Prompt 4 (GPT-5.2)	Prompt 5 (GPT-5.2)
Qwen3-8B-cpt-sft+grpo(Merge)	27.78	22.22	55.56	38.89	44.44	44.44	50.00
Qwen3-8B+grpo(Merge)	27.78	11.11	77.78	55.56	61.11	55.56	55.56
Qwen3-8B-cpt	61.11	72.22	88.89	100.00	94.44	100.00	94.44
Qwen3-8B	50.00	33.33	77.78	83.33	88.89	77.78	77.78
plamo-2-translate	33.33	16.67	50.00	55.56	33.33	27.78	44.44

Success Rate means the percentage of segments in which the reference translation is scored higher than the machine translation output.

Table 4 Average Success Rate on Filtered JA-EN Development Set (%)

Prompt No.	Prompt Name	GPT Model	Average Success Rate (%)
0	GEMBA-MQM-Original	GPT-4	40.00
1	GEMBA-MQM-O-Patent	GPT-4	31.11
1	GEMBA-MQM-O-Patent	GPT-5.2	**70.00**
2	GEMBA-MQM-O-P-Term	GPT-5.2	66.67
3	GEMBA-MQM-O-P-Term-Modified	GPT-5.2	64.44
4	GEMBA-MQM-O-P-T-M-WAT25	GPT-5.2	61.11
5	GEMBA-MQM-O-P-T-M-WAT25-Modified	GPT-5.2	64.44

The average success rate on filtered JA-EN development set across all MT systems for each prompting method
Boldface indicates the highest value in each column.
Prompt 0: GEMBA-MQM-Original; Prompt 1: GEMBA-MQM-O-Patent;
Prompt 2: GEMBA-MQM-O-P-Term; Prompt 3: GEMBA-MQM-O-P-Term-Modified;
Prompt 4: GEMBA-MQM-O-P-T-M-WAT25; Prompt 5: GEMBA-MQM-O-P-T-M-WAT25-Modified

English.

The description of each MT system is as follows:

- **Qwen3-8B-cpt-sft+grpo (Merge):** the Qwen3-8B model[2], trained with CPT, SFT, and Group Relative Policy Optimization (GRPO).
- **Qwen3-8B+grpo (Merge):** the Qwen3-8B model, trained with GRPO.
- **Qwen3-8B-cpt:** the Qwen3-8B model, trained with CPT.
- **Qwen3-8B:** the Qwen3-8B model.
- **plamo-2-translate:** the PLaMo translation model[3].

Patent data are used in the CPT and SFT training stages.

5 Prompts for Evaluating Patent Claim Translation

In this study, we test multiple ways to adapt GEMBA-MQM for patent claim translation, ranging from simply informing the GPT that the task is patent translation to incorporating the Human Evaluation Criteria of the 12th Workshop on Asian Translation (WAT2025). Table 5 in Appendix shows the detailed prompts. First, we simply inform the GPT model that the task is patent translation. However, there is an issue with this prompt: in patent translation, one term may have multiple valid translations. The GPT model may consider the reference translation inadequate and mark it as a "terminology" error. Therefore, we add notes at the end of the prompt to specify how terminology errors should be handled. We test two different ways to describe which terminology errors should be marked: (1) "inappropriate or inconsistent", which adopts the same wording as the GEMBA-MQM-Original prompt shown in Table 5 in Appendix; and (2) simply "incorrect". Then, we modify our prompt specifically for patent claim translation based on the Human Evaluation Criteria of WAT2025 [4]. However, in the reference translations of the development data, "(characterized by)" (the translation of "を特徴とする") is marked as a translation error by the GPT model due

2) https://huggingface.co/Qwen/Qwen3-8B
3) https://huggingface.co/pfnet/plamo-2-translate

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

to the unnecessary parentheses. Hence, we assume that these surrounding parentheses are conventions and modify the prompt to ignore them during evaluation.

6 Evaluation

6.1 Evaluation Methods

We evaluate our prompting method based on Zhang et al.'s study, which utilized the percentage of cases where human translations were scored higher than machine translations as a meta-evaluation metric [9]. We score both the reference translations and the outputs from each MT system. Then, we calculate the percentage of segments where the reference translation is scored higher than the MT output. Moreover, we refer to this percentage as the success rate in Section 6.2.

To compare the prompting methods, we compute pairwise BLEU scores[4] among MT outputs for each source sentence [10]. Then, we filter 18 out of 38 sentences for JA-EN translation whose pairwise BLEU scores are lower than or equal to the average (36.44) to ensure lexical diversity. We calculate the success rate per MT system under each prompt and compute the average success rate on the filtered development set across all MT systems for each prompting method. We compute the average GEMBA-MQM score of the reference translations and MT outputs under different prompts on the filtered development set and test the statistical significance of the GEMBA-MQM score differences between the reference translations and the MT outputs. Furthermore, for reference, we calculate the COMET score of each MT output for each translation directions using Unbabel/wmt22-comet-da[5] [11] on the entire development set.

6.2 Results

Table 2 shows the COMET score of MT outputs under different prompts. Table 3 shows the success rate of each MT system under different prompts. Table 4 shows the average success rate across all MT systems for each prompting method. All the results are based on the filtered development set.

The description of each prompting method is as follows:

- **GEMBA-MQM-Original (Prompt 0):** The prompt used in the original GEMBA-MQM, which is set as the baseline.
- **GEMBA-MQM-O-Patent (Prompt 1):** A modified version of [GEMBA-MQM-Original] for patent translation by simply informing GPT that the task is patent translation.
- **GEMBA-MQM-O-P-Term (Prompt 2):** A modified version of [GEMBA-MQM-O-Patent] by providing notes instructing the LLM to mark terminology errors only when they are inappropriate or inconsistent.
- **GEMBA-MQM-O-P-Term-Modified (Prompt 3):** A further modification of [GEMBA-MQM-O-P-Term] by instructing the LLM to mark terminology errors only when they are clearly incorrect.
- **GEMBA-MQM-O-P-T-M-WAT25 (Prompt 4):** A modified version of [GEMBA-MQM-O-P-Term-Modified] based on the Human Evaluation Criteria of WAT2025.
- **GEMBA-MQM-O-P-T-M-WAT25-Modified (Prompt 5):** A further modification of [GEMBA-MQM-O-P-T-M-WAT25] by treating "(characterized by)" as a convention equivalent to "characterized by".

In summary, using GPT-5.2, GEMBA-MQM-O-Patent (Prompt 1) achieve the highest average success rate for JA-EN translation direction. After excluding the strongest MT system, defined as the one with the lowest success rate, i.e, the smallest percentage of segments where the reference translation is scored higher than the MT output, the same prompt still perform best.

7 Conclusion

This study observes remarkable improvements with GPT-5.2 over GPT-4. According to the average success rate on the filtered JA-EN development set, the simplest adaptation performs best over JA-EN, while most of the adaptations outperform the baseline and effectively evaluate translation quality. At the time of writing this manuscript, the human evaluation results for WAT2026 have not been released. Once this data is released, we intend to use it to develop automated evaluation method that has strong correlation with human evaluations.

4) https://github.com/facebookresearch/fairseq/tree/main/examples/translation_moe
5) https://huggingface.co/Unbabel/wmt22-comet-da

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

References

[1] H. Azami, et al. Patent claim translation via continual pre-training of large language models with parallel data. In **Proc. MTSummit XX: Vol. 1**, pp. 300–314, 2025.

[2] R. Rei and other. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In **Proc. 8th WMT**, pp. 841–848, 2023.

[3] J. Juraska, et al. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In **Proc. 9th WMT**, pp. 492–504, 2024.

[4] T. Nakazawa, et al. Findings of the first patent claims translation task at WAT2025. In **Proc. 12th WAT**, 2025.

[5] T. Kocmi and C. Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In **Proc. 8th WMT**, pp. 768–775, 2023.

[6] T. Kocmi and C. Federmann. Large language models are state-of-the-art evaluators of translation quality. In **Proc. 24th EAMT**, pp. 193–203, 2023.

[7] OpenAI. GPT-4 technical report. https://arxiv.org/abs/2303.08774, 2024.

[8] M. Freitag, et al. Experts, errors, and context: A large-scale study of human evaluation for machine translation. **TACL**, Vol. 9, pp. 1460–1474, 2021.

[9] R. Zhang, et al. LiTransProQA: An LLM-based literary translation evaluation metric with professional question answering. In **Proc. EMNLP**, pp. 29087–29109, 2025.

[10] T. Shen, et al. Mixture models for diverse machine translation: Tricks of the trade. **ICML**, 2019.

[11] R. Rei, et al. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In **Proc. 7th WMT**, pp. 578–585, 2022.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

A Prompts

Table 5 Prompts

GEMBA-MQM-Original (Prompt 0)

```
(system)You are an annotator for the quality of machine translation. Your task is to identify errors and assess the quality of the translation.
(user){source_language} source:\n
```{source_segment}```\n
{target_language} translation:\n
```{target_segment}```\n
Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them.
The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation,
register, spelling), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error.
\nEach error is classified as one of three categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow,
but what the text is trying to say is still understandable.
Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.\n
\n
```

GEMBA-MQM-O-Patent (Prompt 1)

```
(system)You are an annotator for the quality of patent translation. Your task is to identify errors and assess the quality of the
translation
(user){source_language} source:\n
```{source_segment}```\n
{target_language} translation:\n
```{target_segment}```\n
\n
Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them.
The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation,
register, spelling), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error.
\nEach error is classified as one of three categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow,
but what the text is trying to say is still understandable.
Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.\n
```

GEMBA-MQM-O-P-Term (Prompt 2)

```
Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them.
The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation,
register, spelling), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error.
\nEach error is classified as one of three categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow,
but what the text is trying to say is still understandable.
Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.\n
Note: There are often multiple valid translations for terminology in patent translation. Only mark terminology errors when the translation is clearly inappropriate
or inconsistent.
```

GEMBA-MQM-O-P-Term-Modified (Prompt 3)

```
Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them.
The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation,
register, spelling), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error.
\nEach error is classified as one of three categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow,
but what the text is trying to say is still understandable.
Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.\n
Note: There are often multiple valid translations for terminology in patent translation. Only mark terminology errors when the translation is clearly incorrect.
```

GEMBA-MQM-O-P-T-M-WAT25 (Prompt 4)

```
Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them.
Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them.
The categories of errors are:
- accuracy: addition, omission, untranslated text, or mistranslation (which includes sub-categories of: numerals/symbols, article, incorrect dependency, unknown
dependency, ambiguity)
- fluency: character encoding, grammar, inconsistency, punctuation, register, or spelling
- style: awkward
- terminology: inappropriate for context, or inconsistent use
- non-translation
- other
- no-error
Note: There are often multiple valid translations for terminology in patent translation. Only mark terminology errors when the translation is clearly incorrect.
```

GEMBA-MQM-O-P-T-M-WAT25-Modified (Prompt 5)

```
Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them.
Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them.
The categories of errors are:
- accuracy: addition, omission, untranslated text, or mistranslation (which includes sub-categories of: numerals/symbols, article, incorrect dependency, unknown
dependency, ambiguity)
- fluency: character encoding, grammar, inconsistency, punctuation, register, or spelling
- style: awkward
- terminology: inappropriate for context, or inconsistent use
- non-translation
- other
- no-error
Note: There are often multiple valid translations for terminology in patent translation. Only mark terminology errors when the translation is clearly incorrect.
Treat "(characterized by)" as a convention equivalent to "characterized by" but not translation errors.
```

Note: Yellow highlighted portions are shared across all prompts for patent claim translation.
Blue text indicates modified or added content.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).