Towards a Rubric-based LLM-as-a-Judge for Patent Claim Translation Evaluation

Yingyi Fu and Takehito Utsuro
Deg. Prog. Sys.&Inf. Eng.,
Grad. Sch. Sci.&Tech.,
University of Tsukuba

Masaaki Nagata
NTT Communication Science Laboratories,
NTT Corporation, Japan

Abstract

Existing automatic machine translation evaluation metrics, such as COMETKiwi and MetricX, showed relatively low correlations with human evaluation in the WAT2025 Patent Claim Translation/Evaluation Tasks. We propose a rubric-based LLM-as-a-Judge method for patent claim translation evaluation. Based on a patent-specific rubric consisting of five evaluation dimensions with corresponding checklist items, we instruct a GPT model to assess translation quality. On a publicly available human evaluation dataset derived from the WAT2025 Patent Claims Translation/Evaluation Tasks test set, the proposed method achieves the highest Pearson and Spearman correlation coefficients with human evaluation among all evaluated methods.

1 Introduction

Patent claim translation is more difficult than general translation because of complex sentence structures, long sentences, and highly specialized technical terminology. Evaluating patent claim translations is also challenging because seemingly minor errors, such as the omission of a limitation or an incorrect element relationship, may substantially change the meaning or scope of the claimed invention.

Although research on patent translation has a long history, research specifically focusing on patent claim translation has been relatively limited. In 2007, a bilingual Japanese–English patent parallel corpus was created by Utiyama and Isahara (2007) for the NTCIR-6 patent retrieval task (Fujii et al., 2007). More recently, Nagata et al. (2024) built JaParaPat, a Japanese–English parallel patent corpus containing over 300 million sentence pairs, aligning every part of the patent documents including titles, abstracts, descriptions, and claims. Several previous studies on patent translation have focused on constructing patent corpora rather than developing patent-translation-specific methods, e.g., COPPA V2.0 (Junczys-Dowmunt et al., 2016), ParaPat (Soares et al., 2020) and JaParaPat. Moreover, relatively little research has focused specifically on patent claim translation evaluation. In 2025, Nakazawa et al. (2025) organized the first shared task focusing on patent claim translation at WAT2025. However, the organizers of WAT2025 reported that automatic metrics, such as COMETKiwi (Rei et al., 2023) and MetricX (Juraska et al., 2024), showed low correlations with human evaluation. Correlations between general machine translation (MT) LLM-as-a-Judge such as GEMBA (Kocmi and Federmann, 2023b,a) and human evaluation were not reported.

To develop an automatic patent claim translation evaluation method that correlates better with human evaluation, we propose an LLM-as-a-Judge method to assess translation quality based on a patent-specific rubric and checklist. The experimental results show that the proposed method achieves higher correlations with WAT2025 human evaluation than COMETKiwi, MetricX, GEMBA-DA and GEMBA-MQM.

In summary, (1) we develop a rubric-based LLM-as-a-Judge method specific to patent claim translation evaluation, which achieves the highest correlations with human evaluation among the evaluated methods. (2) We evaluate several automatic metrics on WAT2025 human evaluation data, including LLM-as-a-Judge methods and neural metrics such as COMETKiwi and MetricX. (3) We compare the performance of GPT-5.2 and GPT-4 using the same LLM-as-a-Judge methods.

2 Related Work

2.1 Automatic Evaluation of MT

BLEU (Papineni et al., 2002) is a traditional automatic MT evaluation metric that measures MT quality based on modified n-gram precision and a

Proceedings of the 13th Workshop on Asian Translation (WAT 2026), pages 84–103
November 9, 2026 ©2026 Association for Computational Linguistics

brevity penalty. Rei et al. (2020) proposed COMET, a neural MT evaluation metric based on multilingual pretrained language models. COMET uses information from the source text and the reference translation to evaluate MT quality and achieves a high correlation with human judgments. Rei et al. (2022) then presented COMETKiwi, a reference-free MT evaluation metric based on the COMET framework. Freitag et al. (2022) reported that neural evaluation metrics such as COMET outperform overlap-based metrics such as BLEU.

2.2 GEMBA-DA and GEMBA-MQM

To investigate the use of LLMs for MT quality assessment, Kocmi and Federmann (2023b) developed GEMBA. One of its prompt variants, GEMBA-DA, uses a GPT model to directly assess translation quality by assigning a score.

Unlike GEMBA-DA, GEMBA-MQM (Kocmi and Federmann, 2023a) uses a GPT model to identify and classify translation errors into error categories and severity levels based on the Multidimensional Quality Metrics (MQM) framework used by Freitag et al. (2021). Specifically, GEMBA-MQM classifies translation errors into categories such as accuracy, fluency, and terminology. It also marks each translation error as critical, major, or minor, with default weights of 25, 5, and 1, respectively. Table 1 shows the error severity levels and categories. The prompt for GEMBA-MQM includes three few-shot examples.

Both methods take the source text and the MT output as input. GEMBA-DA can optionally use a reference translation, whereas GEMBA-MQM is reference-free. Both methods use predefined prompts. The output of GEMBA-DA is a direct assessment (DA) score, whereas the outputs of GEMBA-MQM are error annotations and a negative score calculated based on the number of translation errors and their severity levels.

2.3 LiTransProQA

Zhang et al. (2025) proposed LiTransProQA, an LLM-based question-answering metric for literary translation evaluation. It provides a prompting template and a list of evaluation questions. The prompt instructs the LLM to answer the questions with "yes," "no," or "maybe," which are mapped to scores of 1, 0, and 0.5, respectively. LiTransProQA achieves adequacy performance comparable to that of trained linguistic students, showing its potential for literary translation evaluation. Inspired by LiTransProQA, we develop a patent-claim-specific checklist with three answer options and a score-mapping rule with the assistance of coding agents.

Error Severity	Error Category
Critical	• Accuracy • Fluency • Terminology • Style • Locale convention (omitted from the general GEMBA-MQM prompt) • Other • Non-translation
Major	Same as above
Minor	Same as above

Table 1: GEMBA-MQM Error Severity Levels and Categories.

2.4 Patent Claims Translation/Evaluation Task at WAT2025

Nakazawa et al. (2025) organized the WAT2025 Patent Claims Translation/Evaluation Tasks, the first shared task concentrating on patent claim translation.

The test set consisted of claims from applications without a corresponding application in the target language when the source texts were distributed to the participants and translations from six MT systems. Three of them were submissions from participants: an open-weight LLM trained using continual pretraining and supervised fine-tuning on JaParaPat (Azami et al., 2025), and two GPT-5-based systems using prompt engineering (Edamatsu et al., 2025). The other three anonymous MT systems were a general MT system, a patent-specific MT system, and an open-weight LLM.

A subset of the test data was manually evaluated using human evaluation criteria modified from the MQM-based framework of Freitag et al. (2021) for patent translation. However, the organizers also reported issues with the quality of the MQM-based human annotations. Therefore, in this study, we use the human evaluation scores and evaluate their correlations with automatic evaluation scores.

3 Proposed Method

3.1 Prompt for Proposed Method

The proposed method, Patent-Rubric-Checklist, assesses translation quality using a rubric specific to patent claim translation evaluation.

Notably, the evaluation dimensions and the checklist items are mainly developed with the assistance

of Copilot. We instructed Copilot to design a checklist for patent claim translation evaluation with reference to LiTransProQA by providing general translation criteria such as semantic equivalence and fluency, as well as patent-specific constraints including the number of claim elements and their logical relationships, the structure of the claim as a single sentence, and consistency of terminology. Moreover, we tried to improve the prompt with ChatGPT because we assumed that it is stronger than Copilot.

First, the prompt for the proposed method instructs a GPT model to evaluate each checklist item according to the rubric. As shown in Figure 1, the first part of the prompt describes the task setting for patent claim translation evaluation and specifies the required evaluation procedure. Specifically, the prompt instructs the GPT model to answer each checklist item as follows:

- "yes": no problem
- "partial": minor or weakening problem
- "no": clear or serious problem

Second, the prompt presents the rubric consisting of the five evaluation dimensions and their corresponding checklist items for patent claim translation evaluation, as shown in Figure 2. Lines 23–27 define Dimension [A] Content Fidelity, which evaluates the preservation of the core inventive idea and technical functions or effects, as well as the absence of hallucinated technical content. Lines 29–33 define Dimension [B] Element Coverage, which evaluates the translation of the claim elements, as well as the preservation of element relationships, quantities, ranges, or conditions. Lines 35–42 define Dimension [C] Claim Scope & Logical Constraints, which evaluates the preservation of claiming scope, the correct attachment of constraint clauses, and the preservation of coordination and modification relations. Lines 44–50 define Dimension [D] Claim Form & Syntactic Requirements, which evaluates whether the translation is written as a single sentence claim, preserves the claim type, maintains consistent antecedent basis, and avoids narrative or explanatory prose style. Lines 52–56 define Dimension [E] Fluency & Grammaticality, which evaluates the grammar correctness, stylistic appropriateness, and readability.

Third, the prompt presents the input and the output formats, shown as Figure 3. The prompt instructs the GPT model to return the answers to the checklist items with justifications and an overall explanation in JSON format.

3.2 Scoring Process

We map the checklist answers to 1.0 for "yes," 0.5 for "partial," and 0.0 for "no." For each evaluation dimension, we average the mapped scores of its checklist items to obtain a dimension score, denoted by A through E in Equation 1. We then aggregate the dimension scores using the following weighted formula:

$$\text{rule_based_score} = 100 \times \big(0.25A + 0.25B + 0.25C + 0.15D + 0.10E\big) \quad (1)$$

Specifically, the weights assigned to each dimension in Equation 1 were suggested by Copilot.

4 Experiment Setup

4.1 Dataset

In this study, we use a manually annotated dataset[1] derived from the WAT2025 patent claims translation evaluation test set[2]. The test set contains 26 documents with 70 claims for the Japanese-to-English (JA–EN) direction and 30 documents with 81 claims for the English-to-Japanese (EN–JA) direction. The JA–EN set contains outputs from six MT systems and the EN–JA set contains outputs from four MT systems. No reference translations are provided for either language direction. The human evaluation dataset contains scores and annotations for the MT outputs of 13 documents with 28 claims in each language direction from the test set described above.

4.2 Prompting Methods

In our experiments, we use three prompting methods to evaluate the translation quality. Both GEMBA-DA and GEMBA-MQM are existing LLM-as-a-Judge methods for MT evaluation developed by Kocmi and Federmann (2023b,a), whereas Patent-Rubric-Checklist is developed in this study for patent claim translation evaluation.

In summary:

- **GEMBA-DA:** Directly assigns a score ranging from 0 to 100 based on meaning and grammar (baseline).

[1] https://github.com/Tzawa/pat-claim-trans
[2] WAT2025 Patent Claims Translation/Evaluation Tasks

```
You are an expert evaluator of MACHINE TRANSLATION OF PATENT CLAIMS.

Your task is to evaluate the translation quality using a
rubric-based evaluation framework,
implemented explicitly as a CHECKLIST.

This is NOT general prose.
The input is a PATENT CLAIM, which has strict legal and structural constraints.

You must:
(1) Answer ALL checklist questions below using:
    - "yes" (no problem)
    - "partial" (minor or weakening problem)
    - "no" (clear or serious problem)
(2) For each question, provide a short justification.
(3) Provide a brief explanation of the evaluation result.
(4) Evaluate SOURCE vs HYPOTHESIS only; no reference is provided.

```

Figure 1: Prompt for Patent-Rubric-Checklist, Part 1: Task Setting and Evaluation Procedure.

```
--------------------
CHECKLIST (PATENT-CLAIM-SPECIFIC)
--------------------

[A] CONTENT FIDELITY (Meaning Preservation)

A1. Is the core inventive idea (what is being claimed) preserved?
A2. Are the technical functions or effects preserved without distortion?
A3. Is the translation free from hallucinated technical content not present in the
    source?

[B] ELEMENT COVERAGE (Completeness)

B1. Are all essential claim elements translated (no omissions)?
B2. Are relationships between elements (having, connected to, arranged on, etc.)
    preserved?
B3. Are quantities, ranges, or conditions (e.g. "at least one", numerical ranges)
    preserved?

[C] CLAIM SCOPE & LOGICAL CONSTRAINTS (CRITICAL)

C1. Is the claiming scope preserved?
    (e.g., "comprising" vs "consisting of", "characterized in that")
C2. Are constraint clauses correctly attached?
    (e.g., "wherein", "前記〇〇が〜")
C3. Are coordination vs modification relations preserved?
    (e.g., A and B that each have C, NOT A and (B that has C))

[D] CLAIM FORM & SYNTACTIC REQUIREMENTS

D1. Is the translation written as a SINGLE SENTENCE claim?
D2. Is the claim type preserved (apparatus / method / composition, etc.)?
D3. Is antecedent basis consistent?
    (a/the/said, 前記〜)
D4. Does the translation avoid narrative or explanatory prose style?

[E] FLUENCY & GRAMMATICALITY

E1. Is the translation grammatically correct in the target language?
E2. Is it stylistically appropriate for patent claims?
E3. Is the sentence readable and not excessively awkward or ambiguous?

```

Figure 2: Prompt for Patent-Rubric-Checklist, Part 2: Rubric and Checklist Items.

```
--------------------
INPUT FORMAT
--------------------
{source_lang} source patent claim:
"{source_seg}"

{target_lang} translated patent claim:
"{target_seg}"

--------------------
OUTPUT FORMAT (JSON ONLY)
--------------------

{{
  "checklist_answers": {{
    "A1": {{"answer": "yes", "justification": "Short justification."}},
    "A2": {{"answer": "yes", "justification": "Short justification."}},
    "A3": {{"answer": "yes", "justification": "Short justification."}},
    "B1": {{"answer": "yes", "justification": "Short justification."}},
    "B2": {{"answer": "yes", "justification": "Short justification."}},
    "B3": {{"answer": "yes", "justification": "Short justification."}},
    "C1": {{"answer": "yes", "justification": "Short justification."}},
    "C2": {{"answer": "yes", "justification": "Short justification."}},
    "C3": {{"answer": "yes", "justification": "Short justification."}},
    "D1": {{"answer": "yes", "justification": "Short justification."}},
    "D2": {{"answer": "yes", "justification": "Short justification."}},
    "D3": {{"answer": "yes", "justification": "Short justification."}},
    "D4": {{"answer": "yes", "justification": "Short justification."}},
    "E1": {{"answer": "yes", "justification": "Short justification."}},
    "E2": {{"answer": "yes", "justification": "Short justification."}},
    "E3": {{"answer": "yes", "justification": "Short justification."}}
  }},
  "overall_explanation": "Brief explanation of the evaluation result in 1-2 sentences."
}}
```

Figure 3: Prompt for Patent-Rubric-Checklist, Part 3: Input and Output Formats.

- **GEMBA-MQM:** Detects translation errors, labels their category and severity, and calculates a score based on error severities (baseline).

- **Patent-Rubric-Checklist:** Instructs a GPT model to evaluate translation quality based on a rubric with five evaluation dimensions and corresponding checklist items. Then maps the checklist answers to numerical values and calculates the final score outside the LLM using predefined weights (proposed method).

The figures in Appendix A provide the detailed prompts for GEMBA-DA and GEMBA-MQM.

To facilitate comparison with the human evaluation score, GEMBA-DA score, and our proposed method's score, we convert the negative raw GEMBA-MQM score into a final score as follows:

$$\text{final_score} = 100 + 4 \times \text{GEMBA-MQM_score} \quad (2)$$

where final_score denotes the final score, and GEMBA-MQM_score denotes the negative raw score calculated from the GEMBA-MQM annotations.

Furthermore, we use GPT-5.2 (OpenAI, 2025) and GPT-4 (OpenAI et al., 2024) as judge models for different prompting methods.

4.3 Evaluation Metrics

We use Pearson's correlation coefficient[3] and Spearman's rank correlation coefficient[4] as evaluation metrics in our experiments. Pearson correlation measures the linear relationship between automatic evaluation scores and human evaluation scores, while Spearman correlation measures the monotonic relationship between them.

Furthermore, we conduct permutation tests[5] for both Pearson and Spearman correlations to compare our proposed method with GEMBA-DA and GEMBA-MQM. All three methods use GPT-5.2.

[3] SciPy documentation for `pearsonr`
[4] SciPy documentation for `spearmanr`
[5] SciPy documentation for `permutation_test`

5 Results and Analysis

5.1 Correlations with Human Evaluation

We compute Pearson and Spearman correlations between human evaluation scores and the scores produced by GEMBA-DA, GEMBA-MQM, and our proposed method, Patent-Rubric-Checklist, using both GPT-4 and GPT-5.2. We also conduct the same correlation analysis for reference-free MetricX-24 and COMETKiwi. Specifically, we use `Unbabel/wmt23-cometkiwi-da-xl`[6] and `google/metricx-24-hybrid-xl-v2p6`[7] in our experiments. For MetricX-24, we negate the scores when computing correlations so that higher scores indicates better quality. As shown in Table 2 , for the JA–EN direction, all correlations between automatic evaluation scores and human evaluation scores are positive and statistically significant ($p < 0.05$ or $p < 0.01$). However, for the EN–JA direction, the Spearman correlation of COMETKiwi and both the Pearson and Spearman correlation for reference-free MetricX-24 and all the LLM-as-a-Judge methods using GPT-4, are not statistically significant. In particular, although GEMBA-MQM using GPT-4 shows negative Pearson and Spearman correlations, neither is statistically significant. Our proposed method, Patent-Rubric-Checklist using GPT-5.2, achieves the highest Pearson and Spearman correlations in both translation directions. Additionally, we observe that results for the JA–EN direction yield higher correlations than those for EN–JA direction. We also observe that GPT-5.2 yields higher correlations than GPT-4 for all prompting methods.

We conduct permutation tests between the proposed method and GEMBA-MQM using GPT-5.2, the proposed method and GEMBA-DA using GPT-5.2, as well as reference-free MetricX-24 and COMETKiwi. For the JA–EN direction, the difference in Spearman correlation between Patent-Rubric-Checklist and GEMBA-DA, both using GPT-5.2, is statistically significant ($p < 0.01$), whereas the difference in Pearson correlation is not. Similarly, the difference in Spearman correlation between Patent-Rubric-Checklist and GEMBA-MQM, both using GPT-5.2, is statistically significant ($p < 0.05$), whereas the difference in Pearson correlation coefficient is not. The differences in Pearson and Spearman correlations between reference-free MetricX-24 and COMETKiwi are not statistically significant.

For the EN–JA direction, the differences in Pearson and Spearman correlations are not statistically significant between Patent-Rubric-Checklist and either GEMBA-DA or GEMBA-MQM, all using GPT-5.2. The differences in both correlations between reference-free MetricX-24 and COMETKiwi are not statistically significant either.

5.2 Dimension Ablation Study

To further analyze the performance of Patent-Rubric-Checklist using GPT-5.2, we conduct a dimension ablation study by computing the Pearson and Spearman correlations for prompt variants containing only one dimension. Specifically, to analyze the effect of each dimension, we remove the other four dimensions and their corresponding checklist items from the prompt shown in Figure 2. We also revise the output format shown in Figure 3 to match the revised rubric. The prompt variants with a single dimension are provided in Figures 6–11 in Appendix A. Additionally, we set the weight of the single dimension to 1.0. We make these modifications with the assistance of ChatGPT.

Table 3 shows the Pearson and Spearman correlations between human evaluation scores and scores calculated using prompt variants with GPT-5.2. In both translation directions, the prompt variant containing only dimension A achieves the lowest Pearson and Spearman correlations. Surprisingly, although dimensions D and E are assigned relatively lower weights in Equation 1, the prompt variant containing only dimension E achieves the highest Pearson correlation, while the variant containing only dimension D achieves the highest Spearman correlation in both translation directions.

We additionally present case studies illustrating differences between the judgments of dimensions A–C and D–E in Appendix B. The single-dimension scores for dimensions D and E correlate more strongly with human evaluation scores, suggesting that the human evaluation may be more sensitive to the quality aspects captured by dimensions D–E (claim form, syntactic requirements, and fluency & grammaticality) than to those captured by dimensions A–C (content fidelity, element coverage, and claim scope & logical constraints). However, based on the available data and results, we cannot determine why the single-dimension scores for dimensions D and E achieve higher correlations with human evaluation scores.

[6] Unbabel/wmt23-cometkiwi-da-xl
[7] google/metricx-24-hybrid-xl-v2p6

Methods	Pearson's r		Spearman's ρ	
	ja–en	en–ja	ja–en	en–ja
COMETKiwi	0.2880**	0.1860*	0.2703**	0.1810
MetricX-24 (reference-free)	0.2436**	0.1397	0.2945**	0.1770
GEMBA-DA (GPT-4)	0.3753**	0.1391	0.3664**	0.0806
GEMBA-DA (GPT-5.2)	0.4685**	0.1996*	0.4210**	0.3017**
GEMBA-MQM (GPT-4)	0.2989**	-0.0297	0.4535**	-0.1291
GEMBA-MQM (GPT-5.2)	0.5533**	0.2461**	0.4741**	0.3135**
Patent-Rubric-Checklist (GPT-4)	0.1906*	0.1329	0.2691**	0.1034
Patent-Rubric-Checklist (GPT-5.2)	**0.6214****	**0.3767****	**0.5931****	**0.4417****

$^{*}p < 0.05$, $^{**}p < 0.01$.

Table 2: Pearson and Spearman Correlations between Automatic Evaluation Scores and Human Evaluation Scores.

Rubric Setting	Pearson's r		Spearman's ρ	
	ja–en	en–ja	ja–en	en–ja
Full	0.6214**	0.3767**	0.5931**	0.4417**
Only A	0.4331**	0.2853**	0.4788**	0.3862**
Only B	0.4856**	0.3352**	0.4878**	0.4039**
Only C	0.5621**	0.3835**	0.5477**	0.4619**
Only D	0.6275**	**0.5239****	0.6050**	**0.5032****
Only E	**0.6314****	0.3535**	**0.6102****	0.3874**

$^{*}p < 0.05$, $^{**}p < 0.01$.

Table 3: Dimension Ablation Study of Patent-Rubric-Checklist (GPT-5.2) using Pearson and Spearman Correlations.

5.3 Comparison of Evaluation Results Using GPT-4 and GPT-5.2

We observe that GPT-5.2 achieves higher Pearson and Spearman correlations with human evaluation than GPT-4 for all LLM-as-a-Judge methods. Specifically, for the JA–EN direction, Patent-Rubric-Checklist using GPT-5.2 achieves Pearson and Spearman correlations that are 0.4308 and 0.3240 higher, respectively, than those achieved using GPT-4. We find that Patent-Rubric-Checklist using GPT-4 assigns a score of 100 to all translations produced by MT systems 2, 3, and 5, indicating that all checklist answers are "yes".

For the EN–JA direction, Patent-Rubric-Checklist using GPT-5.2 achieves Pearson and Spearman correlations that are 0.2438 and 0.3383 higher, respectively, than those achieved using GPT-4. We find that Patent-Rubric-Checklist using GPT-4 assigns a score of 100 to all translations produced from MT systems 1 and 2.

In contrast, when using GPT-5.2, we do not observe Patent-Rubric-Checklist assigning a score of 100 to all translations from any single MT system in either translation direction. Furthermore, the overall proportion of translations receiving a score of 100 using GPT-4 as the judge model is notably higher than that using GPT-5.2 (92.26% vs. 2.98% for JA–EN and 98.21% vs. 5.36% for EN–JA).

Table 4 presents an example of the checklist answers produced by the two GPT models and corresponding externally calculated rule-based scores for the same translation. For the same MT output, GPT-4 answers "yes" to all checklist items, while GPT-5.2 answers "yes" to only one of them. Consequently, the rule-based scores calculated from the two GPT models' checklist answers are substantially different. In particular, although the English translation consists of multiple sentences, GPT-4 fails to detect this error when answering checklist item D1.

6 Discussion

The novelty of our proposed method is that we prompt the GPT model with a rubric consisting of five evaluation dimensions and corresponding checklist items designed specifically for patent claim translation evaluation. The prompt for our proposed method is for Japanese–English translation evaluation. To adapt the prompt to other language pairs, it is important to design a rubric that reflects the language-pair-specific characteristics for patent claim translation evaluation. For example, we can provide patent claims, examples of translation errors or human evaluation criteria in the target language pair to the LLM and instruct it to analyze important evaluation aspects and adapt the rubric, including its checklist items, to the target language pair. The results of this study suggest that designing a rubric with the assistance of coding agents may be a practical approach to patent claim translation evaluation, since documents explaining the guidelines for filing U.S. patent applications and differences between Japanese and U.S. patents are widely available on the internet.

Moreover, the MT outputs in the human evaluation data provided by the WAT2025 organizers include translations generated by a GPT-5-based

Source Text
[請求項 1] 水面下で軸線が前後方向に延びるように配置され、前側を上流側とするとともに後側を下流側とする流路を形成する筒部と、前記筒部の内部に配置され、前記軸線方向に延びる軸部と、前記軸部に装着され、前記流路内で前記軸線の径方向に延びるとともに前記軸線の周方向に配列された複数のプロペラ翼を有し、前記軸線回りに回転可能なプロペラと、前記流路内で前記プロペラに対して前側及び後側の少なくともいずれか一方に設けられ、前記径方向に延びるとともに前記周方向に配列され、前記軸部を支持する複数のストラットと、を備え、前記複数のストラットは、上下方向に延びて前記軸線を通る対称線に対して左右対称に配置され、前記ストラット同士の角度間隔のうち少なくとも１つの角度間隔は、他の前記ストラット同士の角度間隔と異なり、前記軸線を通るように水平方向に延在する水平面に対して、上下方向両側に少なくとも１つの前記ストラットが配置されている、推進装置。
MT Output
[Claim 1] A propulsion device comprising: a cylindrical portion arranged with its axis extending in a forward-rearward direction submerged below the water surface, with one end positioned upstream and the other end positioned downstream to form a flow channel; an axis portion arranged inside the cylindrical portion and extending along the axis direction; a propeller mounted on the axis portion, having multiple propeller blades extending radially along the axis direction and arranged circumferentially around the axis, capable of rotating about the axis; and multiple struts positioned either forward or rearward of the propeller within the flow channel, extending radially and arranged circumferentially to support the axis portion. The multiple struts extend vertically and are arranged symmetrically with respect to a vertical symmetry line passing through the axis direction. At least one angle interval between the struts differs from other angle intervals between the struts. At least one strut is positioned on both sides of the vertical direction relative to a horizontal plane extending horizontally through the axis direction.
Answer to D1 from Patent-Rubric-Checklist (GPT-4)

`"D1": {`
`  "answer": "yes",`
`  "justification": "The translation is written as a single sentence claim."`
`},`

Answer to D1 from Patent-Rubric-Checklist (GPT-5.2)

`"D1": {`
`  "answer": "no",`
`  "justification": "The English claim is written as multiple sentences; patent claims should be a single sentence, and the source is a single-sentence claim."`
`},`

Checklist Answers from Patent-Rubric-Checklist (GPT-4) and Rule-Based Score

`A1: yes, A2: yes, A3: yes,`
`B1: yes, B2: yes, B3: yes,`
`C1: yes, C2: yes, C3: yes,`
`D1: yes, D2: yes, D3: yes, D4: yes,`
`E1: yes, E2: yes, E3: yes`
`rule_based_score: 100`

Checklist Answers from Patent-Rubric-Checklist (GPT-5.2) and Rule-Based Score

`A1: partial, A2: partial, A3: partial,`
`B1: partial, B2: partial, B3: yes,`
`C1: partial, C2: partial, C3: partial,`
`D1: no, D2: yes, D3: partial, D4: partial,`
`E1: partial, E2: partial, E3: partial`
`rule_based_score: 54.1666666666666`

Table 4: Example Answers from the Proposed Method Using GPT-4 and GPT-5.2.

system (Edamatsu et al., 2025; Nakazawa et al., 2025). This raises the possibility that the LLM-as-a-Judge using GPT-4 answered "yes" to every checklist item when evaluating outputs from these systems.

7 Conclusion

This study proposed a rubric-based LLM-as-a-Judge method specific to patent claim translation evaluation. Specifically, our proposed method instructs a GPT model to evaluate MT outputs based on a rubric consisting of five evaluation dimensions and corresponding checklist items. On both the JA–EN and the EN–JA test sets show that our proposed method using GPT-5.2 achieves higher correlations with human evaluation than GEMBA-DA and GEMBA-MQM using either GPT-4 or GPT-5.2, suggesting the potential of the rubric-based LLM-as-a-Judge method for patent claim translation evaluation. In the dimension ablation study, we find that the prompt variants containing only dimension D and E, which have relatively lower weights in the proposed method, achieve the highest Spearman and Pearson correlations, respectively. We also observe that GPT-5.2 achieves higher correlations with human evaluation than GPT-4 for all LLM-as-a-Judge methods. In particular, all translations from some MT systems in each translation direction receive a score of 100 under the proposed method using GPT-4, indicating that GPT-4 answers every checklist item with "yes". In contrast, no similar pattern is observed in the results obtained using GPT-5.2. Moreover, we observe that all automatic evaluation methods yield lower correlations on the EN–JA test set than on the JA–EN test set.

Future work will focus on improving the design of the rubric and checklist items to better align with human evaluation. We will also optimize the dimension weights to develop a scoring rule that yields higher correlation with human evaluation. For example, we could use a held-out set to optimize the dimension weights of the rubric.

Limitations

Although this study suggests the potential of a rubric-based LLM-as-a-Judge method for patent claim translation evaluation, several limitations remain.

First, the dataset used in this study does not include reference translations, therefore, the performance of the proposed method when reference translations are available remains unknown. Second, the dataset is limited, only consisting of 28 claims with translations from six MT systems for the JA–EN direction and 28 claims with translations from four MT systems for the EN–JA direction. Third, we run each prompt only once on the test set, therefore, we cannot assess the stability of the LLM-as-a-Judge methods. Finally, we use only GPT-4 and GPT-5.2 as the judge models. Consequently, the performance of the proposed method using open-weight LLMs such as Qwen and Llama remains unknown.

References

Haruto Azami, Yin Zhang, Futo Kajita, Nobuyori Nishimura, and Takehito Utsuro. 2025. UTSK25 at WAT2025 patent claims translation/evaluation task. In *Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025)*, pages 30–34, Mumbai, India. Association for Computational Linguistics.

Taishi Edamatsu, Isao Goto, and Takashi Ninomiya. 2025. Ehime-U system with judge and refinement, specialized prompting, and few-shot for the patent claim translation task at WAT 2025. In *Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025)*, pages 16–29, Mumbai, India. Association for Computational Linguistics.

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.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007. Overview of the patent retrieval task at the ntcir-6 workshop. In *NTCIR*.

Marcin Junczys-Dowmunt, Bruno Pouliquen, and Christophe Mazenc. 2016. Coppa v2. 0: Corpus of parallel patent applications building large parallel corpora with gnu make. In *4th Workshop on Challenges in the Management of Large Corpora Workshop Programme*, pages 15–19.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. 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, Tampere, Finland. European Association for Machine Translation.

Masaaki Nagata, Makoto Morishita, Katsuki Chousa, and Norihito Yasuda. 2024. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9452–9462, Torino, Italia. ELRA and ICCL.

Toshiaki Nakazawa, Takashi Tsunakawa, Isao Goto, Kazuhiro Kasada, Katsuhito Sudoh, Shoichi Okuyama, Takashi Ieda, and Masaaki Nagata. 2025. Findings of the first patent claims translation task at WAT2025. In *Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025)*, pages 1–15, Mumbai, India. Association for Computational Linguistics.

OpenAI. 2025. Introducing GPT-5.2. `https://openai.com/index/introducing-gpt-5-2/`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de

Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. 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, Online. Association for Computational Linguistics.

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.

Felipe Soares, Mark Stevenson, Diego Bartolome, and Anna Zaretskaya. 2020. ParaPat: The multi-million sentences parallel corpus of patents abstracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3769–3774, Marseille, France. European Language Resources Association.

Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English patent parallel corpus. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark.

Ran Zhang, Wei Zhao, Lieve Macken, and Steffen Eger. 2025. LiTransProQA: An LLM-based literary translation evaluation metric with professional question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29099–29121, Suzhou, China. Association for Computational Linguistics.

A Prompts Used for Each Method in Experiments

This appendix presents the prompts used for GEMBA-DA, GEMBA-MQM and dimension ablation study in our experiments, as shown in Figures 4–11

B Case Studies for Dimension Ablation

We select two cases for each translation direction based on the scores obtained using the single-dimension prompt variants: one where the average single-dimension score across dimensions A–C exceeds that across dimensions D–E by the largest margin, and one in which the average single-dimension score across dimensions D–E exceeds that across dimensions A–C by the largest margin. The five dimensions A–E concern [A] Content Fidelity, [B] Element Coverage, [C] Claim Scope & Logical Constraints, [D] Claim Form & Syntactic Requirements, and [E] Fluency & Grammaticality. Tables 5–8 present the four selected examples. All calculations use unrounded values, while the scores reported in the tables are rounded to two decimal places.

In both cases where the average score for dimensions D–E is lower than that for dimensions A–C, the translation fails to preserve the single-sentence structure of the patent claim (`D1: no`). Moreover, in the EN–JA example where the average score for dimensions A–C is higher than that for dimensions D–E (Table 7), the same sentence-splitting also affects the judgment for C1, resulting in an answer of "partial", because, according to the LLM's explanation, splitting the claim may marginally blur the intended single-claim constraint framing.

```
Score the following translation from {source_lang} to{target_lang} on a continuous scale from 0 to 100, where a score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

{source_lang} source: "{source_seg}"
{target_lang} translation: "{target_seg}"
Score:
```

Figure 4: Prompt for GEMBA-DA.

```
(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_lang} source:
```{source_seg}```
{target_lang} translation:
```{target_seg}```

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.
Each 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.
```

Figure 5: Prompt for GEMBA-MQM.

```
You are an expert evaluator of MACHINE TRANSLATION OF PATENT CLAIMS.
Your task is to evaluate the translation quality using a
rubric-based evaluation framework,
implemented explicitly as a CHECKLIST.
This is NOT general prose.
The input is a PATENT CLAIM, which has strict legal and structural constraints.

You must:
(1) Answer ALL checklist questions below using:
    - "yes" (no problem)
    - "partial" (minor or weakening problem)
    - "no" (clear or serious problem)
(2) For each question, provide a short justification.
(3) Provide a brief explanation of the evaluation result.
(4) Evaluate SOURCE vs HYPOTHESIS only; no reference is provided.

```

Figure 6: Shared Part of the Single-Dimension Prompts for Patent-Rubric-Checklist.

```
--------------------
CHECKLIST (PATENT-CLAIM-SPECIFIC)
--------------------

[A] CONTENT FIDELITY (Meaning Preservation)

A1. Is the core inventive idea (what is being claimed) preserved?
A2. Are the technical functions or effects preserved without distortion?
A3. Is the translation free from hallucinated technical content not present in the
    source?

--------------------
INPUT FORMAT
--------------------
{source_lang} source patent claim:
"{source_seg}"

{target_lang} translated patent claim:
"{target_seg}"

--------------------
OUTPUT FORMAT (JSON ONLY)
--------------------

{{
  "checklist_answers": {{
    "A1": {{"answer": "yes", "justification": "Short justification."}},
    "A2": {{"answer": "yes", "justification": "Short justification."}},
    "A3": {{"answer": "yes", "justification": "Short justification."}}
  }},
  "overall_explanation": "Brief explanation of the evaluation result in 1-2 sentences."
}}
```

Figure 7: Prompt for Patent-Rubric-Checklist with Only Dimension A: Rubric, Checklist Items and Input/Output Format.

```
--------------------
CHECKLIST (PATENT-CLAIM-SPECIFIC)
--------------------

[B] ELEMENT COVERAGE (Completeness)

B1. Are all essential claim elements translated (no omissions)?
B2. Are relationships between elements (having, connected to, arranged on, etc.)
    preserved?
B3. Are quantities, ranges, or conditions (e.g. "at least one", numerical ranges)
    preserved?

--------------------
INPUT FORMAT
--------------------
{source_lang} source patent claim:
"{source_seg}"

{target_lang} translated patent claim:
"{target_seg}"

--------------------
OUTPUT FORMAT (JSON ONLY)
--------------------

{{
  "checklist_answers": {{
    "B1": {{"answer": "yes", "justification": "Short justification."}},
    "B2": {{"answer": "yes", "justification": "Short justification."}},
    "B3": {{"answer": "yes", "justification": "Short justification."}}
    }},
  "overall_explanation": "Brief explanation of the evaluation result in 1-2 sentences."
}}
```

Figure 8: Prompt for Patent-Rubric-Checklist with Only Dimension B: Rubric, Checklist Items and Input/Output Format.

```
--------------------
CHECKLIST (PATENT-CLAIM-SPECIFIC)
--------------------

[C] CLAIM SCOPE & LOGICAL CONSTRAINTS (CRITICAL)

C1. Is the claiming scope preserved?
    (e.g., "comprising" vs "consisting of", "characterized in that")
C2. Are constraint clauses correctly attached?
    (e.g., "wherein", "前記◯◯が〜")
C3. Are coordination vs modification relations preserved?
    (e.g., A and B that each have C, NOT A and (B that has C))

--------------------
INPUT FORMAT
--------------------
{source_lang} source patent claim:
"{source_seg}"

{target_lang} translated patent claim:
"{target_seg}"

--------------------
OUTPUT FORMAT (JSON ONLY)
--------------------

{{
  "checklist_answers": {{
    "C1": {{"answer": "yes", "justification": "Short justification."}},
    "C2": {{"answer": "yes", "justification": "Short justification."}},
    "C3": {{"answer": "yes", "justification": "Short justification."}}
  }},
 "overall_explanation": "Brief explanation of the evaluation result in 1-2 sentences."
}}
```

Figure 9: Prompt for Patent-Rubric-Checklist with Only Dimension C: Rubric, Checklist Items and Input/Output Format.

```
--------------------
CHECKLIST (PATENT-CLAIM-SPECIFIC)
--------------------

[D] CLAIM FORM & SYNTACTIC REQUIREMENTS

D1. Is the translation written as a SINGLE SENTENCE claim?
D2. Is the claim type preserved (apparatus / method / composition, etc.)?
D3. Is antecedent basis consistent?
(a/the/said, 前記〜)
D4. Does the translation avoid narrative or explanatory prose style?

--------------------
INPUT FORMAT
--------------------
{source_lang} source patent claim:
"{source_seg}"

{target_lang} translated patent claim:
"{target_seg}"

--------------------
OUTPUT FORMAT (JSON ONLY)
--------------------

{{
  "checklist_answers": {{
    "D1": {{"answer": "yes", "justification": "Short justification."}},
    "D2": {{"answer": "yes", "justification": "Short justification."}},
    "D3": {{"answer": "yes", "justification": "Short justification."}},
    "D4": {{"answer": "yes", "justification": "Short justification."}}
  }},
 "overall_explanation": "Brief explanation of the evaluation result in 1-2 sentences."
}}
```

Figure 10: Prompt for Patent-Rubric-Checklist with Only Dimension D: Rubric, Checklist Items and Input/Output Format.

```
--------------------
CHECKLIST (PATENT-CLAIM-SPECIFIC)
--------------------

[E] FLUENCY & GRAMMATICALITY

E1. Is the translation grammatically correct in the target language?
E2. Is it stylistically appropriate for patent claims?
E3. Is the sentence readable and not excessively awkward or ambiguous?

--------------------
INPUT FORMAT
--------------------
{source_lang} source patent claim:
"{source_seg}"

{target_lang} translated patent claim:
"{target_seg}"

--------------------
OUTPUT FORMAT (JSON ONLY)
--------------------

{{
  "checklist_answers": {{
  "E1": {{"answer": "yes", "justification": "Short justification."}},
  "E2": {{"answer": "yes", "justification": "Short justification."}},
  "E3": {{"answer": "yes", "justification": "Short justification."}}
  }},
  "overall_explanation": "Brief explanation of the evaluation result in 1-2 sentences."
}},
```

Figure 11: Prompt for Patent-Rubric-Checklist with Only Dimension E: Rubric, Checklist Items and Input/Output Format.

Source Text

[請求項 3] Ｎ本の前記ストラットのそれぞれが、前記軸線回りの領域を前記周方向に均等に分割する（Ｎ × ２ − １）本の分割線のいずれかに配置され、１以上Ｎ未満であるＮ’に対して、前記ストラット同士の角度間隔は、Ｎ’ × ３６０／（Ｎ × ２ − １）度のいずれかである、請求項１又は２に記載の推進装置。

MT Output

[Claim 3] In a propulsion device according to Claims 1 or 2, each of the N aforementioned struts is positioned along one of the (N×2 − 1) dividing lines that evenly divide the area around the axis line in a circumferential direction. For N' where 1 ≤ N' < N, the angular spacing between the struts is either N'×360/(N×2 − 1) degrees.

Dimension A

Checklist answers: `A1: partial, A2: yes, A3: yes`

Overall explanation: The translation largely preserves the claim' s technical meaning and mathematical constraints. The main issue is a minor loss of legal precision regarding whether N' is explicitly an integer/natural number, which could slightly broaden interpretation.

Rule-based score: 83.33

Dimension B

Checklist answers: `B1: yes, B2: yes, B3: yes`

Overall explanation: The translation preserves the claim' s structural elements, inter-element relationships, and mathematical/conditional limitations. Any wording differences (e.g., "area around the axis line") do not materially change the technical/legal constraints expressed in the source.

Rule-based score: 100.00

Dimension C

Checklist answers: `C1: partial, C2: yes, C3: yes`

Overall explanation: The translation is structurally faithful and preserves the main logical constraints of the claim. The only notable issue is a slight potential weakening/ambiguity in expressing the "one of" set-selection over N' values as "either" and as a per-N' phrasing.

Rule-based score: 83.33

Dimension D

Checklist answers: `D1: no, D2: yes, D3: partial, D4: partial`

Overall explanation: The translation generally preserves the technical/legal content and claim type, but it violates a key formal requirement by breaking the claim into two sentences. Minor issues also exist with antecedent tightness and slightly explanatory phrasing.

Rule-based score: 50.00

Dimension E

Checklist answers: `E1: yes, E2: partial, E3: partial`

Overall explanation: The translation is grammatically correct and largely understandable, but it is slightly non-idiomatic for patent-claim style and introduces mild ambiguity/awkwardness in key geometric phrasing and in expressing the set of allowable angular spacings.

Rule-based score: 66.67

Score Summary

A–C average: 88.89

D–E average: 58.33

Gap: 30.56

Human evaluation score: 50

Table 5: A JA–EN Case Where the Average Single-Dimension Score for Dimensions A–C Is Higher than That for Dimensions D–E.

Source Text

[請求項 1] 特定の種別の光の方路切替が可能な第１振分部と、前記第１振分部で取り扱うことができる特定の種別の光と、前記第１振分部で取り扱うことができない特定の種別の光の方路切替が可能な第２振分部と、を備え、前記第２振分部は、入力された前記光を、前記第１振分部を経由する第１経路と、前記第１振分部を経由しない第２経路の少なくともいずれかの経路に出力する光出力部と、前記第２経路上に設けられ、Ｎ（Ｎは１以上の整数）個のポートとｋ（ｋは２以上の整数）個のポートとを有し、前記Ｎ個のポートのいずれかのポートに入力された前記光を、他の前記第２振分部が接続されている前記ｋ個のポートのいずれかのポートから出力するＮ × ｋマトリックス光スイッチと、を備え、前記光出力部は、端末間において波長単位の光パスを確立する場合には入力された前記光を少なくとも前記第１経路に出力し、光ファイバ単位の光パスを確立する場合には入力された前記光を少なくとも前記第２経路に出力する、を備える光ノード装置。

MT Output

1. An optical node apparatus comprising: a first distribution unit configured to switch an optical path of a specific type of light; and a second distribution unit configured to switch optical paths of: (i) a specific type of light that can be handled by the first distribution unit; and (ii) a specific type of light that cannot be handled by the first distribution unit, the second distribution unit comprising: a light output unit configured to output the input light to at least one of: (i) a first path that passes through the first distribution unit; and (ii) a second path that does not pass through the first distribution unit; and an N x k matrix optical switch provided on the second path and having N (N is an integer equal to or greater than 1) ports and k (k is an integer equal to or greater than 2) ports, each of the k ports being connected to another second distribution unit, the N x k matrix optical switch being configured such that light input to any one of the N ports is output from any one of the k ports; wherein: when establishing an optical path between terminals on a wavelength basis, the light output unit is configured to output the input light to at least the first path; and when establishing an optical path on an optical fiber basis, the light output unit is configured to output the input light to at least the second path.

Dimension A

Checklist answers: `A1: yes, A2: partial, A3: partial`

Overall explanation: The translation largely preserves the claimed apparatus and its conditional routing behavior, with only a minor fidelity risk where the connectivity of the k ports is overstated (universalized) compared to the Japanese, slightly narrowing/altering the technical scope.

Rule-based score: 66.67

Dimension B

Checklist answers: `B1: partial, B2: partial, B3: partial`

Overall explanation: The translation largely preserves the claim’s elements and structure, but it introduces a potentially material scope change by strengthening the connectivity requirement for the k ports (from ’any of the k ports’ to ’each of the k ports’), which can affect legal completeness and relationships.

Rule-based score: 50.00

Dimension C

Checklist answers: `C1: partial, C2: partial, C3: yes`

Overall explanation: The translation mostly preserves the claim’s structure and functional limitations, but it introduces a notable scope/logic deviation around the N×k switch’s k-port connectivity/output constraint, potentially altering the claimed network connection requirement.

Rule-based score: 66.67

Dimension D

Checklist answers: `D1: yes, D2: yes, D3: partial, D4: yes`

Overall explanation: Overall, the translation conforms well to patent-claim form and preserves the apparatus nature and structure. The main issue is a minor antecedent-basis looseness around “the input light,” which could be tightened for stricter claim clarity.

Rule-based score: 87.50

Dimension E

Checklist answers: `E1: yes, E2: yes, E3: partial`

Overall explanation: The translation is fluent and largely conforms to patent-claim English, with correct grammar and conventional claim drafting patterns. Minor awkward constructions introduce some interpretive ambiguity, but do not render the claim unreadable.

Rule-based score: 83.33

Score Summary

A–C average: 61.11

D–E average: 85.42

Gap: -24.31

Human evaluation score: 60

Table 6: A JA–EN Case Where the Average Single-Dimension Score for Dimensions D–E Is Higher than That for Dimensions A–C.

Source Text

1. A system including a first electronic device and a cell phone, the first electronic device including a processor and a wired or wireless interface to a network, said first electronic device being adapted to respond to user instructions by changing device state, the cell phone being equipped with a camera and a touchscreen, characterized in that the first electronic device lacks buttons by which the user instructions can be entered, the cell phone being adapted to communicate with the first electronic device over said network, the cell phone further including software instructions adapted to present on the cell phone touchscreen a user interface including Up and Down buttons by which a user can enter user control instructions for transmission to said first electronic device to control said device state, the software instructions also including instructions to obtain identification information corresponding to the first electronic device by decoding information from a 2D machine readable code on the first electronic device that is sensed by a camera of the cell phone, to thereby authorize the cell phone to interact with the first electronic device.

MT Output

1. 第 1 の電子装置と携帯電話を含むシステムであって、第 1 の電子装置はプロセッサと、ネットワークへの有線または無線インターフェースを備えており、この第 1 の電子装置はユーザ指示に応じて装置の状態を変化させることで応答するように構成され、携帯電話にはカメラとタッチスクリーンが装備されている。本システムは、第 1 の電子装置がユーザ指示を入力するためのボタンを有しておらず、携帯電話が当該ネットワークを介して第 1 の電子装置と通信するように構成されている点を特徴とする。さらに、携帯電話には、携帯電話のタッチスクリーン上にユーザインターフェースを表示するように適応されたソフトウェア指示が含まれており、このユーザインターフェースには上下ボタンが設けられており、ユーザはこのボタンを用いて第 1 の電子装置に送信して装置の状態を制御するためのユーザ制御指示を入力することができる。これらのソフトウェア指示には、携帯電話のカメラによって検知される第 1 の電子装置上の 2 次元機械読み取り可能コードから情報を解読することにより、第 1 の電子装置に対応する識別情報を取得する指示も含まれており、これにより携帯電話が第 1 の電子装置と相互作用することを許可する。

Dimension A

Checklist answers: `A1: yes, A2: yes, A3: yes`

Overall explanation: 翻訳は請求項の主要構成要件と機能的関係（UI による制御指示送信、2D コード読取りによる識別情報取得と認可）を適切に保持しており、意味の逸脱や不要な技術的付加は認められない。

Rule-based score: 100.00

Dimension B

Checklist answers: `B1: yes, B2: yes, B3: yes`

Overall explanation: The Japanese claim translation is complete and preserves the structural relationships and functional limitations of the English claim. No material loss of elements, conditions, or inter-element dependencies is observed.

Rule-based score: 100.00

Dimension C

Checklist answers: `C1: partial, C2: yes, C3: yes`

Overall explanation: The Japanese translation preserves the main claim scope and the logical attachment of limitations, with correct nesting of functional features. The only notable issue is a mild structural shift by splitting the claim into multiple sentences around the “characterized” portion, which could marginally blur the intended single-claim constraint framing.

Rule-based score: 83.33

Dimension D

Checklist answers: `D1: no, D2: partial, D3: partial, D4: no`

Overall explanation: 技術内容自体は概ね対応しているものの、訳文が複数文に分割され、説明調の接続語を多用しているため、特許請求の範囲としての単一文・形式要件を満たさず、法的明確性（前記関係の厳密さ）も弱まっている。

Rule-based score: 25.00

Dimension E

Checklist answers: `E1: yes, E2: partial, E3: partial`

Overall explanation: 文法的には概ね良好で内容も追える一方、請求項としては文分割と直訳調表現により、クレームの一体性・限定関係の明確さがやや弱まっている。

Rule-based score: 66.67

Score Summary

A–C average: 94.44

D–E average: 45.83

Gap: 48.61

Human evaluation score: 45

Table 7: An EN–JA Case Where the Average Single-Dimension Score for Dimensions A–C Is Higher than That for Dimensions D–E.

Source Text

1. A method for unmanned aerial vehicle handover, comprising: determining, by a source base station, that one or more base stations meet one or more handover conditions based on a measurement report from an unmanned aerial vehicle; determining, by the source base station, a candidate base station that has completed a handover preparation among the one or more base stations meeting the one or more handover conditions; handing over, by the source base station, the unmanned aerial vehicle to the candidate base station meeting the one or more handover conditions; determining, by the source base station, one or more accessible base stations based on a flight path of the unmanned aerial vehicle and geographical location information of the one or more accessible base stations; selecting, by the source base station, a base station from the one or more accessible base stations based on a flight speed, a flight altitude, or the flight speed and the flight altitude of the unmanned aerial vehicle, wherein the selected base station is to perform a handover preparation operation, and the accessible base station is a base station that the unmanned aerial vehicle is able to access when flying according to the flight path; performing, by the source base station, a handover preparation operation for handing over the base station that is to perform the handover preparation operation; and determining, by the source base station, the base station that has successfully performed the handover preparation as the candidate base station that has completed the handover preparation.

MT Output

１つ又は複数の基地局が無人航空機からの測定レポートに基づいて１つ又は複数のハンドオーバ条件を満たすことを決定することと、前記１つ又は複数のハンドオーバ条件を満たす１つ又は複数の基地局のうち、ハンドオーバ準備を完了した候補基地局を決定することと、前記１つ又は複数のハンドオーバ条件を満たす前記候補基地局に前記無人航空機をハンドオーバすることと、前記無人航空機の飛行経路及び１つ又は複数のアクセス可能な基地局の地理的位置情報に基づいて、前記１つ又は複数のアクセス可能な基地局を決定することと、前記無人航空機の飛行速度、飛行高度、又は前記飛行速度及び前記飛行高度に基づいて、前記１つ又は複数のアクセス可能な基地局から１つの基地局を選択することであって、前記選択された基地局は、ハンドオーバ準備動作を実行する基地局であり、前記アクセス可能な基地局は、前記飛行経路に従って飛行するときに前記無人航空機がアクセス可能な基地局である、選択することと、前記ハンドオーバ準備動作を実行する基地局に対してハンドオーバ準備動作を実行することと、前記ハンドオーバ準備を完了した前記候補基地局として、前記ハンドオーバ準備を正常に完了した基地局を決定することと、を含む、無人航空機ハンドオーバ方法。

Dimension A

Checklist answers: `A1: partial, A2: partial, A3: yes`

Overall explanation: The translation largely preserves the claimed method steps and technical dependencies, but it weakens fidelity by dropping explicit attribution to the source base station and by slightly distorting the phrasing of the handover-preparation step, which could impact claim clarity/scope.

Rule-based score: 66.67

Dimension B

Checklist answers: `B1: partial, B2: partial, B3: yes`

Overall explanation: The Japanese translation largely covers the claim steps and preserves key conditions and quantifiers, but it weakens legal precision by omitting the repeated limitation that actions are performed by the source base station and by introducing some ambiguity in the handover-preparation-for-handover relationship.

Rule-based score: 66.67

Dimension C

Checklist answers: `C1: partial, C2: partial, C3: partial`

Overall explanation: The translation largely preserves the claim' s overall structure and major limitations, but a critical mid-to-late step concerning the handover preparation operation is rendered with an altered object/relationship, and the 'wherein' attachments are somewhat entangled, potentially affecting claim logic and scope.

Rule-based score: 50.00

Dimension D

Checklist answers: `D1: yes, D2: yes, D3: partial, D4: yes`

Overall explanation: 全体として単一文の方法クレームとしての体裁は保たれており、列挙構造も適切です。一方で「前記」の付与が初出要素に及ぶ箇所があり、先行詞の厳密性という特許請求項特有の要件に軽微な弱点があります。

Rule-based score: 87.50

Dimension E

Checklist answers: `E1: partial, E2: yes, E3: partial`

Overall explanation: 特許請求項としての体裁と用語運用は概ね適切で、全体の意味も追える一方、いくつかの箇所で係り受け・目的語関係が不自然で可読性と明確性がやや低下している。

Rule-based score: 66.67

Score Summary

A–C average: 61.11

D–E average: 77.08

Gap: -15.97

Human evaluation score: 65

Table 8: An EN–JA Case Where the Average Single-Dimension Score for Dimensions D–E Is Higher than That for Dimensions A–C.