Ehime-U System with Judge and Refinement, Specialized Prompting, and Few-shot for the Patent Claim Translation Task at WAT 2025

Taishi Edamatsu
Ehime University
edamatsu@ai.cs.ehime-u.ac.jp

Isao Goto
Ehime University
goto.isao.fn@ehime-u.ac.jp

Takashi Ninomiya
Ehime University
ninomiya.takashi.mk@ehime-u.ac.jp

Abstract

The Ehime University team participated in the Japanese-to-English Patent Claim Translation Task at WAT 2025. We experimented with (i) Judge and Refinement, (ii) Specialized Prompting, and (iii) Few-Shot Prompting. We used GPT-5 as the LLM. Evaluation based on the LLM-as-a-Judge framework confirmed improvements for (i), while (ii) and (iii) showed no significant effects. On the other hand, the official human evaluation indicated that the translation quality of method (i) decreased.

1 Introduction

In patent documents, patent claims represent a critically important section defining the scope of rights. Patent claims often consist of extremely long sentences with complex structures, making it difficult to translate them while maintaining correct legal interpretation. Additionally, selecting appropriate translations for patent-specific expressions and technical terminology presents challenges. The emergence of large language models (LLMs) in recent years has enabled machine translation to achieve results surpassing existing tasks. In the patent claim translation task, Azami et al. (2025) performed continued pre-training and fine-tuning of publicly available LLMs using parallel patent-translation data. However, human evaluation was not conducted for patent claim translations, leaving the challenges in patent claim translation unclear.

This paper describes the Ehime-U team's Japanese-to-English translation system for the WAT2025 patent claim translation task. We implemented three approaches in our LLM-based translation system. First, to address the issue that the challenges in patent claim machine translation have not been clearly identified, we introduce (i) Judge and Refinement based on the method of Chen et al. (2024) and (ii) Specialized Prompting . Furthermore, to improve terminology selection and consistency, we search training data for usage examples and employed them as (iii) Few-Shot training. We use GPT-5 as the base LLM. Evaluation using the LLM-as-a-Judge framework confirmed the effectiveness of (i) Judge and Refinement. However, (ii) Specialized Prompting and (iii) Few-Shot showed no discernible effect. On the other hand, the official human evaluation, which assessed only method (i), showed no improvement of method (i). This result indicates that the performance of the LLM-as-a-Judge framework was not sufficient in this case. Although the three methods evaluated in this study improved surface-level quality errors, we observed an increase in errors related to the fidelity of the original patent claims. This suggests that, when constraints are imposed through prompting, the LLM used in this work struggles to satisfy those constraints without degrading the overall fidelity of the content.

2 System Description

In this section, we describe the three techniques incorporated into our system: Judge and Refinement, Specialized Prompting, and Few-Shot.

2.1 Judge and Refinement

Judge and Refinement (Judge&Refinement) consists of three processing stages, and the procedure of each stage is described in order.

(1) Base Translation The Japanese patent claims are translated by an LLM on a per-claim basis while preserving line breaks. We defined a `PROMPT_POLICY` for the model as follows:

- Ensuring fidelity to the source text, including prohibiting additions, omissions, changes in legal meaning or legal scope, alterations of dependencies, and modifications of numerical values;

Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025), pages 16–29

- Enforcing the distinction between independent claims, which are written without referencing preceding claims, and dependent claims, which must explicitly reference preceding claims;
- Standardizing punctuation;
- consistent antecedent references;
- complete preservation of numerical values, units, and formulas;
- consistent terminology across technical domains.

After that, We instructed the model to translate Japanese patent claims into U.S.-style English claims by using the policy as the persona of a professional patent-claim translator. In addition, we instructed the model not to add any annotations and to avoid any addition, omission, splitting, or merging of content. The detailed prompt is shown in Figure 1 in Appendix A. This method is called as Base Translation.

(2) Judge Using the source text and the generated translation, an LLM as a Judge evaluates the translation quality. The evaluation is conducted across the six criteria (Table 1), and an overall score (0–100 points) is calculated by averaging them equally. The detailed prompt is shown in Figure 3 in Appendix A.

(3) Refinement Without using any reference translations, the model is instructed to automatically extract and organize translation errors from the evaluation report, and then retranslate accordingly. From the LLM evaluation results obtained in (2), the model performs knowledge distillation to generalize the insights useful for refinement. Instead of focusing on specific errors (e.g., individual grammar or lexical mistakes), it abstracts recurring error patterns and systematic weaknesses into generalized categories, which serve as revision policies for refinement. Since the goal is to apply generic rather than case-specific corrections, all specific and unique information are removed, and each error is labeled according to one of the six categories used in (2). Common patterns within each category are then rewritten into rule-like sentences, typically following a two-part structure: "Symptom → Expected Form." For example: Symptom: "Range expressions use 'X–Y'." Expected Form: "Write 'X to Y' in ascending order." This design clarifies the purpose of the correction while avoiding semantic changes or redundant fixes.

we provide the extracted evaluation results, the Japanese source text, and the Base Translation as input to the LLM, expecting it to produce an English output with only minimal modifications. Here as well, we instructed the model to translate the text into U.S.-style English patent claims, in the same manner as in the Base Translation. The detailed prompt is shown in Figures 4 and 5 in Appendix A.

2.2 Specialized Prompting

In this section, we describe three methods for improving the translation prompts introduced in Section 2.1 to achieve translations that adhere more closely to U.S. claim conventions.

Specialized Base Translation

Instead of the simple instruction in Section 2.1 (1), "Translate into U.S. claim style," we adopt a strict audit-based translation prompt. The main revisions are as follows:

- **Pre-output audit (SILENT QA)**: The model self-verifies claim type, numbers/units, antecedents, and sentence structure before output.
- **Stronger output constraints**: Restriction to ASCII only, single-sentence structure, and enforcement of "colon + semicolon + ; and" pattern.
- **Explicit prohibitions**: Elimination of "and/or," non-ASCII symbols, ambiguous pronouns, and unnecessary respectively.
- **Fixed terminology and style**: Explicit enforcement of standard phrases such as "apparatus," "configured to," and "equal to or greater than ...".

This enables the translator to function simultaneously as a self-auditing agent, ensuring both legal and structural consistency. The detailed prompt is shown in Figure 6 in Appendix B.

Select of Evaluation Results

The phase that extracts only the information necessary for refinement from the evaluation output is

Criterion	Description
fidelity_legal scope	Fidelity to legal scope and limitations
us_style structure	Conformity to the format and structure of U.S. patent claims
numbers_units ranges	Accuracy of numbers, units, ranges, and formulas
antecedent dependency	Consistency of antecedents and referential dependencies
terminology	Accuracy and consistency of terminology
naturalness	Naturalness and readability of expressions

Table 1: Evaluation criteria used in the LLM-as-a-Judge framework.

	Development Data	Test Data
Number of patents	13	26
Number of claims (sentences)	19	70

Table 2: number of claims

redesigned as a systematic error-category extraction prompt as follows:

- **Priority of extraction**: Fidelity > dependency > numbers/units > legal format.
- **Controlled output volume**: Limited to 10–15 representative issues, merging duplicates and superficial errors.
- **Unified output format**: Exampled as "Symptom > Expected Form" structure.
- **Noise filtering**: Extraction limited to essential issues that affect legal meaning.

This allows the system to identify the core issues to be fixed in refinement using the evaluation results. The detailed prompt is shown in Figure 7 Appendix B.

Refinement
In the refinement phase, a minimal-edit policy is introduced to suppress overcorrection.

- **Two-layered objective**: (1) Maximize semantic and legal consistency, (2) Preserve n-grams for minimal editing (BLEU retention).
- **Limited edit scope**: Revise only the portions listed in "Issues to fix."
- **Format revalidation**: Re-enforce the U.S. claim structure (colon, semicolon, "; and", single-sentence rule).
- **Local correction policy**: Prohibit any rephrasing beyond essential grammatical corrections.

Through this approach, refinement is defined not as a full rewrite but as a localized legal correction phase. The detailed prompt is shown in Figure 8 in Appendix B.

2.3 Few-shot Prompting

We extend the method described in Section 2.1 by incorporating a few-shot mechanism (Brown et al., 2020) using translation examples based on FAISS (Douze et al., 2024) and SentenceTransformer (Reimers and Gurevych, 2019). FAISS is a high-speed library for vector similarity search, designed to efficiently retrieve "similar vectors" from large-scale vector datasets. When constructing the FAISS index, we use bilingual Japanese–English sentence pairs from the Patent Cooperation Treaty (PCT) route portion of the JaParaPat (Nagata et al., 2024) corpus, which is the training data for this task. Under the PCT route, a single international patent application is submitted to multiple national offices through translation, making the resulting multilingual publications effectively parallel. Because these pairs represent direct translations of the same application, they can be regarded as highly reliable parallel data.

The Japanese claim sentences are embedded using the multilingual sentence embedding model (intfloat/multilingual-e5-base), enabling the system to evaluate semantic similarity between sentences based on cosine similarity. Consequently, for a given input claim, semantically similar Japanese–English pairs can be efficiently retrieved and utilized as reference examples in few-shot translation. Few-shot prompting is applied to the

System	LLM as a judge score (%)
Base Translation	91
Base Translation + Judge&Refinement	**92**
Specialized Prompting	80
Specialized Prompting + Judge&Refinement	84
Few-shot (sentence)	84
Few-shot (sentence) + Judge&Refinement	84
Few-shot (sentence) + Specialized Prompting	78
Few-shot (sentence) + Specialized Prompting + Judge&Refinement	79
Few-shot (term)	78
Few-shot (term) + Judge&Refinement	83
Few-shot (term) + Specialized Prompting	82
Few-shot (term) + Specialized Prompting + Judge&Refinement	83

Table 3: Evaluation results based on LLM as a judge for the test data

System	COMET	BLEU
Base Translation	84.59	53.81
Base Translation + Judge&Refinement	84.95	48.89
Specialized Prompting	85.35	**56.55**
Specialized Prompting + Judge&Refinement	**85.48**	55.64
Few-shot (sentence)	84.45	50.09
Few-shot (sentence) + Judge&Refinement	84.67	51.43
Few-shot (sentence) + Specialized Prompting	85.14	53.88
Few-shot (sentence) + Specialized Prompting + Judge&Refinement	85.08	52.43
Few-shot (term)	85.01	53.54
Few-shot (term) + Judge&Refinement	84.97	51.92
Few-shot (term) + Specialized Prompting	85.16	52.92
Few-shot (term) + Specialized Prompting + Judge&Refinement	85.15	52.21

Table 4: Evaluation results based on COMET and BLEU for the development data

translation and refinement stages.

Two types of few-shot examples are used in this study:

Sentence-Level Example The first method performs cosine similarity search against the FAISS index built from full-sentence vectors. The top three most similar Japanese–English pairs are retrieved and inserted into the translation prompt as few-shot (sentence) examples.

Term-Level Example To retrieve translation examples including important terms in the source sentence, the second method uses the LLM to extract three terms from the input sentence and uses them as queries. These queries are used for FAISS retrieval, and the retrieved bilingual sentence pairs are organized into a few-shot sentence. This method is referred to as few-shot (word) in the following evaluation.

Additionally, both of these few-shot methods are combined with the specialized prompt described in Section 2.2 for comparative evaluation. The detailed prompt is shown in Figure 9 in Appendix C.

2.4 LLM

In this system, we use OpenAI's GPT-5[1] as the underlying LLM.

2.5 Dataset

We use the official development and test data provided for the WAT 2025 "Patent Claims Translation / Evaluation Tasks". The development data

[1] https://platform.openai.com/docs/models/gpt-5

consist of source-language patent claims and their corresponding translations, whereas the test data contain only the source-language patent claims. We show the number of patent and patent claims for each data point in the Table 2. In addition, we use data from JaParaPat for the Few-Shot Prompting. JaParaPat is a large-scale Japanese–English parallel corpus aligned between Japanese and English patent application documents. It consists of approximately 107 million Japanese–English sentence pairs automatically extracted from patent document families filed between 2016 and 2020, and includes metadata such as application-type labels and document IDs. From this corpus, we used only the sentence pairs whose document IDs correspond to claim sections.

3 Evaluation

3.1 Our Evaluation

Patent claim translation involves very long and syntactically complex sentences, making it difficult to fully understand the structure of each claim. Furthermore, accuracy must be preserved across multiple dimensions—not only in meaning but also in legal scope and technical terminology—thus, existing automatic evaluation methods struggle to precisely assess translation adequacy. In contrast, LLM-as-a-Judge, which evaluates translations using an LLM, is expected to consistently assess the appropriateness of translations across all parts of a long sentence. Therefore, we employ the LLM-as-a-Judge as our primary evaluation method. The evaluation criteria are the same as those defined in Section 2.1 (2).

For evaluation, we use the test dataset described in Section 2.5, which does not include reference translations. The results are shown in Table 3. The Judge&Refinement configuration achieved a higher score than the Base Translation. On the other hand, the Specialized Prompting score was lower than the baseline, and both Few-shot (sentence) and Few-shot (term) also showed lower scores than the baseline. Therefore, the effectiveness of these few-shot and specialized prompting methods was not confirmed.

For reference, Table 4 presents the results of automatic evaluation using the COMET (wmt22-comet-a; Rei et al., 2022) and BLEU (sacrebleu; Post, 2018) metrics. These scores were calculated using the development dataset, which includes reference translations, instead of the test dataset.

3.2 Official Evaluation

As the official evaluation for WAT 2025, the task organizers conducted human assessment. Manual error annotations and evaluation scores were assigned to each source sentence and its translated output by human evaluators. Error annotations were assigned to problematic segments based on error categories such as mistranslation, omission, and hallucination, with each error being labeled for severity (major or minor). In addition, an official score out of 100 points was assigned to each sentence. After assigning evaluation priorities to the translation results of the test data and submitting all results shown in Table 2, two systems—Base Translation and Judge&Refinement—were evaluated by the organizers. For each system, the 28 sentences out of the 70 test sentences were evaluated by humans. The official human evaluation results for error categories and average scores are presented in Tables 5 and 6. Compared with the Base Translation, the number of major errors in the Refinement output increased from 18 to 42, and the number of minor errors increased from 119 to 150. Therefore, the total number of errors increased from 137 to 192. In addition, the average score decreased from 86.07 for Base Translation to 81.60 for Judge&Refinement.

4 Analysis

4.1 Analysis Based on the Official Evaluation

We analyze the reasons why Judge&Refinement received lower evaluation scores than Base Translation While surface-level errors—such as grammatical errors and punctuation issues involving the use of commas and semicolons—were improved, no improvements were observed for other types of errors. In particular, substantial increases were observed in hallucination, omission, and mistranslation errors, indicating a rise in errors related to the fidelity of the original patent claims. However, many of the mistranslation errors were attributable to article-related issues, such as incorrect selection of “a” or “the” and omitted articles. When these article errors are excluded, the number of remaining mistranslations becomes much closer, with 11 for Base Translation and 13 for Judge&Refinement. Although the change in the number of these errors was not large, many errors related to terminology consistency and contextual inappropriateness were also observed.

Error Category	Base Translation Major	Base Translation Minor	Judge&Refinement Major	Judge&Refinement Minor
Omission	8	13	24	14
Terminology Consistency	1	33	0	36
Grammar	1	8	2	1
Mistranslation	5	18	7	25
Other	0	2	0	5
Contextually Inappropriate	3	11	2	14
Hallucination	0	10	5	34
Source Text Error	0	3	1	2
Punctuation	0	17	0	8
Lack of Consistency	0	0	0	4
Awkward Expression	0	4	0	5
Article Error	0	0	1	2
Total Errors	18	119	42	150
Total (Major+Minor)	137		192	

Table 5: Official human evaluation results (number of error categories).

	Base translation	Judge&Refinement
Average score	86.07	81.60

Table 6: Official human evaluation results (average score)

In the Judge&Refinement method, the initial translation is evaluated using the LLM-as-a-Judge framework, and the output is refined based on the abstract error types extracted from the evaluation report. In addition to the strict U.S.-style constraints defined in the PROMPT_POLICY used for the Base Translation, the model is explicitly instructed to revise the English text in accordance with the identified issues. As a result, while surface-level improvements were made—such as corrections to grammar and punctuation, better adherence to U.S. claim style, and the introduction of common expressions used in patent translation—we consider that there was also an increase in errors related to loss of fidelity to the original text, including incorrect scope or comparison direction, erroneous antecedent references (mistranslations), the addition of elements not present in the source (hallucination), and the omission of obligatory elements (omission). In particular, the refinement step appears to prioritize producing well-formed English over maintaining strict fidelity to the source text, as it tends to rewrite the entire sentence rather than apply minimal edits.

A comparison between the official evaluation and the LLM-as-a-Judge evaluation shows that, although the score for Judge&Refinement improved under the LLM-as-a-Judge framework, its translation quality deteriorated in the human evaluation. Therefore, it was confirmed that the performance of the LLM-as-a-Judge framework was not sufficient in this study.

4.2 Analysis of Results Not Assigned Official Evaluation

For Specialized Prompting and Few-Shot Prompting, we conducted our evaluation using the LLM-as-a-Judge framework. Compared with Judge&Refinement, Specialized Prompting and Few-Shot Prompting improved consistency with U.S. patent-claim style, the naturalness of the English output, and the stability of terminology and unit expressions. As a result, their scores for us_style_structure and naturalness in Table 1 increased. However, incorrect modifications of claim scope and the insertion of erroneous dependency relations led to decreases in fidelity_legal_scope and antecedent_dependency scores. In Specialized Prompting, the model is strongly biased toward producing "natural English" and adhering to "U.S. claim style" whereas essential aspects of patent translation—such as structural preservation and legal fidelity—tend to degrade. We consider that this imbalance led to lower LLM-evaluation scores. Similarly, Few-Shot Prompting showed improvements in stylis-

tic aspects of the translation, including punctuation placement, element enumeration, lexical consistency such as the use of "configured to," and stabilization of U.S.-claim-specific sentence patterns. However, while Few-Shot Prompting improves stylistic consistency and terminology, we consider that it is strongly influenced by the structural bias of the retrieved examples, causing structural distortions in the translated output—such as reorganization of elements, shifts in clause positions, and unnecessary insertions of wherein. These issues likely resulted in substantial score reductions in the fidelity and antecedent_dependency categories of Table 1. We consider that the performance of Few-Shot Prompting declined relative to Specialized Prompting because the model was heavily influenced by the complexity of the retrieved examples. This influence led to several structural distortions, such as subtle alterations of numerical and range expressions, shifts in the positions or antecedents of modifiers and conditional clauses, the splitting of a single original element into multiple parallel components, and the unnecessary insertion of wherein clauses. Since these distortions are treated as structural deviations from the source text in the evaluation, substantial penalties were applied to the fidelity and antecedent dependency categories.

Based on our analysis, when constraints are imposed on the LLM through prompting, it is difficult for the model used in this study to satisfy those constraints without reducing the overall fidelity of the content, indicating that this remains an important challenge for future work.

5 Conclusion

For patent claim translation using LLMs, we explored three different approaches. Among them, Judge and Refinement successfully improved the evaluation scores under the LLM-as-a-Judge framework. the other two approaches—Specialized Prompting and Few-shot did not show any improvement in the LLM-as-a-Judge evaluation. In the official human evaluation, the comparison between Judge and Refinement and Base Translation showed that the total number of errors increased, and the average score dropped from 86.07 for Base Translation to 81.60 for Judge and Refinement, confirming that human-evaluated quality declined. This result indicates that the performance of the LLM-as-a-Judge framework was not sufficient in this study. The analysis showed that although the three methods improved surface-level quality errors, they also led to an increase in errors related to the fidelity of the original patent claims. When constraints are imposed on the LLM through prompting, it is difficult for the model used in this study to satisfy those constraints without reducing the overall fidelity of the content, making this an important challenge for future work.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP24K15071. These research results were obtained from the commissioned research (No.22501) by National Institute of Information and Communications Technology (NICT), Japan.

References

Haruto Azami, Minato Kondo, Takehito Utsuro, and Masaaki Nagata. 2025. Patent claim translation via continual pre-training of large language models with parallel data. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 300–314, Geneva, Switzerland. European Association for Machine Translation.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

A Prompt for Judge and Refinement

```
You are a professional Japanese→English translator for patent claims.
Follow the full policy below strictly. Translate the single input line into exactly one line of
US-style claim English.
Return only the translation (no notes). Do not add/omit/split/merge content.
```

Figure 1: System Prompt for Base Translation

```
Translate this single Japanese claim line into English:
```

Figure 2: User Prompt for Base Translation

```
You are a meticulous patent-claim reviewer. Evaluate an English translation (PRED) against the
original Japanese claim (JA) with NO reference translation.

Rubric categories (each 0–100; all categories have equal weight):
- fidelity_legal_scope
- us_style_structure
- numbers_units_ranges
- antecedent_dependency
- terminology
- naturalness

Output MUST be valid GitHub-flavored Markdown with the following sections:

Findings (>= min_findings items)
- Bullet list of concrete issues or confirmations (legal style, fidelity,
numbers/units/formulas/ranges, terminology consistency, antecedent basis, dependency,
punctuation/format, naturalness).
- Each item starts with a label like [Fidelity], [Numbers/Units], etc., and quotes the exact
snippet(s) from JA/PRED.

Fix Suggestions
- Bullet list mapping to the Findings, each with a minimally-edited corrected English fragment.
- Preserve scope; do not introduce new elements.

One-line Verdict
- One sentence stating whether the PRED is acceptable for filing.

Subscores
Provide a JSON code block with EXACT keys and integer values 0–100:
'''json
{{ "fidelity_legal_scope": 0,
"us_style_structure": 0,
"numbers_units_ranges": 0,
"antecedent_dependency": 0,
"terminology": 0,
"naturalness": 0
}}'''
```

Figure 3: Prompt for Judge

```
あなたは翻訳品質監査官です。入力は“日本語＋翻訳のみで作成された評価レポート”。
この文書から参照文に依存しない抽象的な『問題タイプ/再発しやすい症状』だけを抽出し、
箇条書き（10〜15項目）で**具体的**に要約してください（『症状 → 期待される形』）。
```

Figure 4: Prompt for Error-pattern Distillation

```
You are a professional patent-claims translator performing a second pass.
Obey the full policy below. You do NOT see any reference translation.
Given the Japanese line, your first-pass English line, and abstract issue types, fix the English
**without adding/removing content**.
Return exactly one English line in US claim style.
```

Figure 5: Prompt for Refinement

B Prompt for Specialized Prompting

```
You are a professional Japanese→English translator specialized in **US-style patent claims**.
Translate **each input line** into **exactly one English claim line**.
Output: **ONLY** the final English claim line (no notes, no bullets, no brackets, no extra spaces).

HARD CONSTRAINTS (must all hold):
- **One sentence** per claim; period at the end; ASCII-only characters.
- US claim formatting: colon after the preamble; **semicolons** between parallel elements; **'; and'** before the last element.
- **Do not add/omit/reorder** content; preserve all numbers, units, symbols, ranges ("X to Y"), inequalities (<=, >=, <, >), equations, and dependencies.
- Maintain claim category (apparatus/method/etc.), numbering, and antecedent basis (first mention "a/an/at least one [X]" → thereafter "the [X]"; keep singular/plural consistent).
- For dependent claims: "The [subject] according to claim X (or X or Y/any one of claims X to Y), wherein . . . ." No new elements introduced in dependents.
- Forbidden: "and/or", non-ASCII dashes (—–~), ambiguous pronouns without antecedent, "respectively" unless explicitly warranted by the JP text.

SILENT QA (do internally and **do not print** the checks):
1) **Category map**: identify claim type; keep it unchanged.
2) **Numbers/units audit**: list every value/unit/range/inequality/equation and verify 1:1 preservation; replace wave dashes with "to"; add a space between number and unit (10 mm); % is attached (10%).
3) **Antecedent map**: ensure every "the [X]" has a prior "a/an [X]" (or "first/second [X]").
4) **Format skeleton**: preamble + colon; element list with semicolons; insert "; and" before the last element; final period.
5) **Terminology lock**: prefer "apparatus" (when appropriate), "configured to", "equal to or greater than/less than or equal to", "idle channel", "suction air temperature", "thermo-OFF/ON", etc., as aligned with the policy below.
```

Figure 6: Prompt for Specialized Base Translation

```
あなたは翻訳品質の監査官です。入力は JA と PRED のみで作られた評価レポートです。
人手評価の得点改善に**直結**する 10〜15 件の「問題タイプ（症状→期待形）」を、重複をまとめて一般化して抽出してください。
優先順位:
1) 忠実性の逸脱（主語/述語/条件/比較/包含/選択/否定/数量/因果）
2) 係り受け・依存関係（antecedent、wherein の接続、要素導入/再登場の不整合）
3) 数値・単位・範囲・不等号・式（ASCII/順序/包含条件/桁区切り/単位スペース）
4) 法的フォーマット（コロン/セミコロン/"; and"/一文制/終止）
※ 表層の言い換えのみは除外。意味/法的効果に影響する項目を優先。

出力形式（例）:
- 【範囲表現】"A〜B" を "A to B" に統一。境界の≦/≧は JA に忠実。
- 【antecedent】"the X" には先行 "a/an/first X" を必須化。再登場での冠詞逸脱を是正。
- 【列挙体裁】パラレル要素はセミコロン列挙＋最後に "; and"。. . .
```

Figure 7: Prompt for Specialized Error-pattern Distillation

```
You are a senior patent-claims translator performing a **targeted second pass** with **no reference translation**.
Objectives (in this order):
1) **Semantic adequacy  legal correctness** (maximize human adequacy judgment).
2) **Minimal-edit policy** to preserve n-grams/phrases of the first-pass English **outside the problematic spans** (helps BLEU and perceived consistency).

HARD CONSTRAINTS:
- Do NOT add/remove meaning vs. Japanese; preserve all numbers, units, symbols, ranges ("X to Y"), inequalities, equations, dependencies, and claim category.
- Enforce US claim style: one sentence; colon after preamble; semicolons between parallel elements; "; and" before the last element; final period; ASCII-only.
- Maintain antecedent basis; attach "wherein" to the correct antecedent; do not introduce new elements in dependent claims.
- **Edit only spans implicated by the "Issues to fix" section**; elsewhere keep tokens identical to the first-pass output unless grammar requires a local micro-fix.
- Avoid "and/or" and non-ASCII dashes; keep spacing for numbers/units; keep thousands separators; "μ"→"um".

SILENT QA BEFORE OUTPUT (do not print): numbers/units audit, antecedent map, format skeleton, and dependency sanity check.
```

Figure 8: Prompt for Specialized Refinement

C Prompt for Few-shot Prompting

```
"あなたは日本語特許請求項の専門家です。以下の日本語1行（1クレーム）について、"
"FAISSで高精度に用例を拾うための『日本語クエリ』を**ちょうど3件**、JSON配列で出力してください。"
"各クエリは（A）中核技術語（専門語・化学名・機械要素・電気回路名など）、（B）構成要素/機能語（～部、～手段、～回路、configured to 等）、"
"（C）決定的な制約（wherein条件、数値レンジ、不等号、単位、選択肢列挙、依存関係）を**過不足なく**含めてください。"
"一般語（装置、処理、データ等）や曖昧語を避け、品詞は名詞句中心で**8～24文字程度**に収めます。"
"括弧・全角記号・機種依存文字は使用しません。"
"出力は**厳密に**次のフォーマットのみ："
"...",
"...",
"..."


"説明や余計な文字は一切付けないでください。"
```

Figure 9: Prompt for generating FAISS search queries

Source Sentence

[請求項2] 前記第１推定モデルは、前記第１ユーザ群のそれぞれのユーザの特徴量と前記第１テーブルデータを構成するそれぞれの項目の特徴量とから前記第１ユーザ群のそれぞれのユーザに対応する前記それぞれの項目の値を推定し、
前記第２推定モデルは、前記共通ユーザ群のそれぞれのユーザの特徴量と前記第２テーブルデータを構成するそれぞれの項目の特徴量とから前記共通ユーザ群のそれぞれのユーザに対応する前記それぞれの項目の値を推定し、
前記推定部は、前記第１推定モデルで利用される前記第１ユーザ群のそれぞれのユーザの特徴量と前記第２推定モデルで利用される前記第２テーブルデータを構成するそれぞれの項目の特徴量とに基づいて、前記共通ユーザ群を除く前記第１ユーザ群についての前記第２データの値を推定する、請求項１に記載の推定装置。

Base Translation

2. The estimation apparatus according to claim 1, wherein: the first estimation model is configured to estimate values of respective items corresponding to each user of the first user group based on feature quantities of each user of the first user group and feature quantities of the respective items constituting the first table data; the second estimation model is configured to estimate values of respective items corresponding to each user of the common user group based on feature quantities of each user of the common user group and feature quantities of the respective items constituting the second table data; and the estimation unit is configured to estimate values of the second data for the first user group excluding the common user group, based on the feature quantities of each user of the first user group used by the first estimation model and the feature quantities of the respective items constituting the second table data used by the second estimation model.

Refinement

2. The estimation apparatus according to claim 1, wherein: the first estimation model is configured to estimate values of the respective items corresponding to each user of the first user group based on features of each user of the first user group and features of the respective items that constitute the first table data; the second estimation model is configured to estimate values of the respective items corresponding to each user of the common user group based on features of each user of the common user group and features of the respective items that constitute the second table data; and the estimation unit is configured to estimate values of the second data for users of the first user group excluding users of the common user group based on (i) the features of each user of the first user group that are used by the first estimation model and (ii) the features of the respective items that constitute the second table data and that are used by the second estimation model.

Table 7: Example Output of Judge and Refinement
Examples in which, although punctuation errors were reduced, hallucinations and omissions occurred.

Source Sentence

[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。

Base Translation

1. A propulsion apparatus comprising: a tubular portion arranged below a water surface such that an axis extends in a fore-aft direction, the tubular portion forming a flow path in which a front side is an upstream side and a rear side is a downstream side; a shaft portion disposed inside the tubular portion and extending in an axial direction; a propeller mounted to the shaft portion, the propeller having a plurality of propeller blades that extend in a radial direction of the axis and are arranged in a circumferential direction of the axis, the propeller being rotatable about the axis; and a plurality of struts provided in the flow path on at least one of a front side and a rear side with respect to the propeller, the plurality of struts extending in the radial direction and being arranged in the circumferential direction, the plurality of struts supporting the shaft portion; wherein: (i) the plurality of struts extend in a vertical direction and are disposed left-right symmetrically with respect to a symmetry line passing through the axis; (ii) at least one angular interval among angular intervals between the struts is different from other angular intervals between the struts; and (iii) with respect to a horizontal plane extending horizontally so as to pass through the axis, at least one of the struts is disposed on each of both sides in a vertical direction.

Refinement

1. A propulsion apparatus comprising: a tubular portion arranged below a water surface such that an axis of the tubular portion extends in a fore-aft direction, the tubular portion forming a flow path in which a front side is an upstream side and a rear side is a downstream side; a shaft portion disposed inside the tubular portion and extending in an axial direction; a propeller mounted to the shaft portion, the propeller having a plurality of propeller blades that extend in a radial direction of the axis and are arranged in a circumferential direction of the axis, the propeller being rotatable about the axis; and a plurality of struts provided in the flow path on at least one of a front side and a rear side of the propeller, the plurality of struts extending in the radial direction and being arranged in the circumferential direction, the plurality of struts being configured to support the shaft portion; wherein: (i) the plurality of struts extend in a vertical direction and are disposed left-right symmetrically with respect to a line of symmetry that passes through the axis; (ii) at least one angular interval between the struts differs from the other angular intervals between the struts; and (iii) with respect to a horizontal plane that passes through the axis, at least one of the struts is disposed on each of an upper side and a lower side.

Table 8: Example Output of Judge and Refinement
Examples of increased hallucinations and omissions.