

Coordinate Structure Alignment of Patent Claims by LLMs

Tsukasa Ishimaru¹, Takehito Utsuro¹, Masaaki Nagata²

¹University of Tsukuba ²NTT, Inc.

s2620739_@_u.tsukuba.ac.jp, utsuro_@_iit.tsukuba.ac.jp

masaaki.nagata_@_ntt.com

Abstract

This paper proposes an LLM-based method for cross-lingual alignment of coordinate structures in Japanese–English and Chinese–English patent claims. The approach has two components. First, we construct task-specific alignment datasets by translating Penn Treebank-3 English sentences into Japanese and Chinese and projecting coordinate-structure annotations using syntactic parse trees, word alignments, and the GHKM algorithm. Second, we formulate alignment as a TANL-based text generation task in which LLMs generate the target sentence with the corresponding span tagged. Although the training data are derived from a general-domain corpus, evaluation is conducted on manually annotated real patent claims. Across four translation directions, the proposed approach consistently outperforms SpanAlign, an encoder-based alignment baseline built on SQuAD-style cross-lingual span prediction and ILP-based alignment. Controlled analyses show that the TANL formulation improves performance independently of supervised fine-tuning and that automatically generated demonstrations achieve performance close to manually constructed examples in the tested 3-shot setting. Supervised fine-tuning with the automatically constructed datasets further improves open-weight LLMs, with the best fine-tuned model in each direction achieving an F1 score above 0.8.

1 Introduction

Coordinate structures are common syntactic constructions in many documents. We define them as sequences of conjuncts within a sentence that stand in a syntactically equivalent relationship. In the 2020 portion of JaParaPat (Nagata et al., 2024), approximately 64% of the 1,035,219 patent claim sentences contain at least one coordinate structure, making them a major source of translation inconsistencies and errors in patent translation.

The ultimate goal of this study is to support post-editing that focuses on maintaining structural consistency between source and translated patent claims. In particular, coordinated expressions are often linguistically complex and difficult to translate consistently across languages. Incorrect translation or alignment of coordinated structures may affect the interpretation of patent claims and lead to downstream legal or commercial risks.

A particular challenge arises when semantically corresponding expressions are realized with different coordinate structures across languages. For example, a Japanese patent claim may express “a pair of sensors” as a single noun phrase, whereas its English counterpart may realize the same content as the coordinate structure “a first sensor and a second sensor.” Such cross-lingual structural divergence, including differences in coordination and nesting, makes it difficult to identify exact span correspondences.

As a fundamental step toward this goal, this paper proposes a method for aligning coordinate structures across languages using large language models (LLMs). Unlike prior work that mainly focuses on word alignment or general span alignment, we specifically address cross-lingual alignment of coordinate structures.

First, we propose a method for constructing a high-quality training dataset for coordinate structure alignment by applying the GHKM (Galley et al., 2004) algorithm to bilingual parallel sentences. The proposed pipeline projects coordinate structure annotations obtained from English syntactic parse trees onto Japanese and Chinese translations using word alignment information, enabling scalable construction of task-specific alignment data. This scalability is particularly important for supervised fine-tuning of open-weight LLMs, which requires substantially more task-specific training examples than few-shot prompting.

Second, we formulate cross-lingual coordinate

structure alignment as a translation task between augmented natural languages using the Translation between Augmented Natural Languages (TANL) (Paolini et al., 2021) framework. Unlike conventional approaches that rely on span prediction or optimization algorithms, our method directly generates the aligned target sentence with tagged spans in a single end-to-end stage. By generating the complete target sentence rather than only an isolated span and its offsets, the proposed formulation allows LLMs to exploit their sequence generation capabilities while retaining the surrounding linguistic context.

We evaluate the proposed approach on manually annotated patent claims in four translation directions: Japanese-to-English (ja→en), English-to-Japanese (en→ja), Chinese-to-English (zh→en), and English-to-Chinese (en→zh). Experimental results show that the proposed method outperforms conventional encoder-based approaches such as SpanAlign (Nagata et al., 2020; Chousa et al., 2020). In particular, fine-tuned open-weight LLMs achieve strong performance despite relatively small parameter sizes, while GPT-5 achieves the best overall performance in a few-shot setting.

Controlled analyses further isolate the contributions of the TANL formulation, automatically generated demonstrations, and supervised fine-tuning.

The main contributions of this paper are summarized as follows:

- We propose a scalable method for constructing Japanese–English and Chinese–English coordinate structure alignment datasets using syntactic parse trees, word alignments, and the GHKM algorithm.
- We formulate cross-lingual coordinate structure alignment as a TANL-based text generation task for LLMs and demonstrate through controlled experiments that the formulation itself improves alignment performance independently of supervised fine-tuning.
- We show that GHKM-generated demonstrations achieve few-shot performance close to manually constructed examples in the tested 3-shot setting, while the same data-construction pipeline provides scalable supervision for adapting open-weight LLMs.
- We evaluate the proposed approach across four Japanese–English and Chinese–English

translation directions on manually annotated real patent claims and consistently outperform the SpanAlign baseline (Nagata et al., 2020; Chousa et al., 2020).

2 Related Work

2.1 Cross-lingual Span Alignment

Nagata et al. (2020) proposed a cross-lingual span prediction method based on multilingual BERT-style encoder models. Their method formulates span prediction as a question-answering task similar to SQuAD (Rajpurkar et al., 2016), where the model predicts the start and end positions of the corresponding span in the target sentence. Building on this span-prediction framework, Chousa et al. (2020) proposed SpanAlign, which combines predicted span correspondences with Integer Linear Programming (ILP) to obtain consistent alignments.

These approaches formulate cross-lingual alignment primarily as span extraction from the target sentence. In contrast, our task focuses specifically on the correspondence of coordinate structure conjuncts, which can involve long spans, nested structures, and substantial syntactic reordering across languages. Rather than predicting only span boundaries, our approach generates the target sentence with explicit tags marking the corresponding span.

2.2 LLM-based Alignment

In addition, Miao et al. (2025) proposed a word alignment method using the LLM. Their approach trains the LLM through semi-supervised learning to obtain high-quality word alignments. This method achieves high alignment accuracy while reducing the dependence on an explicitly annotated parallel dataset.

For inference, their approach assigns indices to tokens in the target sentence and generates the tokens corresponding to a marked source span. They also compare aligning individual units independently with jointly aligning multiple units and report better performance for individual alignment. Following this observation, our method also aligns each coordinate structure conjunct independently. However, instead of generating only the corresponding target tokens, we formulate the task so that the LLM generates the complete target sentence with the aligned span explicitly marked by tags.

Our work therefore differs from previous LLM-based alignment studies in both the target structure and the output formulation: we focus on cross-lingual alignment of coordinate structure conjuncts and formulate their alignment as structured text generation rather than token-level alignment.

2.3 TANL-based Structured Prediction

Paolini et al. (2021) proposed the TANL (Translation between Augmented Natural Languages) framework, which reformulates structured prediction tasks as text-to-text generation problems. In this framework, structured outputs such as named entities, relations, and semantic roles are represented as augmented natural language sequences, enabling a unified treatment of various NLP tasks within a sequence generation paradigm. By converting structured prediction into a natural language generation task, TANL enables pre-trained language models to leverage their generative capabilities without relying on task-specific output architectures.

As an application of TANL, Argument Mining as a Text-to-Text Generation Task applied the framework to argument mining tasks (Kawarada et al., 2024). Their study reformulated argument component detection and relation identification as a text generation problem using augmented natural language representations. Experimental results demonstrated that the TANL-based approach could effectively capture argument structures while maintaining flexibility across subtasks, highlighting the potential of text-to-text generation methods for complex structured prediction tasks.

While these studies demonstrate the effectiveness of TANL for structured prediction, our work applies the framework to cross-lingual coordinate structure alignment using decoder-based LLMs. In our formulation, a tagged source-language conjunct and its parallel target sentence are provided as input, and the model generates the complete target sentence with tags inserted around the corresponding span. This formulation enables coordinate structure alignment to be handled directly as a text generation task without introducing a task-specific span prediction architecture.

3 Dataset Construction using the GHKM Algorithm

To train the proposed alignment model, we propose a method for constructing high-quality Japanese–

English and Chinese–English coordinate structure alignment datasets based on the GHKM (Galley et al., 2004) algorithm. The GHKM algorithm extracts synchronous grammar rules from parallel sentences using syntactic parse trees and word alignments. By deriving minimal tree-to-string translation rules that satisfy alignment constraints, the algorithm captures correspondences between syntactic constituents and their translations.

3.1 Dataset Construction Procedure

The procedure for constructing the Japanese–English and Chinese–English coordinate structure alignment datasets using the GHKM algorithm is as follows. Figure 1 provides an overview of the process.

1. English sentences and their syntactic parse trees are randomly selected from Penn Treebank-3 (Marcus et al., 1993).
2. The English sentences are translated into Japanese and Chinese using GPT-5 to construct parallel sentence pairs.
3. Word alignment information between the English sentences and the translated sentences is obtained using WSPAlign (Wu et al., 2023).
4. Coordinate structure spans are extracted from the English parse trees, and special tags are inserted into the corresponding spans in the English sentences.
5. Using the syntactic parse trees and word alignment information, the tagged coordinate structure spans are projected onto the Japanese and Chinese translations based on the minimal translation units defined by the GHKM algorithm.
6. Finally, the same tags are inserted into the corresponding spans in the Japanese and Chinese sentences, resulting in multilingual coordinate structure alignment pairs.

3.2 Dataset Construction Rationale

To construct the training data, we require parallel sentences, word alignments, and syntactic parse trees. To our knowledge, no existing dataset simultaneously provides all three with manual annotations. We therefore use Penn Treebank-3, which

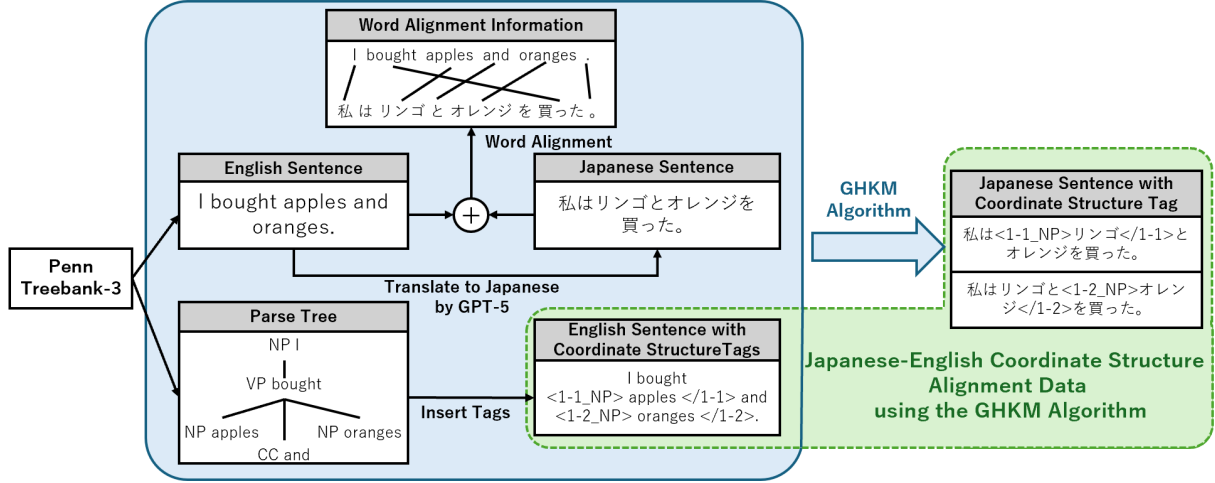


Figure 1: Dataset Construction using the GHKM Algorithm (The Case of ja→en)

provides gold English syntactic parse trees, translate the English sentences into Japanese and Chinese using GPT-5, and obtain word alignments using WSPAlign. Potential biases and distributional differences introduced by this construction procedure are discussed in Section 6.

4 LLM-based Alignment of Coordinate Structures using TANL

4.1 Alignment Formulation

To perform Japanese–English and Chinese–English coordinate structure alignment, we propose an end-to-end single-stage method using LLMs within the TANL (Paolini et al., 2021) framework. TANL is a framework that formulates NLP tasks as translation problems between augmented natural languages, where structural information or tags are embedded directly into natural language.

For each alignment instance, one coordinate structure conjunct in the source sentence is marked with a tag, while the corresponding target sentence is provided without tags. The LLM is then required to reproduce the complete target sentence and insert the same tag around the span corresponding to the marked source conjunct. Thus, alignment is performed through direct generation of an augmented target sentence rather than explicit prediction of token indices or span boundaries.

The alignment task is formulated using the following input-output format:

- Input: A source-language sentence in which one coordinate structure conjunct is marked with tags, together with its parallel target-language sen-

tence without tags.

- Output: The complete target-language sentence with the corresponding span marked using the same tags.

The same formulation is applied to all four translation directions considered in this study: ja→en, en→ja, zh→en, and en→zh. An overview of the alignment procedure for ja→en is shown in Figure 2.

4.2 Independent Conjunct Alignment

In our method, each conjunct is aligned independently rather than jointly predicting all conjuncts simultaneously.

For a coordinate structure containing multiple conjuncts, the same parallel sentence pair is processed separately for each conjunct, with only the conjunct to be aligned marked in the source sentence. The model therefore predicts one source-target conjunct correspondence at a time. This design decomposes the alignment of a complex coordinate structure into a set of simpler span correspondence problems while retaining the full sentential context in both the input and output.

5 Experimental Setup

5.1 Proprietary LLMs (Few-Shot Setting)

To apply the proposed method, we perform Japanese–English and Chinese–English coordinate structure alignment using LLMs within the TANL framework with proprietary models. To leverage the strong syntactic understanding and generation capabilities of recent LLMs, we conduct experi-

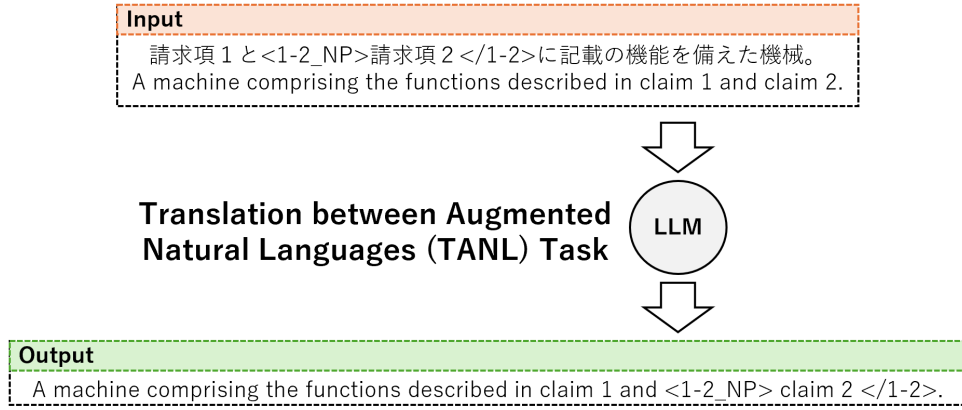


Figure 2: LLM-based Alignment of Coordinate Structures Using TANL (The Case of ja→en)

ments in a 3-shot setting. We use two state-of-the-art proprietary models: Gemini-2.5-Pro¹ and GPT-5², which are accessed via their respective APIs using API keys. The prompts and model parameters are described in Appendix C.

The three in-context demonstrations used in the main experiments are automatically constructed using the GHKM-based pipeline described in Section 3.

5.2 Open-Weight LLMs (Supervised Fine-Tuning)

To optimize the proposed method, we perform supervised fine-tuning (SFT) of the LLM and apply LoRA (Hu et al., 2022) to reduce GPU memory consumption.

For the Japanese–English alignment experiments, we use three open-weight multilingual LLMs with relatively small numbers of parameters as base models: gemma-3-4b-it³, Llama-3.1-8B-Instruct⁴, and llama-3-youko-8b⁵.

For the Chinese–English alignment experiments, we adopt gemma-3-4b-it, Llama-3.1-8B-Instruct, and Qwen3-8B⁶ as the base models.

The training datasets consist of 5,906 Japanese–English and 7,118 Chinese–English coordinate structure alignment sentence pairs constructed in Section 3. The statistics of the dataset are shown in Appendix A. Since we do not perform hyperparameter tuning, we use the entire dataset for supervised fine-tuning and evaluate the models on a manually

constructed test set. The detailed training settings are described in Appendix D.

5.3 Baseline

We use SpanAlign (Nagata et al., 2020; Chousa et al., 2020)⁷ as the baseline. We train its XLM-RoBERTa-based span predictor (Conneau et al., 2020) using the same training data as in Section 5.2, converted into SQuAD 2.0 JSON format, and apply the original ILP-based alignment procedure.

5.4 Evaluation Datasets and Metrics

In this experiment, we construct a Japanese–English evaluation dataset by manually inserting coordinate structure tags into 200 Japanese–English parallel patent claim sentences containing coordinate structures from the 2020 portion of the JaParaPat (Nagata et al., 2024) corpus. In addition, we manually construct a Chinese–English evaluation dataset consisting of 200 parallel patent claim sentences from the 2006–2012 range using DOCDB patent family information provided by the European Patent Office (European Patent Office, 2026).

The inter-annotator agreement for the Japanese–English evaluation dataset achieves an F1 score of 0.9293.

To evaluate the alignment component independently of errors in upstream coordinate structure extraction, we assume that source-side coordinate structure spans are correctly identified and use manually annotated spans as input.

We evaluate alignment accuracy using span-level precision, recall, and F1 after removing the inserted tags. The metric evaluates whether the predicted target span matches the manually annotated structural correspondence; it does not by itself guarantee

¹<https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro>

²<https://openai.com/ja-JP/gpt-5/>

³<https://huggingface.co/google/gemma-3-4b-it>

⁴<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

⁵<https://huggingface.co/rinna/llama-3-youko-8b>

⁶<https://huggingface.co/Qwen/Qwen3-8B>

⁷<https://github.com/nttcs-lab-nlp/spanalign>

preservation of semantic scope or complete semantic equivalence.

Differences from SpanAlign were considered statistically significant using an approximate randomization test (Yeh, 2000). The statistics of the dataset are shown in Appendix A.

6 Discussion

6.1 Overall Alignment Performance

Table 1 shows the precision, recall, and F1 scores for coordinate structure alignment across four translation directions: Japanese-to-English (ja→en), English-to-Japanese (en→ja), Chinese-to-English (zh→en), and English-to-Chinese (en→zh).

Among the proprietary models, GPT-5 achieves the best performance in ja→en, zh→en, and en→zh, while Gemini-2.5-Pro achieves the best result in en→ja. GPT-5 exceeds an F1 score of 0.93 in all four directions, while Gemini-2.5-Pro exceeds 0.92 in three of the four directions.

Among the SFT-based open-weight models, Llama-3.1-8B-Instruct achieves the highest F1 score in all four translation directions, reaching 0.8557, 0.8223, 0.8138, and 0.9173, respectively. All LLM-based settings shown in Table 1 outperform SpanAlign, and the differences are statistically significant at $p < 0.01$.

These main results demonstrate strong performance of the proposed framework, but they do not by themselves distinguish the effects of the TANL formulation, automatically constructed data, and supervised fine-tuning. We therefore conduct controlled analyses below.

6.2 Effect of Supervised Fine-Tuning and Practical Value of Open-Weight LLMs

To analyze the contribution of supervised adaptation, we additionally compare non-fine-tuned base models and fine-tuned models.

Table 2 shows that SFT substantially improves alignment performance across models and translation directions. To isolate this effect while keeping the base model and TANL formulation fixed, Llama-3.1-8B-Instruct improves from 0.5707 to 0.8557 for ja→en, from 0.4252 to 0.8223 for en→ja, from 0.5075 to 0.8138 for zh→en, and from 0.6014 to 0.9173 for en→zh. The improvements range from 0.2850 to 0.3971 F1.

These results show that the automatically constructed task-specific data provide effective supervision for adapting open-weight LLMs. This is

practically important for patent post-editing, where confidentiality, data-governance requirements, and operational constraints may restrict sending patent claims to external proprietary services. In such settings, locally deployable open-weight models are attractive, and SFT using the constructed dataset substantially improves their alignment accuracy. We therefore view proprietary and open-weight LLMs as complementary: proprietary LLMs provide strong few-shot performance and scalable synthetic supervision, while the resulting data can be used to adapt models that can be deployed locally.

6.3 Controlled Analysis of the Proposed Components

To disentangle the contributions of the TANL formulation, GHKM-based demonstrations, and supervised adaptation, we conduct controlled comparisons while keeping the remaining experimental conditions fixed. Table 3 summarizes the average F1 scores over the four translation directions. The contribution of supervised fine-tuning is discussed in Section 6.2, while the following analyses separately examine the TANL formulation and the use and quality of in-context demonstrations. Detailed per-direction results are provided in Appendix B.

6.3.1 Effect of the TANL Formulation

First, we evaluate the contribution of the TANL formulation without supervised fine-tuning. The models, demonstrations, and evaluation data are kept fixed between the two conditions. In the non-TANL condition, the model outputs only the aligned target span and its character offsets in JSON format, whereas TANL generates the complete target sentence with tags around the aligned span.

A concrete example of the TANL and non-TANL output formats is provided in Appendix B (Figure 3). TANL improves F1 in all eight model-direction combinations, with average gains of 0.1560 for GPT-5 and 0.0989 for Llama-3.1-8B-Instruct. Because the models, demonstrations, evaluation data, and absence of SFT are fixed, this comparison isolates the contribution of the output formulation itself.

6.3.2 Effect of In-Context Demonstrations

Second, GPT-5 achieves an average F1 of 0.8388 in the zero-shot setting and 0.9352 with three GHKM-generated demonstrations, corresponding to an average improvement of 0.0965. The 3-shot setting improves performance in all four directions, indi-

Direction	Model	Method	Precision	Recall	F1
ja→en	SpanAlign(Base)	–	0.7363	0.7363	0.7363
	Gemini-2.5-Pro	3-shot	0.8700	0.8657	0.8678**
	GPT-5	3-shot	0.9353	0.9353	0.9353**
	gemma-3-4b-it	SFT (LoRA)	0.7828	0.7711	0.7769**
	Llama-3.1-8B-Instruct	SFT (LoRA)	0.8711	0.8408	0.8557**
	llama-3-youko-8b	SFT (LoRA)	0.8564	0.8308	0.8434**
en→ja	SpanAlign(Base)	–	0.7682	0.7030	0.7341
	Gemini-2.5-Pro	3-shot	0.9546	0.9400	0.9473**
	GPT-5	3-shot	0.9400	0.9400	0.9400**
	gemma-3-4b-it	SFT (LoRA)	0.7801	0.7450	0.7621**
	Llama-3.1-8B-Instruct	SFT (LoRA)	0.8351	0.8100	0.8223**
	llama-3-youko-8b	SFT (LoRA)	0.8325	0.7950	0.8133**
zh→en	SpanAlign(Base)	–	0.8385	0.6089	0.7055
	Gemini-2.5-Pro	3-shot	0.9422	0.9023	0.9218**
	GPT-5	3-shot	0.9372	0.9275	0.9323**
	gemma-3-4b-it	SFT (LoRA)	0.7947	0.7824	0.7885**
	Llama-3.1-8B-Instruct	SFT (LoRA)	0.8361	0.7927	0.8138**
	Qwen3-8B	SFT (LoRA)	0.8242	0.7772	0.8000**
en→zh	SpanAlign(Base)	–	0.7636	0.7826	0.7730
	Gemini-2.5-Pro	3-shot	0.9075	0.9581	0.9321**
	GPT-5	3-shot	0.9115	0.9563	0.9333**
	gemma-3-4b-it	SFT (LoRA)	0.8351	0.8579	0.8464**
	Llama-3.1-8B-Instruct	SFT (LoRA)	0.8958	0.9399	0.9173**
	Qwen3-8B	SFT (LoRA)	0.8848	0.9235	0.9037**

Table 1: Coordinate Structure Alignment Accuracy. ** indicates statistically significant differences ($p < 0.01$) compared to SpanAlign.

cating that the automatically generated demonstrations provide useful in-context supervision beyond the model’s intrinsic zero-shot capability.

6.3.3 Quality of GHKM-Generated Demonstrations

Third, we compare three manually constructed demonstrations with three demonstrations generated by the GHKM-based pipeline. Neither demonstration source consistently outperforms the other. The average absolute F1 differences are only 0.0095 for GPT-5 and 0.0156 for Llama-3.1-8B-Instruct, providing preliminary evidence that the automatically generated demonstrations are of sufficient quality for few-shot prompting. This experiment is limited to three demonstrations and does not establish general equivalence between manually and automatically constructed datasets.

Taken together, these controlled experiments distinguish the roles of the proposed components: the TANL formulation itself improves alignment performance, GHKM-generated examples provide ef-

fective few-shot demonstrations, and large-scale automatically constructed data enable substantial additional gains through supervised adaptation of open-weight LLMs.

6.4 Dataset Bias and Structural Mismatch

The proposed data-construction procedure introduces two important considerations. First, GPT-5 is used both to translate the training data and as one of the evaluated alignment models, which may introduce model-specific bias. The manual-versus-GHKM demonstration comparison in Section 6.3.3 shows only small differences, suggesting that GPT-5’s strong few-shot performance is not solely explained by GPT-5-generated demonstrations. However, this limited 3-shot analysis cannot rule out bias in the full automatically constructed training dataset.

Second, the training and evaluation data differ substantially in structural complexity. For the training data, nesting is identified from the corresponding English Penn Treebank-3 parse trees because

Direction	Model	Method	Precision	Recall	F1
ja→en	gemma-3-4b-it	Base (3-shot)	0.5780	0.4975	0.5348
		SFT (LoRA)	0.7828	0.7711	0.7769
	Llama-3.1-8B-Instruct	Base (3-shot)	0.5904	0.5522	0.5707
		SFT (LoRA)	0.8711	0.8408	0.8557
	llama-3-youko-8b	Base (3-shot)	0.7778	0.1393	0.2363
		SFT (LoRA)	0.8564	0.8308	0.8434
en→ja	gemma-3-4b-it	Base (3-shot)	0.2741	0.1850	0.2209
		SFT (LoRA)	0.7801	0.7450	0.7621
	Llama-3.1-8B-Instruct	Base (3-shot)	0.3624	0.5143	0.4252
		SFT (LoRA)	0.8351	0.8100	0.8223
	llama-3-youko-8b	Base (3-shot)	0.2566	0.2500	0.2533
		SFT (LoRA)	0.8325	0.7950	0.8133
zh→en	gemma-3-4b-it	Base (3-shot)	0.3363	0.3040	0.3193
		SFT (LoRA)	0.7947	0.7824	0.7885
	Llama-3.1-8B-Instruct	Base (3-shot)	0.5484	0.4722	0.5075
		SFT (LoRA)	0.8361	0.7927	0.8138
	Qwen3-8B	Base (3-shot)	0.5000	0.3793	0.4314
		SFT (LoRA)	0.8242	0.7772	0.8000
en→zh	gemma-3-4b-it	Base (3-shot)	0.3644	0.3440	0.3539
		SFT (LoRA)	0.8351	0.8579	0.8464
	Llama-3.1-8B-Instruct	Base (3-shot)	0.5513	0.6615	0.6014
		SFT (LoRA)	0.8958	0.9399	0.9173
	Qwen3-8B	Base (3-shot)	0.5034	0.5827	0.5401
		SFT (LoRA)	0.8848	0.9235	0.9037

Table 2: Comparison between base prompting and supervised fine-tuning (SFT) for coordinate structure alignment.

the machine-translated Japanese and Chinese sentences do not have gold syntactic parse trees. Under this definition, approximately 6% of the training instances contain nested coordinate structures, whereas the proportions in the manually annotated evaluation data range from 52.5% to 78.5%, depending on the language side and dataset. Thus, the distribution shift involves not only domain differences but also substantial structural-complexity mismatch. Detailed statistics and a representative cross-lingual nesting example are provided in Appendix A. A controlled comparison of alignment performance on nested and non-nested subsets remains future work.

7 Error Analysis

The experimental results reveal several characteristic error patterns in multilingual coordinate structure alignment. Although the proposed TANL-based approach substantially improves performance compared with SpanAlign and achieves strong results after supervised fine-tuning, certain

types of alignment errors remain difficult across all language directions.

7.1 Long and Structurally Complex Coordinated Structures

One recurring error pattern involves long and structurally complex coordinated conjuncts. Alignment errors tend to increase when the source coordinated phrases become longer and contain nested modifiers or complex internal syntactic structures. Long conjuncts frequently undergo reordering or syntactic restructuring during translation, making it difficult to identify an exactly corresponding contiguous span in the target sentence. As a result, predictions may cover only part of the gold span or include surrounding modifiers that are not part of the annotated correspondence.

As discussed in Section 6, the substantial structural-complexity mismatch between the training and evaluation data may partly contribute to these errors.

Analysis	Model	Setting A	Setting B	Δ F1
TANL formulation	GPT-5	Non-TANL: 0.7793	TANL: 0.9352	+0.1560
TANL formulation	Llama-3.1-8B-Instruct	Non-TANL: 0.4273	TANL: 0.5262	+0.0989
In-context demonstrations	GPT-5	Zero-shot: 0.8388	GHKM 3-shot: 0.9352	+0.0965
Demonstration source	GPT-5	Manual: 0.9446	GHKM: 0.9352	-0.0094
Demonstration source	Llama-3.1-8B-Instruct	Manual: 0.5359	GHKM: 0.5262	-0.0097
Supervised adaptation	Llama-3.1-8B-Instruct	Base: 0.5262	SFT: 0.8523	+0.3261

Table 3: Controlled analysis of the proposed components. Scores are F1 values averaged over the four translation directions. For the demonstration-source analysis, Δ F1 denotes the difference between the average F1 scores of the two settings, whereas the average absolute per-direction differences are reported in the text.

7.2 Errors in Modifier Boundary Identification

Another recurring error pattern involves incorrect identification of modifier boundaries. In many patent claims, modifiers are attached to coordinated noun phrases in a highly ambiguous manner, making it difficult to determine the exact span boundaries of conjuncts. As a result, the predicted alignments often partially overlap with the gold annotations while failing to exactly match the annotated spans. Such errors are observed in technical patent expressions containing long nominal modifiers and nested coordination structures.

Such boundary errors are important because modifiers may have different scopes across the source and target languages. Even when a predicted span has substantial surface overlap with the gold annotation, the prediction does not necessarily guarantee that the semantic scope of the modifier is preserved. Our span-level evaluation therefore measures structural correspondence rather than complete semantic equivalence. A more detailed semantic analysis of modifier scope remains future work.

8 Conclusion

In this paper, we proposed an LLM-based method for cross-lingual coordinate structure alignment to support post-editing in patent translation. We constructed Japanese–English and Chinese–English alignment datasets using the GHKM algorithm and formulated the task within the TANL framework as a translation problem between augmented natural languages.

Experimental results on four translation directions showed that the proposed LLM-based approach consistently outperformed SpanAlign. Proprietary LLMs achieved strong performance with only a small number of demonstrations, while

supervised fine-tuning using the automatically constructed datasets enabled locally deployable open-weight LLMs to achieve competitive alignment accuracy, which is particularly relevant to confidentiality-sensitive patent workflows.

Controlled analyses further disentangled the contributions of the proposed components. The TANL formulation improved alignment performance even without supervised fine-tuning, indicating that generating the complete target sentence with explicitly tagged spans is itself beneficial. In addition, GHKM-generated demonstrations achieved few-shot performance close to manually constructed demonstrations, while supervised fine-tuning provided substantial additional gains for open-weight LLMs.

These results demonstrate that scalable GHKM-based data construction and TANL-based structured generation provide complementary mechanisms for multilingual coordinate structure alignment. The proposed framework offers a practical basis for supporting structural consistency verification in patent translation, while further analysis of structurally complex cases and semantic equivalence remains important future work.

Limitations

Limited Size of the Evaluation Dataset

Each of the Japanese–English and Chinese–English evaluation sets contains 200 manually annotated patent claim sentence pairs. Although the Japanese–English dataset achieves an inter-annotator agreement F1 score of 0.9293 and provides a useful initial benchmark, 200 sentence pairs per language pair are still limited for comprehensively evaluating LLM-based alignment methods. Constructing larger and more diverse multilingual evaluation benchmarks remains important future work.

Domain Restriction to Patent Claims

Another limitation is that the experiments focus only on patent claims. Patent claims have distinctive syntactic characteristics, such as long sentences and frequent coordinate structures. Therefore, the effectiveness of the proposed method for other domains remains unclear. However, coordinated structures are also highly frequent in other high-value domains such as legal contracts, regulations, and technical specifications. Therefore, we believe the task has practical relevance beyond patent claims.

Synthetic Training Data and Model-Induced Bias

The Japanese and Chinese sentences in the constructed training datasets are generated by GPT-5 from English sentences, and their word alignments are automatically obtained using WSPAlign. Although this pipeline enables scalable data construction, automatically generated translations and alignments may contain systematic artifacts that differ from naturally occurring bilingual texts.

A further concern is that GPT-5 is used both to generate the translated training data and as one of the models evaluated in our experiments, which may introduce model-specific bias. Our controlled comparison between manually constructed and GHKM-generated demonstrations shows only small average absolute F1 differences (0.0095 for GPT-5 and 0.0156 for Llama-3.1-8B-Instruct), suggesting that GPT-5’s strong few-shot performance is not solely explained by the automatically generated demonstrations. However, this analysis is limited to three in-context examples and cannot rule out model-specific bias in the full automatically constructed dataset. Constructing comparable datasets using other LLMs and independent ma-

chine translation systems remains important future work.

Limitations of Surface Span Evaluation

Our evaluation is based primarily on surface span correspondence. The goal of this study is to identify structurally corresponding coordinate elements across languages, and accurate span alignment is an important intermediate step toward structural consistency verification. However, an exact or highly overlapping span match does not necessarily guarantee preservation of semantic scope or semantic roles, particularly when coordination interacts with modifiers or undergoes substantial syntactic reordering during translation.

Therefore, the reported span-level F1 scores should be interpreted as measures of structural correspondence rather than complete semantic equivalence. A more detailed manual evaluation of whether aligned structures preserve modifier scope and semantic relations is left for future work.

Potential Risks

Our approach may pose several risks when applied in real-world settings. Misalignment of coordinate structures could result in incorrect interpretation of patent claims, potentially affecting legal or commercial decisions if used without expert validation. In addition, LLM-based predictions may produce hallucinated or inconsistent tags, particularly for long or syntactically complex spans. Furthermore, while the method is designed to assist human analysis, there is a risk of overreliance on automated outputs, which could amplify the impact of prediction errors in downstream tasks.

Errors in the proposed method may lead to financial loss or reputational damage. Furthermore, employing a system that may produce incorrect outputs as support for human tasks could reduce work efficiency and increase cognitive load, particularly when users must identify and correct errors.

References

- Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. 2020. [SpanAlign: Sentence Alignment Method based on Cross-Language Span Prediction and ILP](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4750–4761, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

- Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised Cross-lingual Representation Learning at Scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- European Patent Office. 2026. [EPO Worldwide Bibliographic Data \(DOCDB\)](#). Accessed: 2026-05-23.
- Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. [What’s in a translation rule?](#) In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 273–280, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-Rank Adaptation of Large Language Models](#). In *International Conference on Learning Representations*.
- Masayuki Kawarada, Tsutomu Hirao, Wataru Uchida, and Masaaki Nagata. 2024. [Argument Mining as a Text-to-Text Generation Task](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2002–2014, St. Julian’s, Malta. Association for Computational Linguistics.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a Large Annotated Corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- Zhongtao Miao, Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2025. [Improving Word Alignment Using Semi-Supervised Learning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19871–19888, Vienna, Austria. Association for Computational Linguistics.
- Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. [A Supervised Word Alignment Method based on Cross-Language Span Prediction using Multilingual BERT](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online. Association for Computational Linguistics.
- 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.
- Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. [Structured Prediction as Translation between Augmented Natural Languages](#). In *International Conference on Learning Representations*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ Questions for Machine Comprehension of Text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2023. [WSPAlign: Word Alignment Pre-training via Large-Scale Weakly Supervised Span Prediction](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11084–11099, Toronto, Canada. Association for Computational Linguistics.
- Alexander Yeh. 2000. [More accurate tests for the statistical significance of result differences](#). In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

A Details of Datasets

Lang	Dataset	# Sent. Pairs	Conj. Length Avg.(Max) tokens	# Conj. per CS Avg.(Max)
ja	Train	5,906	5.2 (37)	2.6 (5)
	Eval	200	21.9 (70)	6.0 (16)
zh	Train	7,118	4.9 (44)	2.2 (10)
	Eval	200	6.6 (41)	4.8 (10)
en(-ja)	Train	5,906	3.8 (27)	2.0 (8)
	Eval	200	8.8 (72)	4.4 (12)
en(-zh)	Train	7,118	4.0 (37)	1.8 (11)
	Eval	200	15.5 (41)	4.7 (13)

Table 4: Statistics of the Training and Evaluation Datasets

Table 4 shows the statistics of the training and evaluation datasets used in our experiments. The table reports the number of sentence pairs, the average and maximum length of conjuncts, and the average number of conjuncts contained in each coordinate structure. The length of conjuncts is measured in tokens. Since the number of tokens varies depending on the tokenizer used by each model, we compute token lengths using the tokenizer of Llama-3.1-8B-Instruct.

A.1 Nested Coordinate Structures

We define a nested coordinate structure as a coordinate structure containing another coordinate structure within one of its conjuncts or aligned spans. Table 5 shows that nested structures are substantially more frequent in the evaluation datasets than in the training datasets. For the Japanese and Chinese training sentences, nesting counts are derived from the corresponding English Penn Treebank-3

Lang	Dataset	# Pairs	# Nested	Ratio
ja	Train	5,906	372	6.3%
ja	Eval	200	105	52.5%
en(-ja)	Train	5,906	372	6.3%
en(-ja)	Eval	200	110	55.0%
zh	Train	7,118	457	6.4%
zh	Eval	200	144	72.0%
en(-zh)	Train	7,118	457	6.4%
en(-zh)	Eval	200	157	78.5%

Table 5: Statistics of sentences containing nested coordinate structures.

parse trees because the translated sentences do not have gold syntactic parse trees.

Equivalent meanings can also be realized using different nesting structures across languages. For example:

Japanese:

<1-1_NP>送信部</1-1>と、<1-2_NP>一対のセンサに接続された制御部</1-2>とを備える通信装置。

English:

A communication apparatus comprising <1-1_NP>a transmitter </1-1> and <1-2_NP>a controller connected to <2-1_NP>a first sensor </2-1> and <2-2_NP>a second sensor </2-2> </1-2>.

In this example, Japanese expresses the sensors as the single phrase “a pair of sensors,” whereas English realizes the same content as the coordination “a first sensor and a second sensor.” Consequently, the two sides contain different numbers of nested coordinate structures despite expressing corresponding content.

B Detailed Results of Controlled Analyses

Tables 6–8 provide the per-direction results of the controlled analyses summarized in Table 3.

Table 6 compares the TANL and non-TANL output formulations under the same 3-shot setting. TANL improves F1 in all eight model–direction combinations. The average gains over the four translation directions are 0.1560 for GPT-5 and 0.0989 for Llama-3.1-8B-Instruct, indicating that the TANL formulation itself contributes to alignment performance independently of supervised fine-tuning.

Figure 3 illustrates the difference between the TANL and non-TANL output formulations using the same Japanese–English alignment instance. In the TANL condition, the model generates the complete target sentence with tags around the aligned span. In the non-TANL condition, the model in-

stead outputs only the corresponding target span and its character offsets in JSON format.

Table 7 compares zero-shot and 3-shot alignment using GPT-5. The GHKM-based 3-shot setting improves F1 in all four translation directions, with improvements ranging from 0.0250 to 0.1595 and an average improvement of 0.0965. These results show that the automatically generated demonstrations provide useful in-context information beyond zero-shot prompting.

Table 8 compares manually constructed and GHKM-generated demonstrations while keeping the model, output formulation, number of demonstrations, and evaluation data fixed. Neither demonstration type consistently outperforms the other. The average absolute F1 differences are 0.0095 for GPT-5 and 0.0156 for Llama-3.1-8B-Instruct. This provides preliminary evidence that the automatically generated demonstrations are of sufficient quality for few-shot prompting, although the comparison is limited to three demonstrations.

C Configuration for Proprietary LLMs

We evaluate two proprietary large language models, GPT-5 and Gemini-2.5-Pro, using the prompt shown in Figure 5. Both models are provided with the same instructions and 3-shot examples.

GPT-5 GPT-5 is accessed through the OpenAI Chat Completions API. The model receives a system prompt followed by 3-shot examples and the evaluation input. The decoding parameters are as follows. Model: gpt-5, API: OpenAI Chat Completions API, Maximum completion tokens: 4096, Temperature: default.

Gemini-2.5-Pro Gemini-2.5-Pro is accessed through the Google GenAI API. The model receives the same prompt and few-shot examples as GPT-5. The decoding parameters are as follows. Model: gemini-2.5-pro, API: Google GenAI API, Temperature: 0.0.

D Configuration for Fine-Tuning Open-Weight LLMs

We fine-tune the open-weight large language models using Supervised Fine-Tuning (SFT) with Low-Rank Adaptation (LoRA). Training is conducted using the SFTTrainer provided by the TRL library. The model is loaded using the utilities provided by Unsloth, which enable efficient fine-tuning of large language models with reduced memory usage.

Model	Direction	Non-TANL	TANL	Δ F1
GPT-5	ja $\rightarrow$ en	0.8150	0.9353	+0.1203
	en $\rightarrow$ ja	0.8010	0.9400	+0.1390
	zh $\rightarrow$ en	0.6842	0.9323	+0.2481
	en $\rightarrow$ zh	0.8168	0.9333	+0.1165
Llama-3.1-8B-Instruct	ja $\rightarrow$ en	0.4748	0.5707	+0.0959
	en $\rightarrow$ ja	0.4198	0.4252	+0.0054
	zh $\rightarrow$ en	0.3479	0.5075	+0.1596
	en $\rightarrow$ zh	0.4666	0.6014	+0.1348

Table 6: Per-direction F1 comparison between non-TANL and TANL output formulations under the same 3-shot setting.

Direction	Zero-shot			3-shot			Δ F1
	P	R	F1	P	R	F1	
ja $\rightarrow$ en	0.8872	0.8026	0.8428	0.9353	0.9353	0.9353	+0.0925
en $\rightarrow$ ja	0.9150	0.9150	0.9150	0.9400	0.9400	0.9400	+0.0250
zh $\rightarrow$ en	0.7789	0.7668	0.7728	0.9372	0.9275	0.9323	+0.1595
en $\rightarrow$ zh	0.8031	0.8470	0.8245	0.9115	0.9563	0.9333	+0.1088

Table 7: Zero-shot and 3-shot alignment results for GPT-5. The 3-shot demonstrations are generated using the GHKM-based pipeline.

D.1 Training Data Format

The input prompt follows a role-based instruction format consisting of a system instruction, a user instruction, and the assistant response prefix. The prompt is shown in Figure 4.

D.2 Training Objective

During training, the prompt tokens are masked from the loss computation so that the model learns only to generate the complete target-language sentence with the corresponding alignment tags. Specifically, tokens corresponding to the prompt are assigned a label value of -100 , following the standard causal language modeling training procedure.

D.3 Parameter-Efficient Fine-Tuning

We apply LoRA to the transformer layers of the model. The LoRA configuration is applied to the following projection modules: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj.

The LoRA rank is set to 32, and the scaling factor (α) is also set to 32.

D.4 Training Configuration

The model is trained with the following hyperparameters: batch size: 2, gradient accumulation steps: 4, number of epochs: 1, learning rate: 3×10^{-4} , optimizer: AdamW (8-bit), weight decay:

0.01, learning rate scheduler: linear, warmup steps: 5.

Mixed precision training is enabled using either FP16 or BF16, depending on hardware support. The maximum input sequence length is set to 2048 tokens.

Model	Direction	Manual	GHKM-Based	Δ F1
GPT-5	ja→en	0.9352	0.9353	+0.0001
	en→ja	0.9724	0.9400	-0.0324
	zh→en	0.9347	0.9323	-0.0024
	en→zh	0.9362	0.9333	-0.0029
Llama-3.1-8B-Instruct	ja→en	0.5865	0.5707	-0.0158
	en→ja	0.4252	0.4252	0.0000
	zh→en	0.4957	0.5075	+0.0118
	en→zh	0.6361	0.6014	-0.0347

Table 8: F1 comparison between manually constructed and GHKM-generated 3-shot demonstrations.

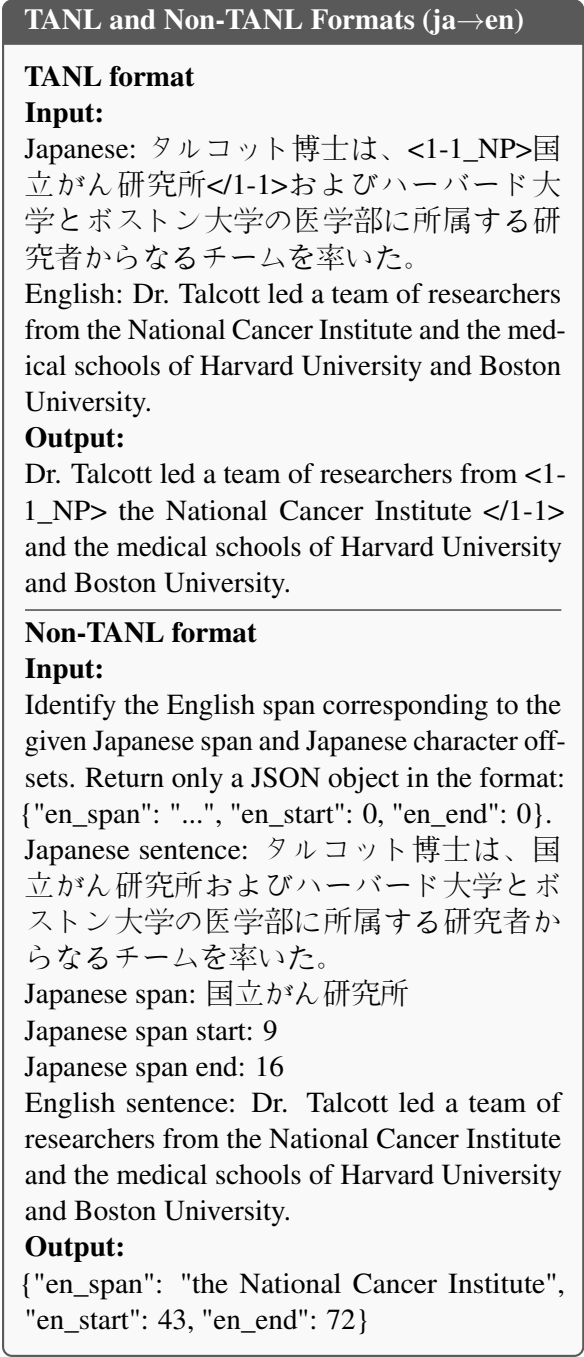


Figure 3: Example of the TANL and non-TANL output formulations for the same Japanese–English alignment instance.

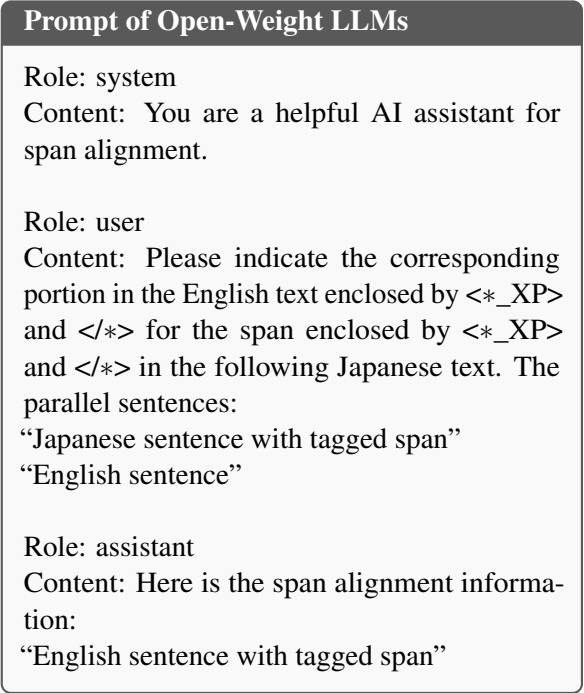


Figure 4: Prompt of Fine-tuned LLMs (The Case of ja→en)

3-Shot Prompt of Proprietary LLMs

role: system

content: You are a helpful AI assistant for span alignment.

role: user

content: Please indicate the corresponding portion in the English text enclosed by <*_XP> and </*> for the span enclosed by <*_XP> and </*> in the following Japanese text. The parallel sentences:

タルコット博士は、<1-1_NP>国立がん研究所</1-1>およびハーバード大学とボストン大学の医学部に所属する研究者からなるチームを率いた。

Dr. Talcott led a team of researchers from the National Cancer Institute and the medical schools of Harvard University and Boston University.

role: assistant

content: Dr. Talcott led a team of researchers from <1-1_NP> the National Cancer Institute </1-1> and the medical schools of Harvard University and Boston University.

role: user

content: Please indicate the corresponding portion in the English text enclosed by <*_XP> and </*> for the span enclosed by <*_XP> and </*> in the following Japanese text. The parallel sentences:

ロリラード社の広報担当者は、1950年代初頭にはフィルター用の紙の製造に石綿が「ごくわずかな量」用いられており、<1-2_VP>1956年には別の種類のフィルターに置き換え</1-2>られたと述べた。

The Lorillard spokeswoman said asbestos was used in “very modest amounts” in making paper for the filters in the early 1950s and replaced with a different type of filter in 1956.

role: assistant

content: The Lorillard spokeswoman said asbestos was used in “very modest amounts” in making paper for the filters in the early 1950s and <1-2_VP> replaced with a different type of filter in 1956 </1-2>.

role: user

content: Please indicate the corresponding portion in the English text enclosed by <*_XP> and </*> for the span enclosed by <*_XP> and </*> in the following Japanese text. The parallel sentences:

<1-1_NP>悪性中皮腫</1-1>、肺がん、石綿肺による死亡者数の合計は18人で、予想よりはるかに多かったと研究者らは述べた。

The total of 18 deaths from malignant mesothelioma, lung cancer and asbestosis was far higher than expected, the researchers said.

role: assistant

content: The total of 18 deaths from <1-1_NP> malignant mesothelioma </1-1>, lung cancer and asbestosis was far higher than expected, the researchers said.

Figure 5: 3-Shot Prompt of Proprietary LLMs and Non-Fine-Tuned Base LLMs (The Case of ja→en)