Findings of the 2nd Patent Claims Translation/Evaluation Tasks at WAT2026

Toshiaki Nakazawa
National Institute of Informatics (NII)
nakazawa@nii.ac.jp

Takashi Tsunakawa
Shizuoka University
tuna@inf.shizuoka.ac.jp

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

Katsuhito Sudoh
Nara Women's University
sudoh@ics.nara-wu.ac.jp

Shoichi Okuyama
Patent Attorney in Japan
okuyama@quon-ip.jp

Takashi Ieda
ArtIntell Corp.
takashi.ieda@artintell.tech

Masaaki Nagata
NTT, Inc.
masaaki.nagata@ntt.com

Abstract

This paper presents the results and findings of the 2nd Patent Claims Translation/Evaluation Task at WAT2026. Building on the first task, which collected human error annotations for machine-translated patent claims, this year we introduce an *evaluation* task: given the source claims and the machine translations gathered last year, participants automatically annotate translation errors (spans, categories, and severities) and assign a holistic quality score to each translation. We received submissions from three teams covering the Japanese–English (JA–EN) and English–Japanese (EN–JA) directions. We evaluated the submissions against the organizers' reference annotations in terms of error-span detection, severity, and quality-score correlation and correlated the results with the design of each participating system. Our analysis highlights two observations. First, giving an LLM explicit word alignments mainly helps it detect errors that hinge on the source–target correspondence (mistranslation and terminology), while leaving target-only fluency errors unaffected. Second, locating errors and producing a good overall quality score behave as largely independent abilities: the system whose score correlated best with the reference detected the fewest errors, because it scored quality with a module separate from its error annotator.

1 Introduction

Machine translation (MT) based on neural models and large language models (LLMs) has reached a level where, in some domains and language pairs, its output rivals human translation. However, there is still no universally reliable way to *evaluate* MT quality. Even widely used neural metrics such as COMET (Rei et al., 2020) have been reported to behave unstably outside their training domains (Kocmi et al., 2025). This problem is especially acute for patent claims, which are long, single-sentence, structurally complex, and terminologically demanding, so that a seemingly small error—an omitted limitation or a wrong dependency between claim elements—can change the legal scope of the claimed invention.

The first Patent Claims Translation Task at WAT2025 (Nakazawa et al., 2025) addressed this difficulty by collecting *human* error annotations and holistic quality scores for machine-translated Japanese–English patent claims. That study confirmed that reference-free automatic metrics correlate poorly with human judgments in this domain, and it produced a set of human-annotated translation errors intended as training and development material for automatic evaluation.

Building directly on that effort, the 2nd task at WAT2026 shifts the focus from *translation* to *evaluation*. The machine translations collected in the first task serve as the material to be evaluated: given a source claim and its machine translation, a participating *evaluation* system must (i) annotate translation errors with their spans, MQM-style categories, and severities, and (ii) assign a holistic

Proceedings of the 13th Workshop on Asian Translation (WAT 2026), pages 155–163
November 9, 2026 ©2026 Association for Computational Linguistics

quality score from 0 to 100. The human annotations from WAT2025 are released as training data, and the organizers prepare new reference annotations on held-out translations as the test set.

This paper reports the task setup and the results of the three participating teams. Our contributions are: (1) we define and run the first automatic patent-claim evaluation task; (2) we propose an evaluation protocol that compares heterogeneous submissions against gold reference annotations at the character, span, severity, and score levels; and (3) we analyze how the design of each system—in particular, word-alignment conditioning and the separation of error annotation from quality scoring—manifests in the results.

2 Task Definition

The 2nd task is an *evaluation* task in two translation directions, Japanese-to-English (JA–EN) and English-to-Japanese (EN–JA). For each translated claim, a participating system is given the source claim and its machine translation (the output of an MT system collected in the first task) and must produce, in the same JSON format as the released training data, the following:

- **Error annotation.** The system marks error spans on either the source side (`input`) or the MT-output side (`output`), and assigns to each span a category from a three-level MQM-based hierarchy and a severity. The hierarchy follows the MQM framework (Lommel et al., 2014; Freitag et al., 2021) as adapted to patent translation in WAT2025 (Nakazawa et al., 2025); its top-level categories are *accuracy*, *fluency*, *terminology*, *style*, *other*, *source error*, and *non-translation*. For example, an incorrect article is annotated as *article* under *accuracy* > *mistranslation*.
- **Quality scoring.** The system assigns a holistic quality score from 0 to 100 to each translated claim.

The reference (gold) annotations are produced by the organizers in the same format. In the gold annotations, error spans are marked directly on a character grid aligned with the source and target text, categories are drawn from the MQM-based hierarchy, and severities take three levels (small, medium, large) including intermediate ranges.

The ultimate goal of the task is to develop automatic evaluation methods that assess patent-claim translations as reliably as expert human evaluators, both by localizing concrete errors and by producing calibrated quality scores.

	Training		Test	
	JA–EN	EN–JA	JA–EN	EN–JA
Source documents	13	13	2	2
Source claims	28	28	2	3
MT outputs / claim	6	4	5	4
Translated claims	168	112	10	12
Error annotations	1,130	1,017	93	66
Avg. quality score	75.5	66.0	34.4†	53.7†

Table 1: Statistics of the training and test data. The training data are the WAT2025 human-evaluation results; the test data are newly annotated by the organizers. †The test average is the reference MQM score derived from the gold annotations (Section 5), not a human score.

3 Data

Table 1 summarizes the data. The **training data** are the human-evaluation results of the first task (Nakazawa et al., 2025): for each direction, 28 claims from 13 documents were translated by six (JA–EN) or four (EN–JA) MT systems, and every translation was annotated with error spans and a holistic quality score, yielding 168 JA–EN and 112 EN–JA annotated translations. These annotations are released so that participants can develop and tune their evaluation systems.[1]

The **test data** consist of machine translations held out from the first task, re-annotated by the organizers to serve as the gold reference for this year's evaluation. The JA–EN test set is derived from two single-claim source documents, each paired with five different MT outputs (10 translated claims); the EN–JA test set is derived from two source documents containing three claims in total, each paired with four MT outputs (12 translated claims). Because the gold annotations are produced on the character grid, every error span can be mapped to a character range in the source or target text, which we use for the evaluation described in Section 5. In total the organizers annotated 93 (JA–EN) and 66 (EN–JA) errors on the test translations; the corresponding average reference MQM scores (Section 5) are 34.4 and 53.7 (Table 1).

4 Participants

Three teams submitted evaluation results (Table 2). We summarize each system below based on the

[1] https://github.com/Tzawa/pat-claim-trans

Team	Organization	Directions
Ehime-U	Ehime University	JA–EN
UTSK26	University of Tsukuba	JA–EN, EN–JA
Anonymous	(participant)	JA–EN, EN–JA

Table 2: Participating teams and the directions they submitted.

teams' system-description papers; Ehime-U and UTSK26 are analyzed in detail in Section 7.

Ehime-U Wang et al. (2026) predicts translation errors with an LLM (GPT-6-Astra) that is additionally given explicit word-level correspondences between the source and target claims. The intuition is that many patent-claim errors are alignment-level phenomena: a source word with no target counterpart is an omission candidate, a target word with no source counterpart an addition candidate, and a source word linked to the wrong target word a mistranslation candidate. The team compares six settings that differ in how the correspondence is obtained: four alignments from Unbalanced Optimal Transport (UOT; Arase et al. 2023)—Forward, Reverse, their intersection (AND) and union (OR)—one alignment produced directly by the LLM (GPT6), and a no-alignment baseline (GPT6-noAlignment). The LLM outputs the error list and the quality score in a single response. The team submitted all six settings for JA–EN.

UTSK26 Fu et al. (2026) separates error annotation from quality scoring. Error annotation uses *GEMBA-Patent-Hybrid*, which combines a GEMBA-MQM-style prompt (Kocmi and Federmann, 2023) to GPT-5.2 that detects spans, categories, and severities, with a fallback that labels encoder-detected spans (Buma et al., 2025) when the LLM output is invalid. Quality scoring uses *Patent-Rubric-Checklist*, a rubric of five patent-claim-specific dimensions with sixteen checklist items, whose yes/partial/no answers are aggregated into a 0–100 score outside the LLM. Crucially, the quality score is computed independently of the error annotation.

Anonymous submitted error annotations and quality scores for both directions but did not submit a system-description paper. We therefore report its results without a detailed method analysis, and the team is anonymized here.

5 Evaluation Methodology

Submissions and gold annotations are compared at the claim level. Because submissions differ from the gold in coordinate system (raw-text offsets vs. character-grid offsets), in whether the claim number is included, and in whitespace handling, raw `start`/`end` offsets are not directly comparable. We therefore **re-anchor** every annotation to a non-whitespace character position in the gold text: using each annotation's verbatim span text together with its reported offsets, we align the submission's non-whitespace character sequence to the gold's with a standard sequence-alignment procedure and map the span onto gold coordinates. As a sanity check, evaluating the gold against itself yields $F_1 = 1.0$ on all metrics.

We report the following metrics.

Error detection. Treating each character as "marked as an error or not," we compute character-level precision, recall, and F_1 (on the `input` side, the `output` side, and combined). Precision is the fraction of the system's marked characters that are also marked in the gold (fewer false alarms); recall is the fraction of gold-marked characters recovered (fewer misses). We additionally report a span-level detection F_1 in which a predicted and a gold span are matched when their intersection-over-union is at least 0.5.

Category and severity. Because the gold and the submissions use different category inventories, we do not score category identity directly; instead we report per-gold-category recall (how well each gold category is *located*, regardless of the predicted label) and, where useful, a confusion mapping from gold to predicted categories. For severity, we report the agreement rate over characters marked by both sides.

Quality score. From the gold annotations we derive a reference MQM score in the GEMBA-MQM style (Kocmi and Federmann, 2023): each error contributes a weight (small $=1$, medium $=3$, large $=5$; intermediate ranges take the midpoint; spans left unlabeled count as minor), the per-claim penalty is the sum of weights, and the score is $100 + 4 \cdot \max(-25, -\text{penalty})$, so that scores lie in $[0, 100]$. We correlate this reference score with each team's submitted score using Pearson's r and Spearman's ρ at the claim level ($n = 10$ for JA–EN, $n = 12$ for EN–JA). We emphasize that the reference MQM score is a construct of this formula

and weighting, not a human score, and that the small sample size makes the correlations indicative rather than definitive.

6 Results

Table 3 reports the official results. For **error detection**, the Ehime-U settings lead in JA–EN, with character-level F_1 between 0.454 and 0.512; the best setting is GPT6 (the dense LLM alignment). Anonymous achieves high recall but low precision (0.302), indicating substantial over-annotation, whereas UTSK26 annotates far fewer errors than the gold (3.1 vs. 9.3 annotations per claim) and consequently has the lowest recall and F_1. Span-level F_1 (IoU$\geq$0.5) is uniformly lower than character-level F_1 (0.07–0.28), showing that systems localize the approximate region of an error more reliably than its exact boundaries.

For **score correlation**, the picture is markedly different from detection. In JA–EN, UTSK26 attains the highest Spearman correlation (0.727) despite having the weakest error detection, and Ehime-U/GPT6 achieves the best overall balance ($r = 0.649$, $\rho = 0.633$). We analyze these two contrasting outcomes—the effect of alignment on detection, and the detection–scoring split—in the next section.

The gold MQM scores derived for the test set average 34.4 (JA–EN) and 53.7 (EN–JA); submitted scores are substantially higher on average (around 75 in JA–EN), so absolute score agreement (mean absolute error) is poor even where the rank correlation is moderate-to-high. This systematic leniency is discussed in Section 7.

7 Analysis

7.1 Ehime-U: alignment conditioning selectively helps alignment-level errors

Table 4 breaks the Ehime-U settings down by gold category. Although the combined F_1 varies only slightly across settings—consistent with the team's own observation that the choice of correspondence mainly changes *how many* errors are reported rather than the dominant error types—the per-category view reveals a structured effect. Moving from the no-alignment baseline to the dense LLM alignment (GPT6) raises Accuracy recall from 0.34 to 0.40 and Terminology recall from 0.17 to 0.34 (roughly doubling), while Fluency recall is essentially unchanged (0.51).

This pattern matches the intuition behind the system: alignment information helps errors that depend on the source–target correspondence—mistranslation (Accuracy) and inconsistent terminology—but offers no help for errors that can be judged monolingually on the target, such as fluency. The result thus provides gold-grounded support for Ehime-U's central hypothesis that a subset of patent-claim errors are alignment-level phenomena. Among the UOT variants, the sparse-but-reliable intersection (AND) is strongest, whereas the single-direction Forward alignment falls below the no-alignment baseline, suggesting that dense, bidirectional, high-precision correspondence is the most useful.

7.2 Error localization and holistic scoring behave as separate abilities

In WMT metric evaluation, error-span detection and segment-level quality scoring are treated as distinct subtasks (Freitag et al., 2024; Rei et al., 2023), and a system can be strong on one without the other. Our submissions provide a concrete illustration of the same decoupling in the patent-claim domain. UTSK26 has the lowest detection F_1 in both directions, yet its submitted quality score attains the best Spearman correlation with the gold MQM score in JA–EN (0.727). The explanation lies in its design: UTSK26 computes the quality score with *Patent-Rubric-Checklist*, a rubric evaluated independently of the error annotation, so a sparse or incomplete error list does not prevent the score from tracking overall quality. This is consistent with the UTSK26 paper's own report that the rubric-based score correlates best with human scores on the training data. By contrast, Ehime-U emits errors and score from a single prompt, which yields balanced but not dominant performance on both axes.

Two implications follow. First, a reliable holistic quality score does not require exhaustive error localization; the two capabilities can and perhaps should be optimized separately. Second, if the task goal includes explainability—pointing to the concrete errors behind a score—then UTSK26's sparse annotation is a practical weakness. Indeed, its low annotation density on the test set (Table 3) mirrors the "empty annotation" behavior the team identifies in its own paper, where an empty error list from the LLM is accepted as valid and the encoder-based fallback is not invoked.

Dir.	Team	Ann./ claim	Char-level detection F_1	P	R	Span F_1 (IoU≥.5)	Score corr. Pearson	Spearman
JA–EN	Ehime-U/GPT6	11.2	**0.512**	0.606	0.443	0.283	0.649	0.633
JA–EN	Ehime-U/AND	11.2	0.486	0.588	0.414	0.254	0.635	0.590
JA–EN	Ehime-U/Reverse	10.3	0.485	0.600	0.407	0.276	0.601	0.652
JA–EN	Ehime-U/OR	10.6	0.482	0.584	0.410	0.261	0.613	0.590
JA–EN	Ehime-U/GPT6-noAlignment	11.1	0.475	0.580	0.402	0.255	0.634	0.623
JA–EN	Ehime-U/Forward	11.0	0.454	0.565	0.379	0.256	0.633	0.561
JA–EN	Anonymous	15.8	0.375	0.302	0.494	0.104	0.674	0.364
JA–EN	UTSK26	3.1	0.111	0.121	0.102	0.097	0.654	**0.727**
EN–JA	Anonymous	22.5	0.154	0.089	0.562	0.071	0.604	0.600
EN–JA	UTSK26	4.2	0.139	0.183	0.112	0.155	0.383	0.320

Table 3: Official results per team and direction. "Ann./claim" is the average number of annotations per claim (for reference, the gold density is 9.3 for JA–EN and 5.5 for EN–JA). Detection is character-level against the gold; score correlation is between the gold MQM score and the submitted score ($n = 10$ for JA–EN, $n = 12$ for EN–JA). Best values in each column group are in bold.

Setting	Det F_1	Det F_1^{out}	Recall by gold category Acc.	Term.	Flu.
GPT6 (dense)	**0.512**	**0.555**	**0.40**	**0.34**	0.51
UOT-AND	0.486	0.528	0.36	0.17	0.48
UOT-Reverse	0.485	0.523	0.35	0.17	0.51
UOT-OR	0.482	0.526	0.35	0.34	0.52
no-align baseline	0.475	0.510	0.34	0.17	0.51
UOT-Forward	0.454	0.485	0.28	0.31	0.52

Table 4: Ehime-U six settings (JA–EN). "Det F_1^{out}" is the output-side character F_1; "Recall by gold category" is how well the system locates gold errors of each top-level category (Accuracy, Terminology, Fluency).

7.3 Systematic leniency of submitted scores

Across teams, submitted quality scores are considerably higher than the gold MQM scores (e.g., a JA–EN gold mean of 34.4 against submitted means around 75). Because Pearson and Spearman correlations are invariant to such an offset, the moderate correlations coexist with large mean absolute errors. This indicates that current systems rank translations by quality more reliably than they calibrate the absolute score, and it highlights the sensitivity of the reference MQM score to the severity weighting and the per-segment clipping at -25 (which places several heavily penalized claims at 0).

7.4 Reference-score gap between the two directions

The reference MQM scores differ markedly between directions (JA–EN mean 34.4, EN–JA 53.7). The gap is mechanical: JA–EN test translations carry about 70% more error annotations per claim than EN–JA (9.3 vs. 5.5) and a larger share of high-weight errors (14 large and 5 medium–large, vs. 7 and 8), so the per-claim penalty is larger and more JA–EN claims hit the -25 floor (three of ten, vs. one of twelve), pulling the mean down. We caution against reading this as a general quality gap between directions: each direction has only two source documents and 10–12 translations, so a few heavily penalized claims dominate the average, and the reference score reflects *how many* errors the organizers marked rather than a human holistic judgment. Indeed, it runs opposite to the WAT2025 training human scores, where JA–EN averaged higher than EN–JA (75.5 vs. 66.0; Table 1).

8 Discussion

Heterogeneity of submission formats. The submissions varied in coordinate system (raw-text vs. character-grid offsets), in whether the leading claim number was retained, and in whitespace, so that a naive comparison of `start`/`end` offsets was impossible. We absorbed these differences by re-anchoring every span to the gold text (Section 5), but for a shared task a stricter output specification—fixed coordinate system, a shared category vocabulary, and a fixed set of severity levels—would make submissions directly comparable and reduce the risk of scoring artifacts.

Mismatch of category inventories. The gold annotations use abstract MQM top-level categories (accuracy, fluency, terminology), whereas the submissions use finer labels (omission, mistranslation, article, incorrect dependency, etc.). We therefore evaluated *where* errors were found rather than *whether the exact label matched*. A confusion mapping from gold to predicted categories (produced

as part of the evaluation) shows, for example, that the gold *accuracy* class maps mostly to predicted *mistranslation* and *dependency* errors, and *fluency* to *article* and *grammar*. For future editions, agreeing on a single label set—or an explicit crosswalk between label sets—would allow category-level scoring.

Severity granularity. The gold uses three severity levels (small, medium, large) plus intermediate ranges, while the submissions use two (major, minor). The treatment of "medium" in the MQM score (we used a midpoint weight of 3) directly affects the reference scores and hence the correlations; aligning the severity scales across gold and submissions is desirable.

Detection vs. scoring. Error-span detection and segment-level scoring are already separate subtasks in WMT metric evaluation (Freitag et al., 2024). The observation in Section 7.2 suggests that the patent task should likewise report the two capabilities on separate leaderboards: error localization (span/character F_1, severity) and holistic scoring (correlation with the reference). A system can be strong on one and weak on the other, and conflating them would obscure where the state of the art actually stands.

Small test set. The current test set is small (10 JA–EN and 12 EN–JA claims). While it is sufficient to expose qualitative effects—such as the selective benefit of alignment and the detection/scoring dissociation—larger and more diverse test data are needed to draw robust quantitative conclusions and to compute stable correlations.

9 Conclusion and Future Perspective

We described the 2nd Patent Claims Translation/Evaluation Task at WAT2026, which reframes the first task's translation effort as an *evaluation* task: participating systems annotate translation errors and assign quality scores to machine-translated patent claims, and are compared against the organizers' reference annotations. Three teams participated, covering both translation directions.

Our analysis yields two main findings. First, conditioning an LLM on explicit word alignments (Ehime-U) selectively improves the detection of alignment-dependent errors—accuracy and terminology consistency—while leaving monolingual fluency errors unchanged, giving gold-grounded support to the hypothesis that some patent-claim errors are alignment-level phenomena. Second, consistent with the separate error-detection and scoring subtasks in WMT metric evaluation, error localization and holistic scoring behaved as largely independent abilities here: a system with a separate rubric-based scorer (UTSK26) ranked translations best despite detecting the fewest errors.

In future editions we plan to enlarge and diversify the test set, standardize the submission format and the category/severity inventories so that submissions are directly comparable, and report error localization and holistic scoring as separate tracks. We also intend to use the growing body of human and reference annotations as training data for more accurate and better-calibrated automatic evaluation of patent-claim translation.

Limitations

The findings reported here are subject to several limitations. The test set is small (10 JA–EN and 12 EN–JA claims), so detection scores and especially correlations are indicative rather than definitive and are sensitive to a few rank swaps. The reference MQM quality score is a construct of this paper (severity weights small/medium/large $= 1/3/5$, ranges at their midpoint, unlabeled spans counted as minor, and per-segment clipping at -25); the choice of the "medium" weight and of clipping affects the reported correlations, and the score is not a human judgment. Because the gold and the submissions use different category inventories, we evaluate error *location* rather than exact category identity. Input-side (source) gold annotations are very sparse (only two in JA–EN), so the input-side metrics are low-signal and our evaluation is effectively output-centric. Span-level F_1 requires boundary agreement at IoU ≥ 0.5 and is therefore conservative; the character-level metric better reflects whether an error location was found. Finally, each system's LLM output is essentially a single run, so output stability is not assessed.

Acknowledgments

We thank the participating teams for their submissions and system descriptions. The human annotation of the reference data is supported by the Asia-Pacific Association for Machine Translation (AAMT) and the AAMT/Japio[2] Special Interest Group on Patent Translation.

[2]Japan Patent Information Organization

References

Yuki Arase, Han Bao, and Sho Yokoi. 2023. Unbalanced optimal transport for unbalanced word alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3966–3986, Toronto, Canada. Association for Computational Linguistics.

Kosei Buma, Takehito Utsuro, and Masaaki Nagata. 2025. Two step automatic post editing of patent machine translation based on pre-trained encoder models and LLMs. In *The 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 218–231, 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, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Yingyi Fu, Kosei Buma, and Takehito Utsuro. 2026. UTSK26 at WAT2026 patent claims translation/evaluation tasks. In *Proceedings of the 13th Workshop on Asian Translation (WAT 2026)*. System description paper.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. 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.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):455–463.

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.

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

Xu Yin Wang, Isao Goto, and Takashi Ninomiya. 2026. Ehime-U system for the WAT 2026 patent claims evaluation task: LLM-based translation error prediction using word alignments. In *Proceedings of the 13th Workshop on Asian Translation (WAT 2026)*. System description paper.

A Re-anchoring Procedure

Submissions differ from the gold in coordinate system (raw-text vs. character-grid offsets), in whether the leading claim number is kept, and in whitespace, so a span's reported `start`/`end` cannot be compared to the gold directly (Section 5). We map every annotation onto a common coordinate—the index of a non-whitespace character in the gold text—as follows.

1. For an annotation, pick the reference string whose `[start:end]` slice equals its verbatim `text`: the character grid if present, otherwise the raw text. This identifies which coordinate system the submission used.

2. Convert the span's endpoints to *non-whitespace* offsets by counting non-whitespace characters before each endpoint. This removes all whitespace differences.

3. Build the non-whitespace character sequences of the submission and the gold for that side and align them with a standard sequence

Dir.	Claim	Penalty	Score
JA–EN	01	19	24
JA–EN	02	9	64
JA–EN	03	19	24
JA–EN	04	50	0
JA–EN	05	17	32
JA–EN	06	6	76
JA–EN	07	0	100
JA–EN	08	25	0
JA–EN	09	42	0
JA–EN	10	19	24
EN–JA	01-1	16	36
EN–JA	01-2	0	100
EN–JA	02-1	20	20
EN–JA	02-2	6	76
EN–JA	03-1	10	60
EN–JA	03-2	14	44
EN–JA	04-1	11	56
EN–JA	04-2	0	100
EN–JA	05	11	56
EN–JA	06	47	0
EN–JA	07	15	40
EN–JA	08	11	56

Table 5: Reference MQM penalty and score per test claim (Eq. 1). Means: 34.4 (JA–EN), 53.7 (EN–JA).

matcher (Python's `difflib`); matched positions give a map from submission non-whitespace indices to gold non-whitespace indices. This absorbs the claim-number and minor text differences.

4. Map the span's non-whitespace endpoints through this map to obtain its position in the gold coordinate.

Gold annotations are expressed in the gold coordinate directly (step 3 is the identity). As a sanity check, evaluating the gold against itself with this procedure yields $F_1 = 1.0$ on every metric, confirming that the re-anchoring is lossless.

B Reference MQM Score

The reference quality score follows the GEMBA-MQM convention. Each error contributes a severity weight (small $= 1$, medium $= 3$, large $= 5$; a severity range, e.g. medium-to-large, takes the midpoint of its endpoint weights; spans left unlabeled count as minor). The per-claim penalty is the sum of weights, and the score is

$$\text{score} = 100 + 4 \cdot \max(-25, -\text{penalty}), \quad (1)$$

so that a clipped score lies in $[0, 100]$. Table 5 lists the penalty and score for every test claim; EN–JA files with two claims are shown as `file-block`.

Dir.	Team	Acc.	Term.	Flu.	Unlab.
JA–EN	Ehime-U GPT6	0.40	0.34	0.51	0.51
JA–EN	Ehime-U AND	0.36	0.17	0.48	0.53
JA–EN	Ehime-U Reverse	0.35	0.17	0.51	0.51
JA–EN	Ehime-U OR	0.35	0.34	0.52	0.48
JA–EN	Ehime-U GPT6-noAlignment	0.34	0.17	0.51	0.51
JA–EN	Ehime-U Forward	0.28	0.31	0.52	0.51
JA–EN	Anonymous	0.30	0.14	0.74	0.83
JA–EN	UTSK26	0.15	0.17	0.06	0.01
EN–JA	Anonymous	0.42	0.86	0.96	0.61
EN–JA	UTSK26	0.08	0.03	0.17	0.18

Table 6: Per-gold-category recall (Accuracy, Terminology, Fluency, Unlabeled).

Gold category	Predicted categories (char counts)
Ehime-U/GPT6	
Accuracy	mistranslation (190), dependency (157), numerals/symbols (25), article (21), grammar (17)
Fluency	article (128), grammar (21), punctuation (3)
Terminology	mistranslation (17), terminology (17)
Unlabeled	omission (99), dependency (78), over-translation (41), mistranslation (30), article (10)
UTSK26	
Accuracy	numerals/symbols (73), omission (38), context/word choice (38), grammar (10)
Fluency	omission (12), article (6)
Terminology	context/word choice (17)
Unlabeled	mistranslation (7)

Table 7: Gold-to-submission category confusion (overlapping characters) for two JA–EN systems. Submission labels are glossed in English (e.g. *dependency*, *over-translation*, *omission*).

C Per-Category Recall

Table 6 gives, for every team, the recall for each top-level gold category—how well the system *locates* gold errors of that category, regardless of the predicted label (Section 5). "Unlab." denotes gold spans highlighted without a category label.

D Category Confusion (Gold → Submission)

Because the gold and the submissions use different category inventories, we do not score label identity; instead, for characters marked by both sides, we count which predicted category a gold category maps to. Table 7 shows this mapping (character counts) for two representative JA–EN systems. The gold's abstract *accuracy* class maps mostly to the

submissions' concrete *mistranslation* / *dependency* labels, and *fluency* maps to *article* and *grammar*, confirming that the two inventories are compatible once the abstraction gap is accounted for.