Tagged Span Annotation for Detecting Translation Errors in Reasoning LLMs

Taemin Yeom[1,2], Yonghyun Ryu[2], Yoonjung Choi[2†], JinYeong Bak[3†]

[1]Department of Digital Media and Communications Engineering, Sungkyunkwan University,
[2]Samsung Research,
[3]Department of Computer Science and Engineering, Sungkyunkwan University
taemin.yeom@g.skku.edu
{yonghyun.ryu, yj0807.choi}@samsung.com
jy.bak@skku.edu

Abstract

We present the submission of the AIP team to the WMT 2025 Unified MT Evaluation Shared Task, focusing on the span-level error detection subtask. Our system emphasizes response-format design to better harness the capabilities of OpenAI's o3, the state-of-the-art reasoning LLM. To this end, we introduce Tagged Span Annotation (TSA), an annotation scheme designed to more accurately extract span-level information from the LLM. On our refined version of WMT24 ESA dataset, our reference-free method achieves an F1 score of approximately 27 for character-level label prediction, outperforming the reference-based XCOMET-XXL at approximately 17.[1]

1 Introduction

With the recent widespread use of the LLM-as-a-judge approach, research on human-like translation quality evaluation using large language models (LLMs)—such as GEMBA-DA (Kocmi and Federmann, 2023b), MQM (Kocmi and Federmann, 2023a), ESA (Zouhar et al., 2025), EAPrompt (Lu et al., 2024), and AutoMQM (Fernandes et al., 2023)—has grown rapidly. Fine-grained error detection enables explainable translation evaluation and informs the design of post-editing systems, making it widely applicable across diverse systems that leverage machine translation. Nevertheless, research on fine-grained translation error detection remains scarce; only a handful of studies exist despite growing interest. Moreover, the LLM-based studies mentioned above focused less on fine-grained error detection itself and more on using it to compute final machine translation (MT) evaluation scores. Additionally, most prior work (Kocmi and Federmann, 2023b,a; Fernandes et al., 2023) has been limited to non-reasoning LLMs. With the recent emergence of reasoning models that have achieved state-of-the-art performance across a wide range of tasks, there is, therefore, a need to investigate their use for fine-grained error detection as well.

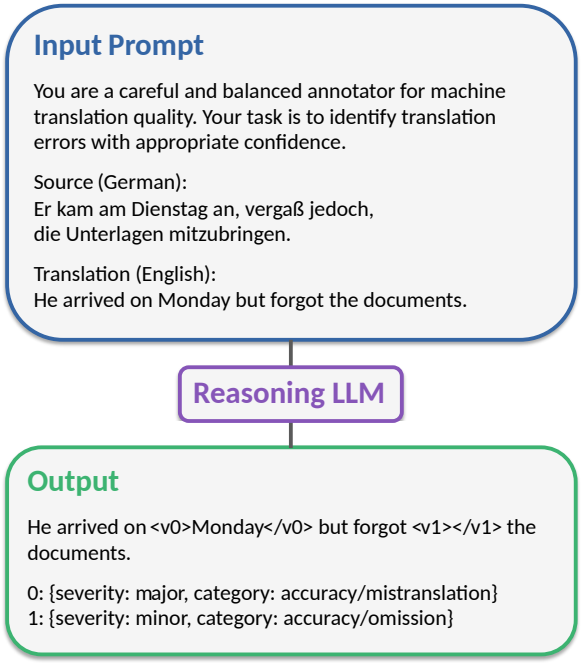

Figure 1: The overview of **Tagged Span Annotation (TSA)** system. The reference translation is *He arrived on Tuesday, but forgot to bring the documents.* The LLM tags each error span in the hypothesis with inline tags <vN> and returns a severity–category pair for every tag.

In this paper, we propose a translation error span detection system that leverages OpenAI o3 (OpenAI, 2025a), the state-of-the-art reasoning large language model, for the WMT'25 MT evaluation span-level error detection task. Our system adopts

[†]Corresponding authors
[1]Code and dataset repo: https://github.com/TaeminYeom/Tagged_Span_Annotation

Proceedings of the Tenth Conference on Machine Translation, pages 878–886
November 8-9, 2025

Lang	TSA (Ours)			XCOMET-XXL-QE			XCOMET-XXL			GEMBA			GEMBA (fixed)		
	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
en-cs	**25.24**	**23.32**	27.49	15.07	9.73	**33.42**	15.90	10.44	33.33	4.25	12.13	2.57	18.26	14.24	25.44
en-ja	**32.15**	**21.91**	**60.35**	7.84	4.69	23.98	8.63	5.24	24.46	1.42	1.70	1.21	5.73	3.66	13.20
en-zh	**17.86**	**13.07**	28.21	7.32	4.07	**36.03**	7.58	4.25	35.24	4.11	5.38	3.33	9.72	6.35	20.65
en-is	32.09	**47.66**	24.19	33.77	30.74	37.46	**35.48**	33.27	**38.01**	1.19	26.43	0.61	30.95	38.68	25.79
en-uk	**31.10**	**29.52**	**32.85**	15.83	11.52	25.28	17.01	12.72	25.67	1.48	7.69	0.82	14.92	13.41	16.82
en-ru	**23.87**	**28.54**	20.51	18.05	16.12	20.50	19.18	16.85	**22.26**	3.79	18.85	2.11	19.39	19.59	19.19
Avg.	**27.05**	**27.34**	**32.27**	16.31	12.81	29.45	17.30	13.79	29.83	2.71	12.03	1.78	16.50	15.99	20.18

Table 1: Span-level detection performance of TSA method compared with baseline systems on WMT24 ESA dataset six language pairs. TSA consistently achieves the highest average F1, Precision, and Recall in most languages.

structured output to ensure reliable parsing of responses and represents error locations using tagged-span annotations. On our refined version of the WMT24 ESA dataset, our reference-free system achieved an F1 score of 27.05, outperforming the reference-based XCOMET-XXL at 17.30.

2 Method

2.1 Background

Transformer-encoder approaches such as XCOMET (Guerreiro et al., 2024) typically cast span-level error detection as a token-classification task on the translation, assigning a severity label—*no-error*, *minor*, or *major*—to every token.

In generative-response settings, error spans can be expressed in three canonical ways:

(i) by returning the erroneous text itself

(ii) by providing its start- and end-character indices

(iii) by inserting inline tags directly into the translation

Method (i) is the format adopted by both GEMBA-MQM and GEMBA-ESA. In GEMBA-ESA, the returned span string is used only as a hint, and in MQM the sentence-level score is computed solely from the number of spans and their severity labels. Hence, for *sentence-level scoring* this representation is sufficient. For a *span-detection* task, however, it is inadequate: the same substring can occur multiple times in a sentence, so the actual error location remains ambiguous.

Method (ii)—returning start and end character indices looks attractive for its simplicity and precision. In practice, however, generative LLMs struggle with producing *exact* numerical values: they operate on sub-word (BPE) tokens rather than raw characters, so mapping token boundaries back to UTF-8 offsets is non-trivial (Zhang and He, 2024)

Method (iii) follows the inline-tag scheme used in the publicly released *WMT MQM Human-Evaluation* dataset (Freitag et al., 2021). In that corpus, each record contains exactly one error span; when a translation exhibits multiple errors, it is split into multiple records, each highlighting a single span.

2.2 Tagged Span Annotation (TSA)

In our system we adopt **method (iii)** and extend it with *numbered inline tags*: each error span is wrapped in a unique <v$_k$>–</v$_k$> pair, allowing multiple errors to be annotated simultaneously within the same sentence.

When a source segment is omitted in the hypothesis, we mark the omission by inserting a zero-length tag pair <v$_k$></v$_k$> at the exact insertion point.

To elicit the desired response format, we composed a detailed system prompt, supplied few-shot examples that explicitly illustrate the target schema, and leveraged the model's *structured-output* interface to guarantee machine-parseable answers. (see Appendix A; OpenAI, 2023a).

We evaluate three OpenAI models—**GPT-4.1** (OpenAI, 2025b), **o3** (OpenAI, 2025a), and **o4-mini** (OpenAI, 2025a). Decoding parameters follow the API defaults, except that GPT-4.1 uses a lower temperature of 0.2 to reduce variance.

3 Experiments

3.1 Datasets

We use the our refined version of **WMT24** ESA Annotations datasets[2] for six language pairs: **en→{cs,**

[2] https://github.com/wmt-conference/wmt24-news-systems

is, ja, ru, uk, zh}. Each pair contains approximately **6–8k** segments, for a total of about **39,685** segments. We preserve the human **ESA gold** annotations as the reference for span-level evaluation. [3]

3.2 Metrics

We follow the official WMT25 Task 2 ESA scorer[4]. Span detection is evaluated by **character-level overlap**, not span matching: for each character position, the counts of *major/minor* errors in gold and prediction are compared, and the true-positive score is calculated as the sum of the minimum counts at each character position. Partial overlaps that match in location but not in severity receive **partial credit** (0.5). Omissions marked as `start="missing"` are ignored by the scorer. MQM categories are not used for scoring. We report precision, recall, and F1 as micro-averages over segments within each language, and then macro-averages across languages.

3.3 Baselines

We compare our methods against two established baselines: **XCOMET** (Guerreiro et al., 2024) and **GEMBA** (Kocmi and Federmann, 2023a). Both XCOMET and GEMBA are also used as official baselines in the WMT25 Shared Task. All scores are character-level and macro-averaged across the six MQM language pairs; within each language we compute micro averages over segments. For GEMBA, we adhere to the original prompt and settings with one modification: the few-shot prompt is revised to require outputs in a strict JSON format to facilitate reliable parsing. In preliminary experiments, the unmodified prompt frequently elicited explanatory text in addition to the JSON output, resulting in parsing failures and degraded performance. To mitigate this issue, we explicitly instructed the model to omit any explanatory content, which yielded improved performance. We denote this variant as GEMBA (fixed).

[3]The script to refine the WMT24 ESA has been merged into the WMT25 official repository: https://github.com/wmt-conference/wmt25-mteval/blob/3332614/scripts/devset/create_tsv_from_wmt24_esa.sh

[4]https://github.com/wmt-conference/wmt25-mteval/blob/3332614/scripts/scoring/task2/scoring_esa.py

Method	F1	Prec.	Rec.
GEMBA	19.70	20.70	21.54
Direct-index	22.57	27.32	23.50
TSA (no precision emphasis)	25.43	21.99	**38.19**
TSA	**27.05**	**27.34**	32.27

Table 2: Ablation on output format and precision-emphasis prompting on o3 model. GEMBA, Direct-index, and TSA use an **identical prompt**, differing only in the directive that specifies the output format. *TSA (no precision emphasis)* employs the same TSA format but removes the precision-oriented instruction, isolating its contribution.

4 Results

Main results Table 1 presents the primary results. Our best system combines the **o3** reasoning model with the *Tagged Span Annotation* output design and achieves the highest F1 score. Under the same evaluation setting, it surpasses the *reference*-based **XCOMET-XXL** by **+9.65** F1 and the **GEMBA (fixed)** baseline by **+10.55** F1.

4.1 Ablation Study

We perform extensive experiments to identify the factors that influence response quality, and we report our findings here.

4.1.1 Span Annotation

We compare three output formats (i) **TSA**, (ii) a **GEMBA** format that returns the error-span text itself and (iii) a **Direct-index** format that outputs the character indices of each span using the **o3** model.

Except for the directive specifying the output format, every part of the prompt is kept identical, allowing us to isolate performance differences attributable solely to the span-annotation scheme. As shown in Table 2, the performance gap between the GEMBA and Direct-index baselines and our TSA method confirms this effect.

4.1.2 Precision Emphasis

In the TSA setting, we observe a pronounced precision–recall imbalance: recall was consistently much higher than precision in most language pairs.

Because the F1 score is the harmonic mean of precision and recall, it is dominated by the lower of the two components; consequently, it attains higher values only when the two metrics are balanced. To mitigate the observed imbalance, we explicitly

Method	GPT-4.1			o3			o4-mini		
	F1	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**
GEMBA	**21.16**	**20.84**	21.16	19.70	20.70	21.54	17.42	15.28	23.50
Direct-index	13.46	14.92	14.95	22.57	27.32	23.50	18.13	19.30	20.76
TSA	20.03	18.45	**26.82**	**27.05**	**27.34**	**32.27**	**21.26**	**19.95**	**26.86**

Table 3: Span-level detection Performance by model and output format. Tagged Span Annotation (TSA) shows no noticeable performance gains over GEMBA or Direct-index on the non-reasoning baseline `GPT-4.1`, but it delivers substantial improvements on the reasoning models `o3` and `o4-mini`, highlighting the pronounced benefit of span markup for models with stronger reasoning capability.

reinforced the requirement of promoting higher precision.

We instruct the model to assign an error label only when it is *confident* the translation is incorrect. Moreover, because the model tends to select spans wider than the actual error, the prompt explicitly requires it to tag only the *minimal* substring that covers the core error.

For comparison, Table 2 also reports a TSA variant in which the prompt *did not* include the precision-oriented instructions.

4.1.3 Reasoning Impact

As shown in Table 3, the gap between TSA and the GEMBA format is most pronounced for **reasoning models**. We interpret that this stems from the two-stage decision process enforced by TSA, which closely resembles the human ESA scoring protocol: human annotators **first** indicate the mark spans and severity of errors, and then score the translation quality. ESA has been reported to achieve inter-annotator agreement comparable to full MQM annotation (Kocmi et al., 2024). Similarly, TSA requires the LLM to mark error spans in the `annotated_translation` field before predicting their severity and type, thus externalising an intermediate reasoning step.

The non-reasoning baseline GPT-4.1 shows no performance improvement from TSA over GEMBA, whereas the reasoning models `o3` and `o4-mini` exhibit substantial gains. This suggests that models with stronger reasoning ability are better able to exploit the intermediate span markup to conduct a more fine-grained error analysis.

5 Related Work

5.1 Encoder-Based Evaluation

COMET (Rei et al., 2020) is the first to attach a *regression head* to a pretrained cross-lingual encoder (e.g., XLM-R) in order to predict Direct Assessment (DA) scores, achieving higher human correlation than traditional metrics such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). **COMET-MQM** (Rei et al., 2021) added an *auxiliary token-classification task* that predicts whether each token falls inside an error span based on MQM annotations.

XCOMET (Guerreiro et al., 2024) unifies and generalizes this line of work by (i) retaining the DA regression head and (ii) replacing COMET-MQM's token classifier with a *span head* that emits start–end indices. Yet these index-based spans lack error *types/categories*, limiting diagnostic power and compatibility with MQM/ESA labels, and the primary training objective remains sentence-level scoring, leaving boundary accuracy secondary. We treat XCOMET-XXL as the strongest encoder-based baseline and, inspired by its index span annotation, test an Direct-index format variant in our LLM-as-judge pipeline (§4.1.1); however, vulnerability to character-index drift ultimately leads us to adopt *Tagged Span Annotation (TSA)* for our final design.

5.2 LLM-as-a-judge

GEMBA (Kocmi and Federmann, 2023a) frames MT evaluation as an LLM-as-judge task: a prompted GPT-4 produces Direct Assessment (DA) scores or MQM/ESA error spans and severities from the source and translation (optionally a reference). However, GEMBA's few-shot, free-form outputs can yield duplicated instructions across exemplars, brittle parsing. To improve robustness, we convert the output to a JSON format and adopted structured-output prompting.

MQM-APE (Lu et al., 2025) augments GEMBA with a downstream *automatic post-editing* (APE) pass: after the initial GPT-4 annotation, a second LLM rewrites the MT segment with the pro-

posed fixes and discards spans whose removal does not change the edited meaning. This post-filtering raises span precision with little effect on recall. Because MQM-APE operates after span generation, it is orthogonal to our TSA, JSON-validated spans (with minimal fixes) could be plugged into MQM-APE without modifying its algorithm. We leave a full integration and evaluation to future work.

6 Limitations

This study has the following limitations. First, the experiments were conducted primarily on proprietary large language models (LLMs) such as `gpt-4.1`, `o3` and `o4-mini`, without providing comparable results for large-scale open-source models. This limits the generalizability of the proposed approach to other model families. Second, although per-language results were reported, we did not conduct in-depth analyses of performance variations or error patterns across specific language pairs. In particular, the causes of performance differences for low-resource languages remain underexplored. Third, the monetary cost of using proprietary LLMs can hinder large-scale adoption. Using `tiktoken` (OpenAI, 2023b) to count tokens, our evaluation on six WMT24 ESA language pairs (39,684 segments, 56.82M input tokens, 3.26M output tokens) is estimated to have cost about $139 in total ($3.5 per 1,000 segments) under the current `o3` pricing ($2/M input, $8/M output as of September 2025), which may hinder large-scale adoption despite the accuracy gains.

7 Conclusion and Future Work

We introduced **Tagged Span Annotation** (TSA), a structured-output framework that pairs numbered inline tags with a JSON schema to enable reliable, fine-grained translation-error detection. Across six WMT language pairs, TSA achieved the highest span-level F1 scores, surpassing the strongest GEMBA variant by +10.55 and XCOMET-XXL by +9.65 absolute points on average.

Reasoning vs. non-reasoning models. We observed performance differences between reasoning and non-reasoning LLMs. As future work, we will systematically study how prompting style and intermediate reasoning affect span detection by comparing: (i) direct answers vs. chain-of-thought (CoT; Wei et al. 2022) prompting, (ii) single-pass decoding vs. self-consistency with multiple CoT paths, and (iii) unguided CoT vs. CoT *path selection* aided by a span-level verifier. We will evaluate not only span F1 (overall and by severity/type), but also invalid-output rate, span-length bias, latency, and cost to characterize accuracy–efficiency trade-offs.

Open-source models trained on evaluation data. To improve reproducibility and broaden applicability, we will fine-tune strong open-source LLMs on human-annotated MQM/ESA data (and carefully curated synthetic data), comparing SFT vs. parameter-efficient adapters (e.g., LoRA; Hu et al. 2021) under multilingual vs. per-language regimes. We will analyze cross-lingual transfer (high- to low-resource), domain shift, and calibration quality, and report effect sizes alongside standard correlations. Where licensing permits, we plan to release training code, prompts, and evaluation harnesses to facilitate future benchmarking.

Coupling with post-editing (MQM-APE). We further plan to couple our span detector with post-editing systems that explicitly target MQM categories (e.g., MQM-aware APE). Concretely, we will explore (i) using predicted spans and severities to guide edit proposals, (ii) verifier- or reward-based reranking of edits, and (iii) joint training where an APE loss encourages span-consistent corrections. We will report pre-/post-edit score deltas, human correlation, and downstream adequacy/fluency gains to quantify the benefit of integrating detection with correction.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful questions and comments. JinYeong Bak was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-RS-2025-00523385).

References

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. 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. *Preprint*, arXiv:2104.14478.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. *Preprint*, arXiv:2406.11580. ArXiv preprint.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2023a. Function calling and other api updates. https://openai.com/index/function-calling-and-other-api-updates/. Accessed 2025-09-26.

OpenAI. 2023b. tiktoken. https://github.com/openai/tiktoken. Accessed 2025-09-26.

OpenAI. 2025a. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed 2025-09-26.

OpenAI. 2025b. System card: Gpt-4.1. https://openai.com/index/gpt-4-1/. Accessed 2025-09-26.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yidan Zhang and Zhenan He. 2024. Large language models can not perform well in understanding and manipulating natural language at both character and word levels? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11826–11842, Miami, Florida, USA. Association for Computational Linguistics.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

A Prompt

A.1 System Prompt

We used the following system prompt. It described the essential rules for marking error spans and placed emphasis on addressing the issue of precision being significantly lower than recall.

```
You are a careful and balanced annotator for machine translation quality.
Your task is to identify translation errors with appropriate confidence.

## EVALUATION GUIDELINES:
- Be thorough but precise: Only mark errors when you are confident they are
↪  incorrect
- Consider context and domain: Some variations may be acceptable depending on
↪  context
- Distinguish between errors and acceptable alternatives: Multiple valid
↪  translations may exist
- Focus on clear, objective errors rather than subjective preferences
- Verify each potential error against the source text before marking
- When in doubt, err on the side of not marking an error

## Error Categories:
- Accuracy: addition, mistranslation, omission, untranslated text
- Fluency: character encoding, grammar, inconsistency, punctuation, register,
↪  spelling
- Style: awkward phrasing
- Terminology: inappropriate for context, inconsistent use
- Other: non-translation, other issues

## Severity Classification:
- Major: Errors that impact meaning or usability but do not render the text unusable
- Minor: Errors that do not impact meaning or usability

## CRITICAL OUTPUT REQUIREMENTS:
- Mark errors only when you have clear evidence they are incorrect
- Consider whether alternative translations could be equally valid
- Apply strict standards: better to miss a minor error than create a false positive
- Wrap error spans in the translation with tags: <v0>, <v1>, <v2>, etc.
- Use tag numbers in sequential order starting from <v0>. Do not skip numbers or use
↪  them out of order
- For omissions, place empty tags <vN></vN> where the missing text should have been
- NO comments, explanations, or additional text
- Mark only the minimal substring that contains the clear error; do not include
↪  extra context
```

A.2 Few-shot Examples

We revised the GEMBA-MQM (Kocmi and Federmann, 2023a) few-shot examples to align with the TSA format. In addition, we added one example without any error spans.

```
Source (English):
```

```
I do apologise about this, we must gain permission from the account holder to
↪  discuss an order with another person, I apologise if this was done previously,
↪  however, I would not be able to discuss this with yourself without the account
↪  holders permission.

Translation (German):
Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung
↪  mit einer anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor
↪  geschehen wäre, aber ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der
↪  Lage, dies mit dir involvement.

Answer:
{
"annotated_translation": "Ich entschuldige mich dafür, wir müssen die
↪  Erlaubnis<v0></v0> einholen, um eine Bestellung mit einer anderen Person zu
↪  besprechen. Ich entschuldige mich, falls dies zuvor geschehen <v1>wäre</v1>,
↪  aber ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit
↪  <v2>dir</v2> <v3>involvement</v3>.",
"errors": [
    {"severity": "Major", "category": "accuracy/omission"},
    {"severity": "Minor", "category": "fluency/grammar"},
    {"severity": "Minor", "category": "fluency/register"},
    {"severity": "Major", "category": "accuracy/mistranslation"}
]
}

---

Source (English):
Talks have resumed in Vienna to try to revive the nuclear pact, with both sides
↪  trying to gauge the prospects of success after the latest exchanges in the
↪  stop-start negotiations.

Translation (Czech):
Ve Vídni se ve Vídni obnovily rozhovory o oživení jaderného paktu, přičemž obě
↪  partaje se snaží posoudit vyhlídky na úspěch po posledních výměnách v jednáních.

Answer:
{
"annotated_translation": "Ve Vídni se <v0>ve Vídni</v0> obnovily rozhovory o
↪  oživení jaderného paktu, přičemž obě <v1>partaje</v1> se snaží posoudit
↪  vyhlídky na úspěch po posledních výměnách v<v2></v2> jednáních.",
"errors": [
    {"severity": "Major", "category": "accuracy/addition"},
    {"severity": "Minor", "category": "terminology/inappropriate for context"},
    {"severity": "Major", "category": "accuracy/omission"}
]
}

---

Source (Chinese):
```

```
大点木家居道提供高居然之家地址，等最新商信息，修公司，就上大点

Translation (English):
Urumqi Home Furnishing Store Channel provides you with the latest business
↪  information such as the address, telephone number, business hours, etc., of
↪  high-speed rail, and find a decoration company, and go to the reviews.

Answer:
{
"annotated_translation": "Urumqi Home Furnishing Store Channel provides you with
↪  the latest business information such as the address, telephone number, business
↪  hours, <v0>etc.,</v0> <v1>of high-speed rail,</v1> and find a decoration
↪  company, and <v2>go to the reviews</v2>.",
"errors": [
    {"severity": "Minor", "category": "style/awkward"},
    {"severity": "Major", "category": "accuracy/addition"},
    {"severity": "Major", "category": "accuracy/mistranslation"}
]
}

---

Source (English):
According to the terms outlined in the agreement, the supplier shall deliver all
↪  components no later than thirty days after receiving the initial purchase order,
↪  and any delays must be communicated in writing at least five business days in
↪  advance.

Translation (Spanish):
De acuerdo con los términos establecidos en el acuerdo, el proveedor deberá
↪  entregar todos los componentes a más tardar treinta días después de recibir la
↪  orden de compra inicial, y cualquier retraso deberá comunicarse por escrito con
↪  al menos cinco días hábiles de antelación.

Answer:
{
"annotated_translation": "De acuerdo con los términos establecidos en el acuerdo,
↪  el proveedor deberá entregar todos los componentes a más tardar treinta días
↪  después de recibir la orden de compra inicial, y cualquier retraso deberá
↪  comunicarse por escrito con al menos cinco días hábiles de antelación.",
"errors": [
]
}
```
