Closing the Bilingual Gap: Cross-lingual Semantic Scaffolding for Fine-Grained Machine Translation Evaluation in Taiwanese Hokkien

You-Wei Wu[1] **En-Shiun Annie Lee**[2,3] **Richard Tzong-Han Tsai**[1]*

[1]Department of Computer Science and Information Engineering, National Central University, Taiwan
[2]Department of Computer Science, Faculty of Arts and Science, University of Toronto
[3]Computer Science Program, Faculty of Science, Ontario Tech University
thtsai@g.ncu.edu.tw

Abstract

Despite the growing adoption of large language models (LLMs) for machine translation (MT) evaluation, reliably detecting fine-grained error spans in low-resource languages remains challenging. A key limitation is that LLMs often lack sufficient cross-lingual knowledge, making it difficult for them to accurately localize errors at the span level. In this work, we introduce **Cross-lingual Semantic Scaffolding**, a two-part prompting framework that incorporates sentence-level parallel examples and word-level lexical mappings at inference time. We evaluate this approach on five LLMs using controlled ablation experiments on Taiwanese Hokkien. The proposed framework consistently improves performance across all tested models, increasing average Span F1 by $+0.175$ ($+33.6\%$) and average Span Recall by $+0.351$ ($+75.9\%$), and shows that the gains are mainly manifest as improved error coverage. It also supports AI-assisted human annotation, reducing expert annotation time from 450 seconds to 90.92 seconds per sentence, with 98.2% of AI-proposed spans accepted without modification. These results demonstrate that Cross-lingual Semantic Scaffolding is both an effective detection method and a practical tool for reducing expert annotation costs in low-resource MT evaluation.



Figure 1: Overview of the cross-lingual semantic scaffolding pipeline. Given a Mandarin source sentence and a Taiwanese Hokkien machine translation, the method retrieves sentence-level parallel examples and word-level lexical mappings as scaffold context, which is injected into the LLM prompt to predict error spans. The expert post-editing shown here is part of our separate AI-assisted annotation study (Section 6) and is not used to produce the evaluation gold labels used in this paper.

1 Introduction

Machine translation (MT) evaluation has increasingly moved beyond sentence-level scoring toward fine-grained diagnosis of translation errors (Lommel et al., 2013; Kocmi et al., 2024). To reduce the cost of human evaluation, recent work has explored using large language models (LLMs) as automatic evaluators or pre-annotation tools for identifying translation errors (Kocmi and Federmann, 2023b,a; Zouhar et al., 2025). However, automatically detecting error spans remains difficult for low-resource languages like Taiwanese Hokkien.

This paper focuses on MT evaluation for Taiwanese Hokkien, a Sinitic language mainly spoken in Taiwan and in parts of southern China and Southeast Asia (Lu et al., 2024). Although it has a large spoken community, Taiwanese Hokkien remains low-resource in NLP because its written forms have long been less standardized and less consistently used in formal domains (Lu et al., 2024; Chen et al., 2023). This creates a gap between everyday language use and the written resources needed to train and evaluate language technologies.

A further challenge specific to Taiwanese Hokkien lies in its relationship with Mandarin. Two

* Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 664–678
November 6-10, 2026 ©2026 Association for Computational Linguistics

properties make error span detection especially difficult. First, shared Han characters can mask semantic divergence: Taiwanese Hokkien and Mandarin share substantial orthographic overlap, yet the same character string may carry different meanings. For example, 手指 (lit. *hand finger*) means *finger* in Mandarin but *ring* in Taiwanese Hokkien (Lu et al., 2024). Second, LLMs with limited exposure to Taiwanese Hokkien are prone to language confusion (Marchisio et al., 2024), producing outputs that resemble Taiwanese Hokkien on the surface while retaining Mandarin lexical items. For instance, given the Mandarin source 今天天氣很好 (*The weather is nice today*), an erroneous translation may be 今仔日天氣很好. The sentence contains the correct Taiwanese Hokkien expression 今仔日 for 今天 (*today*), but 很 remains a Mandarin adverb; the Taiwanese Hokkien equivalent is 真. This error reflects Mandarin interference rather than a simple mistranslation: the model preserves a Mandarin lexical choice in an otherwise Hokkien-like sentence. Such errors are hard to detect by focusing on the surface characters alone and require explicit knowledge about Mandarin–Hokkien correspondence.

In this work, we address these challenges by providing explicit cross-lingual signals at inference time. We introduce **Cross-lingual Semantic Scaffolding**, a two-part prompting method that enhances LLM-based error span detection by adding supporting cross-lingual context to the prompt. As shown in Figure 1, the method retrieves sentence-level and word-level context before the LLM predicts error spans.[1]

Our contributions are as follows:

1. We introduce Cross-lingual Semantic Scaffolding, which supports LLM-based error span detection with two types of context: (1) sentence-level parallel examples that help the model learn grammatical and semantic correspondences between Mandarin and Taiwanese Hokkien, and (2) word-level lexical mappings that provide explicit Mandarin–Hokkien term correspondences.

2. We show that the proposed method consistently improves performance across five LLMs, with average gains of $+0.175$ in Span F1 ($+33.6\%$) and $+0.351$ in Span Recall ($+75.9\%$), mainly through better error coverage.

3. We find that the two scaffolding layers play complementary roles: sentence-level examples mainly improve error coverage, while word-level mappings strengthen lexical precision. Their combination consistently outperforms either layer alone across all tested models.

4. We demonstrate that performance gains come from semantic relevance rather than bilingual context length alone: random bilingual context yields only small and inconsistent gains, while semantically retrieved context improves average Span F1 by $+0.136$ ($+24.0\%$) and average Span Recall by $+0.278$ ($+52.5\%$).

5. We additionally show that AI-generated pre-annotations support efficient human annotation, reducing expert annotation time from 450 seconds to 90.92 seconds per sentence, while 98.2% of AI-proposed spans are accepted without modification.

2 Methodology

We introduce our method and dataset below.

2.1 Cross-lingual Semantic Scaffolding

As shown in Figure 2, Cross-lingual Semantic Scaffolding is an inference-time prompting strategy that adds explicit Mandarin–Taiwanese Hokkien supporting context to the prompt before the LLM predicts error spans. The term *scaffolding* refers to this added context: it does not change the model parameters, but provides bilingual cues that help the model reason about Mandarin–Taiwanese Hokkien correspondences, compare the source with its translation, and locate possible errors. Our scaffolding has two layers.

Sentence-level Scaffolding Given a Mandarin source sentence, we retrieve the top-ranked semantically similar Mandarin–Taiwanese Hokkien parallel pairs from the sentence-level corpus and prepend them to the prompt as bilingual exemplars, enabling the model to reason about translation patterns by analogy.

Word-level Scaffolding We segment the source sentence using `jieba-zh_TW`[2] and retain tokens

[1] Code, data, prompts, and the annotation interface: https://github.com/350016z/Closing-the-Bilingual-Gap

[2] https://github.com/ldkrsi/jieba-zh_TW

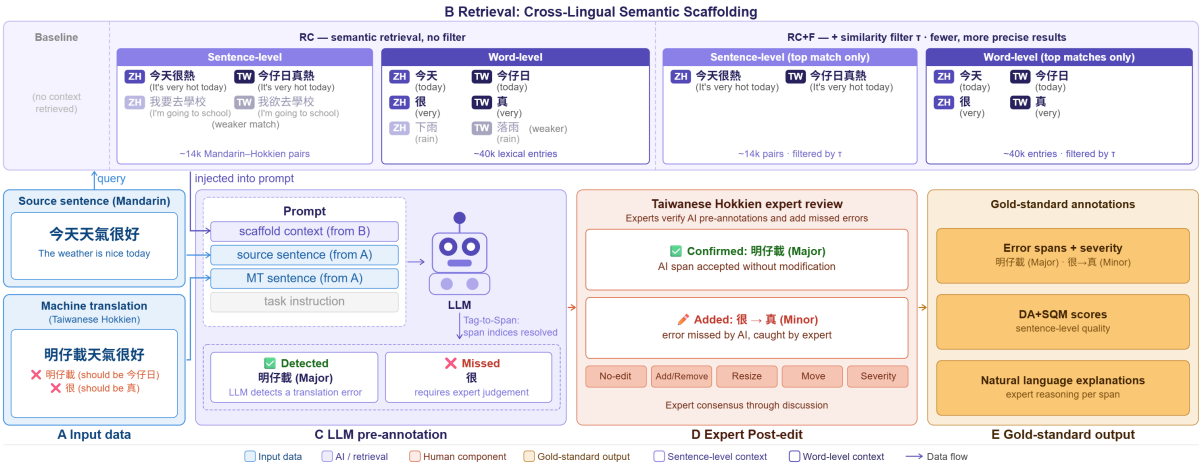


Figure 2: Detailed pipeline of cross-lingual semantic scaffolding and expert post-editing. **(A)** Input: source and MT sentence pair. **(B)** Three retrieval configurations: Baseline uses no external context; RC retrieves related sentence-level examples and word-level mappings, but may include some less relevant matches; RC+F adds a similarity filter (threshold τ) to keep only the closest matches. **(C)** LLM pre-annotation: scaffold context is injected into the prompt for error-span prediction. **(D)** Expert post-editing (part of our separate AI-assisted annotation study, Section 6): Taiwanese Hokkien experts apply five operations (No-edit, Add/Remove, Resize, Move, Severity). **(E)** Annotation output of that study: error spans with severity labels, SQM scores, and natural-language explanations. The evaluation gold labels used for all quantitative results in this paper are not produced by this post-editing pipeline.

of length $\geq$ 2 characters to reduce noise from short function words. For each token, we retrieve the top-ranked semantically similar entry from the word-level corpus, yielding Mandarin–Taiwanese Hokkien lexical correspondences that are injected into the prompt as a lexical reference table, providing explicit semantic anchors for detecting mistranslations arising from cross-lingual divergence. The prompt structure is detailed in Appendix B.

2.2 Dataset

Evaluation Data and Annotation Protocol Our evaluation benchmark consists of 154 Mandarin-to-Taiwanese-Hokkien sentence pairs, together with their human error-span annotations, taken from the Taiwanese Hokkien subset of Liu et al. (2026). We use these expert annotations as the evaluation gold *as released*, with no model involvement in their creation. In that benchmark, the annotation follows an ESA-style protocol (Kocmi et al., 2024) (error spans with Major/Minor severity, rather than the full MQM taxonomy): two expert linguists first calibrated on shared examples and then annotated each sentence to consensus. Our AI-assisted annotation study (Section 6) is a *separate* efficiency experiment that reuses the same sentences to measure annotation effort; its post-edited outputs are never used as evaluation labels. We discuss the implications of the benchmark's consensus-based design in the Limitations section.

Retrieval Corpora The retrieval corpus has two levels, all converted into a unified JSON schema (Appendix G); no filtering is applied at construction time, the similarity filter in Section 3.1 is applied only at retrieval time. For **sentence-level** retrieval (14,243 pairs), we use 11,582 entries from the Ministry of Education (MoE) Taiwanese Dictionary and 2,661 sentence pairs from the TAT Corpus.[3] For **word-level** retrieval (∼40,000 entries), we compile 34,380 entries from the MoE Taiwanese Dictionary, 1,203 recommended-character-form entries,[4] and 2,677 subject-domain terminology entries,[5] and further employ Gemini 2.5 Pro (Comanici et al., 2025) for a semi-automatic lexical expansion that supplies Mandarin correspondences for entries lacking one. This expansion is a *Mandarin-to-Mandarin* extraction step: Gemini extracts Mandarin synonym terms from the existing Mandarin definitions of a dictionary entry (for example, extracting the Mandarin terms 人力、人手 ("manpower; labor force") from the Mandarin definition 人的勞動力或是指可以工作的人。 ("a person's labor capacity, or someone able to work")); it generates no Taiwanese Hokkien content, so concerns about LLM-generated Hokkien quality do not apply to this step.

[3] https://sites.google.com/speech.ntut.edu.tw/fsw/home/tat-corpus
[4] https://language.moe.gov.tw/index
[5] https://mhi.moe.edu.tw/

2.3 Retrieval Encoder

We use `intfloat/multilingual-e5-base` (Wang et al., 2024b) as the semantic retrieval encoder for both scaffolding layers, computing cosine similarity between the source sentence (or segmented token) and corpus entries to rank candidates.

2.4 Tag-to-Span Post-processing

Direct character-index prediction by LLMs is prone to offset and boundary errors, especially given the Han-character overlap between Mandarin and Taiwanese Hokkien. We therefore adopt an inline **Tag-to-Span** mechanism: the model wraps each error span with `<spanN>...</spanN>` tags directly inside the translation string, and a regex-based post-processor strips the tags to recover a clean string and computes character-level start/end indices for each span. This avoids direct index prediction and yields precise, reproducible span boundaries. Full details and the output JSON schema are in Appendix C.

2.5 Model Output Format

For each translation, the model produces three types of output: (1) error spans with inline tagging (i.e., inserting `<spanN>...</spanN>` markers directly into the translation string) and severity labels (Minor/Major); (2) a segment-level quality score on a continuous 0–100 scale, prompted with SQM-style descriptive anchors to stabilize judgment consistency (Freitag et al., 2021; Kocmi and Federmann, 2023b); and (3) a natural language explanation summarizing all flagged error spans collectively, designed to support efficient expert post-editing. The 0–100 SQM score is part of the output format but is not our primary evaluation target: the segment-level correlations (Section 3.4) are computed from the predicted error severities (mapped to MQM scores), not from the raw SQM output.

3 Experimental Setup

3.1 Model Configurations

Five LLMs are included in the evaluation: GPT-5.2 (OpenAI, 2025), Gemini-3-Pro (Google DeepMind, 2025), Qwen3-Max (Yang et al., 2025), DeepSeek-V3.2 (DeepSeek-AI et al., 2024), and Claude Sonnet 4.5 (Anthropic, 2025). Each model is tested under three configurations:

1. **Baseline**: No retrieved context is provided.
2. **Retrieved Context (RC)**: Sentence-level and word-level scaffolding are both injected.
3. **Retrieved Context + Filter (RC+F)**: RC with a similarity filter. Using the encoder in Section 2.3, we compute the cosine similarity between the query and each retrieved candidate and discard candidates below a threshold τ.

Retrieval hyperparameters The similarity thresholds and top-k values are treated as configuration choices selected via the inner loop of nested cross-validation (Appendix E): $\tau_{\text{sent}} \in [0.70, 0.85]$, $\tau_{\text{word}} \in [0.80, 0.95]$, top-$k_{\text{sent}} \in \{3, 5, 7\}$, and top-$k_{\text{word}} \in \{1, 3, 5\}$.

3.2 Rationale for Metric Selection

Our evaluation framework targets two complementary dimensions: span-level diagnostic accuracy and segment-level alignment with human quality judgments. We do not use traditional surface overlap metrics (BLEU, Papineni et al. 2002; chrF, Popović 2015) and supervised neural evaluators (xCOMET, Guerreiro et al. 2024; MetricX, Juraska et al. 2023) as baselines, for the following reasons:

- **Absence of diagnostic localization**: BLEU and chrF measure lexical overlap against a reference and cannot provide the error location information required for error span detection.
- **Fundamental language coverage limitations**: The MQM fine-tuning data for xCOMET and MetricX covers only a small set of high-resource language pairs (e.g., Mandarin–English, English–German, English–Russian) (Guerreiro et al., 2024; Juraska et al., 2023). Taiwanese Hokkien is absent from both fine-tuning data and, to a large extent, from the underlying pre-training corpora, making these metrics unreliable for our evaluation setting.

We therefore use the Baseline LLM configuration (no retrieved context) as our control, isolating semantic scaffolding as the sole experimental variable. Span-level metrics and segment-level correlation analyses serve as the primary evaluation criteria.

To mitigate overfitting on our 154-instance dataset, we employ repeated stratified nested cross-validation (5-fold outer, 4-fold inner, 10 repeats); details are provided in Appendix E.

3.3 Span-level Metrics

Following Fernandes et al. (2023) and Huang et al. (2024), we adopt a token-set overlap framework: model-predicted span tokens are compared against gold span tokens to compute Precision, Recall, and F1. Let G denote the union of gold span tokens and P the union of predicted span tokens:

$$\text{SP} = \frac{|P \cap G|}{|P|} \tag{1}$$

$$\text{SR} = \frac{|P \cap G|}{|G|} \tag{2}$$

$$\text{SF1} = \frac{2 \cdot \text{SP} \cdot \text{SR}}{\text{SP} + \text{SR}} \tag{3}$$

For translation-major errors, we additionally define Major-only metrics (MP, MR, MF1), computed exclusively over spans labeled as Major in the gold annotation, measuring the model's ability to capture semantically significant errors.

3.4 Segment-level Correlation Evaluation

To assess whether model annotations align with expert quality judgments at the segment level, we adopt the WMT-standard annotation mapping and conversion pipeline: (1) gold and predicted annotations are mapped to numeric MQM scores using standardized severity weights (Minor $= 1$, Major $= 5$) established by Freitag et al. (2021) and adopted in WMT human evaluation (Kocmi et al., 2024); (2) we use Google's `mt-metrics-eval`[6] toolkit to convert fine-grained annotations to segment-level MQM scores; (3) we compute Pearson's r (linear correlation), Spearman's ρ (rank correlation), and Kendall's τ (ordinal concordance) between model-derived and expert gold scores.

4 Results and Discussion

4.1 Span-level Performance

Tables 1 and 2 present span-level performance across all five models under each configuration. Introducing cross-lingual semantic scaffolding yields consistent improvements in fine-grained detection capability across all tested models.

4.2 Key Findings and Analysis

Dramatic Improvement in Span Recall The most consistent contribution of semantic scaffolding is a substantial increase in error coverage (span recall and major-span recall). Claude Sonnet 4.5, for instance, achieves a span recall of 0.8861 under RC, up from 0.2895 in the Baseline configuration. We interpret this cautiously. While the pattern is consistent with improved cross-lingual grounding, our experiments cannot exclusively attribute the gains to cross-lingual knowledge transfer, as opposed to lexical priming, the retrieval of reference-like targets, or a general tendency to propose more spans. We therefore report the observed effect, namely, higher recall at a small precision cost, and treat the underlying mechanism as an association rather than an established cause.

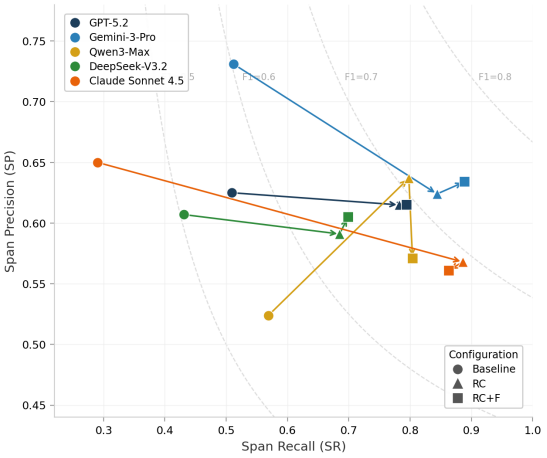


Figure 3: Span Precision (SP) vs. Span Recall (SR) for all five models under three configurations. Each color represents one model; Arrows show the trajectory Baseline→RC→RC+F. Dashed curves are iso-F1 contours at $F1 \in \{0.5, 0.6, 0.7, 0.8\}$. Baseline: no scaffolding; RC: word- and sentence-level retrieval combined; RC+F: RC with similarity-threshold filtering (τ).

Precision–Recall Trade-off and the Role of Filtering As shown in Figure 3, the precision drop under RC relative to Baseline is small but consistent: for example, GPT-5.2 SP falls from 0.625 to 0.615, and Gemini-3-Pro from 0.731 to 0.624. This is a direct consequence of how the scaffolding works: by providing richer bilingual context, the model becomes less conservative and flags more spans, which drives recall up but allows some false positives through. RC+F addresses this by applying a similarity threshold τ to remove retrieved entries with low relevance scores, reducing noise in the scaffold. For GPT-5.2 and Gemini-3-Pro, RC+F recovers precision relative to RC while maintaining or further improving recall. Qwen3-Max is an exception (its SF1 drops under RC+F), suggesting that for some models, broader context including

[6] https://github.com/google-research/mt-metrics-eval

Model	Setting	Span-Level			Major Span-Level			Segment-Level Corr.		
		Precision	Recall	F1	Precision	Recall	F1	Pearson r	Spearman ρ	Kendall τ
GPT-5.2	Base	.625±.045	.509±.067	.558±.047	.526±.078	.505±.085	.510±.063	.436±.136	.391±.139	.299±.109
	RC	.615±.053	.782±.045	.686±.035	.498±.076	.619±.088	.548±.065	.448±.156	.440±.154	.337±.126
	RC+F	.615±.053	.794±.040	.691±.038	.511±.084	.594±.075	.546±.070	.429±.183	.429±.154	.327±.127
Gemini-3-Pro	Base	**.731±.046**	.512±.067	.599±.050	**.620±.102**	.464±.093	.526±.088	.479±.106	.444±.129	.335±.103
	RC	.624±.060	.844±.047	.715±.041	.469±.097	.702±.092	**.556±.084**	.524±.156	.497±.142	.390±.120
	RC+F	.634±.051	**.889±.040**	**.739±.039**	.431±.087	.678±.075	.521±.071	**.558±.157**	**.542±.130**	**.429±.115**
Qwen3-Max	Base	.524±.048*	.569±.051	.543±.034	.374±.066	.641±.084	.468±.059	.332±.181	.294±.162	.223±.125
	RC	.637±.050	.798±.036	.707±.034	.393±.070	.709±.071	.501±.060	.496±.156	.476±.119	.355±.095
	RC+F	.571±.048	.804±.068	.665±.040	.365±.064	.741±.063	.485±.055	.315±.168	.332±.165	.254±.127
DeepSeek-V3.2	Base	.607±.058	.431±.065	.502±.053	.437±.101	.428±.066	.428±.076*	.366±.166	.392±.152	.299±.109
	RC	.591±.057	.685±.066	.632±.045	.506±.082	.557±.082	.527±.068	.374±.172	.331±.152	.259±.120
	RC+F	.605±.054	.699±.069	.646±.047	.515±.086	.546±.087	.523±.062	.416±.141	.475±.124	.360±.098
Claude Sonnet 4.5	Base	.650±.074	.290±.055*	.398±.060*	.572±.109	.364±.092*	.438±.084	.291±.242*	.224±.202*	.174±.165*
	RC	.568±.052	.886±.031	.690±.038	.362±.070	**.761±.065**	.486±.067	.462±.111	.528±.106	.398±.085
	RC+F	.561±.046	.863±.043	.678±.037	.361±.071*	.717±.063	.476±.065	.402±.144	.401±.146	.302±.112

Table 1: Model performance on span-level diagnostic metrics and correlation coefficients with segment-level human MQM scores. RC: Retrieved Context; RC+F: Retrieved Context + Filter. **Bold**: best overall per metric; underline: second best; *: worst.

Model	Best Setting	Span F1 Score			Span Recall		
		Base	Best	Δ [95% CI]	Base	Best	Δ [95% CI]
GPT-5.2	RC+F	.558	.691	+.133 [.118, .148]	.509	.794	+.285 [.259, .311]
Gemini-3-Pro	RC+F	.599	**.739**	+.139 [.125, .153]	.512	**.889**	+.377 [.350, .403]
Qwen3-Max	RC	.543	.707	+.164 [.151, .178]	.569	.798	+.229 [.205, .253]
DeepSeek-V3.2	RC+F	.502	.646	+.145 [.128, .162]	.431	.699	+.268 [.241, .295]
Claude Sonnet 4.5	RC	.398	.690	+.292 [.271, .314]	.290	.886	+.597 [.567, .626]

Table 2: Per-model performance summary (Baseline → best configuration). Δ values report absolute gains; 95% CIs are computed as $\bar{x} \pm 1.96\,\sigma/\sqrt{50}$ over 50 outer cross-validation folds. **Bold**: highest Best score per metric. Underline: largest absolute gain per metric.

weaker matches is beneficial rather than noisy.

Segment-level Alignment As shown in the rightmost columns of Table 1, semantic scaffolding not only improves span-level detection performance, but also improves segment-level correlation with expert human ratings for most models. Gemini-3-Pro achieves the highest Pearson $r = 0.5579$ under RC+F, a notable improvement over its Baseline configuration ($r = 0.4790$). GPT-5.2 is an exception under RC+F ($\Delta r = -0.007$), though it shows a small positive gain under RC ($\Delta r = +0.012$), suggesting sensitivity to retrieval noise at the segment level.

The moderate level of segment-level correlation (Pearson r ranging from 0.29–0.56) warrants careful interpretation. Span-level annotation and segment-level scoring inherently capture different quality dimensions: the former focuses on the existence and location of local errors, while the latter attempts to quantify translation quality by linearly aggregating span-level penalties. An information gap between the two dimensions is therefore expected, even for human annotators. Additionally, pragmatic properties of Taiwanese Hokkien (e.g., sentence-final particles, aspect markers) may introduce subjective perceptual judgments into expert scores that fall outside the MQM conversion framework's capture range. We therefore recommend treating segment-level correlation as a supplementary consistency check rather than the primary performance criterion for error span detection.

5 Ablation Study

We decompose the scaffolding framework to quantify the independent contributions of sentence-level (Sent-only) and word-level (Word-only) information and verify that semantic relevance (not context length) drives performance gains. We compare six configurations: Baseline, Random (length-controlled baseline with randomly sampled bilingual context), Word-only, Sent-only, RC, and RC+F. Ablation results for three representative high-performing models selected from the main evaluation are reported in Table 3, with detailed discussion in Appendix D.

Model	Setting	Span-Level			Major Span-Level			Segment-Level Corr.		
		Precision	Recall	F1	Precision	Recall	F1	Pearson r	Spearman ρ	Kendall τ
GPT-5.2	Base	.625±.045	.509±.067	.558±.047	.526±.078	.505±.085	.510±.063	.436±.136	.391±.139	.299±.109
	Random	.589±.043	.487±.061*	.531±.042*	.529±.079	.485±.085	.502±.071	.445±.173	.403±.153	.313±.121
	Word-only	.589±.043	.610±.060	.597±.037	.499±.081	.564±.086	.526±.073	.412±.164	.401±.147	.305±.115
	Sent-only	.608±.048	.769±.044	.677±.031	.518±.085	.584±.092	.544±.073	.528±.150	.512±.131	.395±.107
	RC	.615±.053	.782±.045	.686±.035	.498±.076	.619±.088	.548±.065	.448±.156	.440±.154	.337±.126
	RC+F	.615±.053	.794±.040	.691±.038	.511±.084	.594±.075	.546±.070	.429±.183	.429±.154	.327±.127
Gemini-3-Pro	Base	.731±.046	.512±.067	.599±.050	**.620±.102**	.464±.093*	.526±.088	.479±.106	.444±.129	.335±.103
	Random	**.741±.056**	.629±.055	.679±.046	.557±.105	.654±.086	.595±.077	**.604±.120**	**.607±.115**	**.477±.101**
	Word-only	.706±.055	.651±.061	.675±.046	.583±.097	.670±.083	**.618±.071**	.559±.138	.559±.115	.439±.102
	Sent-only	.575±.058	**.931±.026**	.709±.044	.375±.071	**.798±.072**	.506±.070	.597±.119	.591±.123	.464±.107
	RC	.624±.060	.844±.047	.715±.041	.469±.097	.702±.092	.556±.084	.524±.156	.497±.142	.390±.120
	RC+F	.634±.051	.889±.040	**.739±.039**	.431±.087	.678±.075	.521±.071	.558±.157	.542±.130	.429±.115
Qwen3-Max	Base	.524±.048	.569±.051	.543±.034	.374±.066	.641±.084	.468±.059	.332±.181	.294±.162*	.223±.125*
	Random	.544±.051	.570±.058	.554±.039	.348±.063	.652±.088	.450±.061*	.361±.190	.300±.175	.231±.137
	Word-only	.515±.049*	.610±.056	.556±.036	.351±.061	.663±.097	.455±.061	.322±.219	.317±.188	.236±.143
	Sent-only	.549±.048	.729±.048	.625±.034	.331±.055*	.733±.064	.452±.052	.350±.179	.331±.160	.250±.122
	RC	.637±.050	.798±.036	.707±.034	.393±.070	.709±.071	.501±.060	.496±.156	.476±.119	.355±.095
	RC+F	.571±.048	.804±.068	.665±.040	.365±.064	.741±.063	.485±.055	.315±.168*	.332±.165	.254±.127

Table 3: Ablation results on span-level metrics and segment-level correlation across scaffolding configurations for three representative high-performing models. Random: randomly sampled bilingual context (length-controlled baseline); Word-only / Sent-only: single-layer scaffolding; RC: both layers combined; RC+F: RC with similarity filtering. **Bold**: best overall per metric; underline: second best; *: worst. Metric abbreviations follow Table 1.

Figure 4 visualizes SR and SF1 trends across all configurations. The Random Context baseline confirms that indiscriminate context addition does not reliably improve performance: for GPT-5.2, Random (SF1 $= 0.5309$) falls below Baseline (SF1 $= 0.5582$), suggesting that semantically irrelevant context can mildly disrupt reasoning, while RC reaches SF1 $= 0.6864$. The gap between Random and RC confirms that content relevance, not context length, accounts for the improvement.

Word-level scaffolding improves lexical precision by providing explicit cross-lingual correspondences, while sentence-level scaffolding yields larger gains in span recall by supporting global semantic and syntactic reasoning. For GPT-5.2, Sent-only raises SF1 by $+21.2\%$ over Baseline. The full dual-layer RC configuration consistently outperforms either single-layer variant, with Qwen3-Max RC reaching SF1 $= 0.7070$ versus Word-only (0.5561) and Sent-only (0.6245), confirming strong complementarity between the two components.

The effect of similarity-based filtering is heterogeneous: GPT-5.2 and Gemini-3-Pro benefit from noise reduction under RC+F, while Qwen3-Max shows a clear decline (SF1 from 0.7070 to 0.6651), suggesting that some models exploit broader, weakly aligned signals. Filtering should therefore be tuned to model-specific characteristics rather than applied uniformly.

These findings extend prior work on low-resource in-context learning (Li et al., 2025; Cahyawijaya et al., 2024): the Random baseline confirms that cross-lingual relevance (not context length) is the operative factor, consistent with the finding that alignment-based prompting is especially effective where standard fine-tuning is unreliable. Our results add controlled span-level evidence for that claim in a previously untested low-resource language, and suggest that the gain comes from the cross-lingual *content* of the scaffold, not from having additional context alone.

6 AI-Assisted Human Annotation

Following recent work on AI-assisted human evaluation of MT (Zouhar et al., 2025), We evaluate annotation efficiency using a paired experimental design in which the same experts annotated identical data with and without AI assistance (using pre-annotations from Gemini-3-Pro RC+F, the best-performing configuration by SF1). As shown in Table 4, AI-assisted annotation reduces average annotation time by nearly $5\times$ (from 450 to 90.92 seconds per sentence), shifting the annotation paradigm from exhaustive error search to rapid verification and reducing cognitive load while preserving expert judgment.

Annotators and blinding The two annotators are native Taiwanese Hokkien speakers and Taiwanese Hokkien teachers, recruited through the authors' academic network, and followed a written guideline based on the ESA protocol (Kocmi et al., 2024). As this is a post-editing study, full blinding

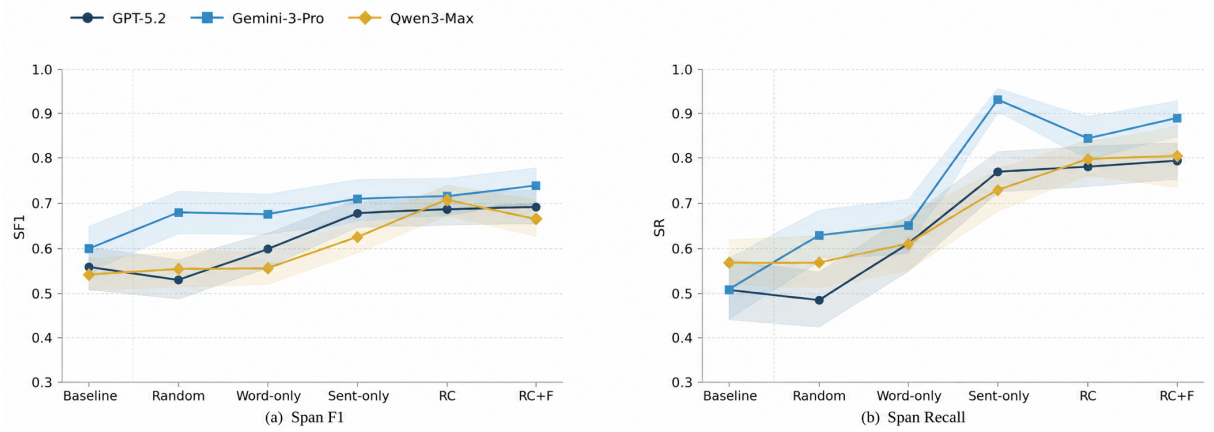


Figure 4: Ablation results across six configurations for three representative models. **(a)** Span F1 (SF1) and **(b)** Span Recall (SR), each shown as a line chart; the dotted vertical line separates the no-scaffolding Baseline from scaffolding configurations, and shaded bands show ± 1 standard deviation across 50 folds. Baseline: no scaffolding; Random: randomly sampled bilingual context (length-controlled); Word-only / Sent-only: single-layer scaffolding; RC: both layers combined; RC+F: RC with similarity filtering.

Metric	Manual	AI-Assisted
Avg. Time per Sentence (s)	450.00	**90.92**
Annotation Paradigm	Search	**Verify**
Cognitive Load	High	**Low**

Table 4: Efficiency comparison between manual and AI-assisted annotation. AI assistance shifts the annotation workflow from error searching to error verification.

Action Type	Count	Percentage
No-edit (Adopted)	436	98.20%
Add	17	3.83%
Remove	9	2.03%
Resize	7	1.58%
Reset	5	1.13%
Move	1	0.23%
Severity Modification	0	0.00%
Total Initial Errors	**444**	

Table 5: Distribution of expert post-editing actions on AI-generated pre-annotations. The high adoption rate (98.2%) underlines the practical reliability of the system. Percentages are computed independently for each action type and therefore do not sum to 100%.

is not feasible, since annotators necessarily see the AI-proposed spans, which we note as a limitation; the evaluation gold itself (Section 2.2) involves no model output and is therefore unaffected.

Table 5 shows that 98.20% of AI pre-annotated spans were directly adopted without modification. Resize (1.58%) and Move (0.23%) rates are extremely low, reflecting two complementary contributions: Tag-to-Span post-processing ensures precise boundary extraction, while word-level scaffolding improves the semantic precision of initial span localization, reducing the need for boundary adjustment before expert review. Critically, the Severity Modification rate is 0.00%, providing a stable foundation for downstream MQM score conversion. These results suggest an efficient division of cognitive labor: AI handles high-volume error localization while experts focus on pragmatic refinement, with system-generated explanations enabling rapid logical alignment without interrupting the annotation workflow. We acknowledge that the high adoption rate (98.20%) may partly reflect anchoring effects: experts reviewing pre-generated spans may be more likely to accept them than to independently produce identical annotations. While the paired design ensures that the same experts perform both manual and AI-assisted annotation, it does not fully isolate anchoring from genuine pre-annotation accuracy. A more rigorous future evaluation would include a held-out subset with deliberately perturbed AI suggestions to quantify how readily experts override incorrect pre-annotations. A detailed description of the annotation interface is provided in Appendix F.

7 Related Work

Prior work has shown that large language models can serve as machine translation evaluators, from system-level scoring with GEMBA (Kocmi and Federmann, 2023b) to MQM-style error span prediction with GEMBA-MQM (Kocmi and Federmann, 2023a). Other studies have examined the effects of prompting, in-context learning, and fine-

tuning on LLM-based machine translation error analysis (Fernandes et al., 2023). Another relevant study is COMET-poly (Züfle et al., 2025), which incorporates additional candidate translations or retrieved examples to produce segment-level quality scores. More broadly, retrieval and in-context learning have been explored for low-resource language processing and machine translation, including retrieval-augmented prompting with parallel sentences and lexical resources (Cahyawijaya et al., 2024; Li et al., 2025; Merx et al., 2024). Unlike prior work that primarily focuses on translation quality scoring, generation, or general language understanding, we study whether such context can improve LLM-based error span detection for Mandarin–Taiwanese Hokkien MT evaluation.

8 Conclusion

We presented **Cross-lingual Semantic Scaffolding**, an inference-time prompting method for LLM-based error span detection in low-resource MT evaluation. Applied to Mandarin–Taiwanese Hokkien MT evaluation, the method consistently improves performance across five LLMs, increasing average Span F1 by $+0.175$ ($+33.6\%$) and average Span Recall by $+0.351$ ($+75.9\%$). These gains mainly come from improved error coverage, showing that explicit cross-lingual context helps LLMs find error spans that are often missed without scaffolding.

Ablation results show that the gains come from semantic relevance rather than bilingual text length alone. Random bilingual context produces small and inconsistent improvements, while semantically retrieved context yields much larger gains. Sentence-level examples and word-level mappings provide complementary support, and similarity filtering helps some models but not others, suggesting that context filtering should be tuned by model.

The method also supports AI-assisted human annotation. AI-generated pre-annotations reduce expert annotation time from 450 to 90.92 seconds per sentence (-79.8%), with 98.2% of proposed spans accepted without modification. These findings suggest that cross-lingual scaffolding can improve low-resource error span detection while also reducing the cost of expert annotation. Future work will extend this approach to more genres, systems, and low-resource language pairs.

Limitations

We summarize the primary limitations of our study below:

1. **Data scale and source diversity**: Our gold span dataset consists of 154 sentence pairs, sufficient for repeated nested cross-validation and AI-assisted annotation analysis, but limited in scale. The predominance of dictionary-sourced sentences may not represent error distributions in other genres. Incorporating news, conversational, and spontaneous speech text would strengthen the framework's robustness across domains.

2. **Language scope and Han-character-centric design**: Our experiments are limited to Mandarin-to-Taiwanese-Hokkien translation, where the source and target share substantial Han-character orthographic overlap. The effectiveness of cross-lingual scaffolding on language pairs without shared writing systems (e.g., Mandarin-to-Indonesian, Mandarin-to-Vietnamese) remains an important direction for future work to verify cross-linguistic generalizability. In addition, the benchmark provides annotations only for the Mandarin-to-Taiwanese-Hokkien direction, so our study is limited to this direction by data availability rather than by method.

3. **Consensus-based annotation without independent IAA**: Due to the high cost of recruiting expert annotators for Taiwanese Hokkien, our annotation protocol relies on two-annotator consensus through discussion rather than independent double annotation. While the consensus mechanism ensures that the final gold labels reflect agreement between two experts, it precludes the calculation of inter-annotator agreement (IAA) statistics in the conventional sense, since the annotators were not independent. Such pragmatic adaptations to standard MQM annotation protocols are common in low-resource language evaluation under expert scarcity: prior work has likewise adopted simplified MQM guidelines for African languages (Wang et al., 2024a) or relied on consensus-based annotation for Sinitic languages (Liu et al., 2026). We acknowledge that quantitative IAA measurement would nevertheless strengthen future iterations of this benchmark.

4. **Retrieval resource dependency**: The framework's performance is directly contingent on the relevance and coverage of external retrieval resources relative to the evaluation data. In languages with sparser lexical or parallel resources than Taiwanese Hokkien, scaffold benefits may diminish. Practitioners should carefully assess external resource coverage before applying this framework to other low-resource languages, as insufficient coverage represents a core risk. Relatedly, the retrieval corpora and the evaluation sentences partly rely on some of the same dictionary sources, so limited lexical overlap between retrieved context and gold references cannot be fully ruled out; disentangling this from genuine cross-lingual grounding is left to future work.

References

Anthropic. 2025. Claude Sonnet 4.5 system card. Technical report, Anthropic.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, Hirofumi Inaguma, Sravya Popuri, Changhan Wang, Juan Pino, Wei-Ning Hsu, and Ann Lee. 2023. Speech-to-speech translation for a real-world unwritten language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4969–4983, Toronto, Canada. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. 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. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Google DeepMind. 2025. Gemini 3 pro model card. Technical report, Google DeepMind.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. In *Transactions of the Association for Computational Linguistics*, volume 12, pages 979–995, Cambridge, MA. MIT Press.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the source language: How large language models evaluate the quality of machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3546–3562, Bangkok, Thailand. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

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

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

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja

Popovic, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yue Li, Zhixue Zhao, and Carolina Scarton. 2025. It's all about in-context learning! teaching extremely low-resource languages to LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29544–29559.

Hannah Liu, Junghyun Min, En-Shiun Annie Lee, Ethan Yue Heng Cheung, Shou-Yi Hung, Elsie Chan, Shiyao Qian, Runtong Liang, Kimlan Huynh, Wing Yu Yip, York Hay Ng, Tsz Fung Yau, Ka Ieng Charlotte Lo, You-Wei Wu, and Richard Tzong-Han Tsai. 2026. SiniticMTError: A machine translation dataset with error annotations for sinitic languages. *Preprint*, arXiv:2509.20557.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*.

Bo-Han Lu, Yi-Hsuan Lin, Annie Lee, and Richard Tzong-Han Tsai. 2024. Enhancing Taiwanese Hokkien dual translation by exploring and standardizing of four writing systems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6077–6090, Torino, Italia. ELRA and ICCL.

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. Association for Computational Linguistics.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

OpenAI. 2025. Update to GPT-5 system card: GPT-5.2. Technical report, OpenAI.

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

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, and 39 others. 2024a. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual E5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. COMET-poly: Machine translation metric grounded in other candidates. In *Proceedings of the Tenth Conference on Machine Translation*, pages 887–904, Suzhou, China. Association for Computational Linguistics.

A Background and Related Work

A.1 Fine-grained Machine Translation Diagnostics and Evaluation

Prior work on fine-grained Machine Translation (MT) evaluation has converged on MQM-based frameworks (Lommel et al., 2013) and the more annotation-efficient ESA protocol (Kocmi et al., 2024). A parallel line of work explores using LLMs themselves as MT evaluators. Kocmi and Federmann (2023b) introduce GEMBA, showing that prompting large generative models can achieve strong system-level MT evaluation performance. Kocmi and Federmann (2023a) extend this to fine-grained diagnostics with GEMBA-MQM, demonstrating that a fixed three-shot, language-agnostic prompt can elicit MQM-style error spans in a reference-free setting. Concurrently, Fernandes et al. (2023) investigate the complementary question of how in-context demonstrations and fine-tuning shape LLMs' ability to identify and categorize translation errors. More recent work observes that LLM-predicted spans, while useful for system-level ranking, often diverge from human annotations, and proposes post-editing-based filtering to improve span reliability (Lu et al., 2025).

These studies demonstrate the promise of LLM-based fine-grained MT evaluation, but do not explicitly address the cross-lingual grounding problem that arises in low-resource settings such as Taiwanese Hokkien, where pre-trained bilingual knowledge is sparse. Our work addresses this gap by investigating whether sentence-level parallel examples and word-level lexical mappings can serve as inference-time cross-lingual scaffolds for LLM-based fine-grained error span detection.

A.2 Taiwanese Hokkien NLP and Low-Resource MT Evaluation

Taiwanese Hokkien is a widely spoken but under-resourced language variety in Taiwan. Recent NLP work has begun to explore translation and writing-system standardization for Taiwanese Hokkien; for example, Lu et al. (2024) develop dual translation between Taiwanese Hokkien and Mandarin/English and study the role of orthographic similarity across writing systems. Liu et al. (2026) introduced fine-grained error annotations for Sinitic languages, including Taiwanese Hokkien as one language variety. Building on this resource, our work conducts the first dedicated investigation into LLM-based fine-grained error span detection for Mandarin-to-Taiwanese-Hokkien MT.

A.3 In-Context Learning and Retrieval Augmentation for Low-Resource Languages

Recent work suggests that in-context learning (ICL; Dong et al., 2024) and retrieval-based support can be especially valuable for low-resource languages. Li et al. (2025) show that language alignment signals are particularly effective in extremely low-resource settings, where standard fine-tuning is often less reliable. Cahyawijaya et al. (2024) likewise show that semantically relevant in-context information and cross-lingual query alignment improve low-resource language understanding. In the MT domain, Merx et al. (2024) demonstrate that prompting with retrieved parallel sentences and dictionary entries can improve translation quality for Mambai. However, prior work mainly focuses on generation quality or general task performance, rather than on fine-grained error localization. Building on these findings and the general paradigm of Retrieval-Augmented Generation (RAG; Lewis et al., 2020), we investigate a dual-layer semantic scaffold that combines sentence-level parallel retrieval with word-level lexical mappings for low-resource diagnostic evaluation.

A.4 AI-Assisted Annotation

Zouhar et al. (2025) show that AI-assisted human evaluation can substantially reduce annotation time while maintaining comparable quality. We extend this paradigm to the low-resource setting via cross-lingual semantic scaffolding, achieving an approximately $5\times$ reduction in expert annotation time and a 98.2% span adoption rate.

Appendix

B Prompt Structure for Cross-lingual Scaffolding

The scaffolding context is appended to the user-facing prompt as a structured *Alignment Hints* block, divided into two sections corresponding to the two scaffolding layers. To prevent the model from inserting error span tags within the hints block, an explicit instruction is prepended stating that the hints are for reference only and that all annotation rules remain in effect.

Alignment Hints

[Reference only; do not insert <spanN> tags here.]
Sentence-level alignment (few-shot, zh↔tai)

[Mandarin: {M_1}
Taiwanese (HAN): {T_1_han}
Taiwanese (TL, optional): {T_1_tl},
Mandarin: {M_2}
Taiwanese (HAN): {T_2_han}
Taiwanese (TL, optional): {T_2_tl},...]

Word-level alignment (lexical hints, zh→tai)

[{zh_1} = tai(HAN: {han_1}; TL: {tl_1}, optional),
{zh_2} = tai(HAN: {han_2}; TL: {tl_2}, optional),...]

Multiple sentence-level pairs and word-level entries are concatenated within their respective sections and together form the complete Alignment Hints block appended to the prompt.

C Implementation Details of Tag-to-Span Mechanism

Direct character index prediction by LLMs is prone to severe offset and boundary inconsistencies, especially in the Han-character overlap between Mandarin-to-Taiwanese-Hokkien translation. We address this with an inline tagging (Tag-to-Span) mechanism. The model is required to wrap error spans with <spanN>...</spanN> tags directly within the translation string and return the following JSON structure:

```
{
 "errors_in_mt": "...<span1>error</span1>...",
 "errors": [
   {
     "error_id": 1,
     "error_text": "<span1>",
     "severity": "<Major|Minor>"
   },...
 ]
}
```

Tagged spans are then parsed by a regex-based post-processor: tags are stripped to obtain a clean translation string, and precise character-level start and end indices of each span are computed with respect to this tag-stripped string. All error spans are finally consolidated into a standardized format suitable for span-level evaluation.

D Extended Ablation Analysis

This appendix provides per-model analyses for the three representative models used in the ablation study, complementing the summary in Section 5. Figure 5 gives a heatmap overview of SF1 across all ablation configurations for these models, followed by model-specific observations not accommodated in the main text.

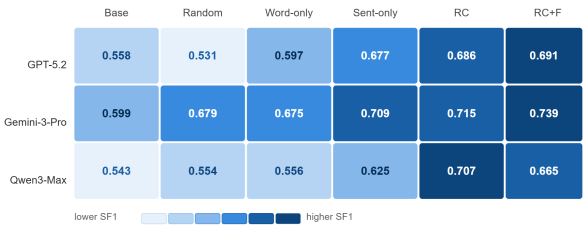


Figure 5: SF1 heatmap across ablation configurations (columns) and models (rows). Darker green indicates higher SF1.

D.1 Random Context Control: Verifying Semantic Relevance

A fundamental question for any retrieval-augmented system is whether performance gains stem from semantic relevance or simply from increased context length. Our Random Context baseline isolates this factor by injecting randomly sampled bilingual context without semantic alignment to the input. For GPT-5.2, Random Context (SF1 = 0.5309) does not surpass Baseline (SF1 = 0.5582) and in fact shows a slight decrease, suggesting that semantically irrelevant context mildly disrupts reasoning. By contrast, RC achieves SF1 = 0.6864, a +0.1555 margin over Random, confirming that semantic alignment is a primary driver of observed gains. Qwen3-Max shows a similar pattern: Random (SF1 = 0.5541) barely exceeds Baseline (SF1 = 0.5427), while RC achieves 0.7070 (+0.1529 over Random).

Gemini-3-Pro presents an interesting exception: its Random Context performance (SF1 = 0.6788) substantially exceeds Baseline (SF1 = 0.5992) and approaches Word-only (SF1 = 0.6751), suggesting that this model may integrate contextual structure more broadly regardless of semantic alignment.

D.2 Word-level Scaffolding: Lexical Precision

Word-level scaffolding consistently improves span-level performance relative to Baseline across all models, while generally underperforming relative to Sent-only. These results are consistent with the hypothesis that word-level mappings function as semantic anchors: by providing explicit Mandarin-to-Taiwanese-Hokkien correspondences, the model can more reliably adjudicate whether a given Han character sequence constitutes a mistranslation or a terminological inconsistency. Word-level information also shows meaningful contributions to

segment-level Pearson r, suggesting that precise lexical grounding helps calibrate holistic quality scoring.

D.3 Sentence-level Scaffolding: Global Recall

Relative to word-level scaffolding, sentence-level context produces larger performance jumps, particularly in span recall. For GPT-5.2, Sent-only raises SF1 from 0.5582 (Baseline) to 0.6769 (+21.2%), substantially exceeding the Word-only contribution. Parallel sentences function as translation exemplars, enabling the model to reason analogically about Taiwanese Hokkien syntactic structures and pragmatic conventions. This global semantic scaffolding is particularly decisive for capturing grammar-level and fluency-level errors that require understanding of sentence-wide structure rather than isolated lexical properties.

D.4 Dual-layer Scaffolding Synergy

The full dual-layer RC configuration consistently outperforms either single-layer variant across all three models. For Qwen3-Max, RC achieves SF1 = 0.7070, substantially exceeding both Word-only (0.5561) and Sent-only (0.6245). The combined effect reflects complementary mechanisms: sentence-level scaffolding provides macro-level syntactic guidance, while word-level scaffolding supplies micro-level lexical constraints. Together, they target complementary error types, covering both lexical and structural dimensions of Taiwanese Hokkien translation.

D.5 Heterogeneous Effects of Similarity Filtering

Similarity-based filtering affects different model architectures in opposing directions. GPT-5.2 and Gemini-3-Pro benefit from filtering (SF1 improves under RC+F), indicating sensitivity to retrieval noise: removing low-similarity context helps focus these models on the most relevant evidence. Qwen3-Max, by contrast, shows a clear decline from RC (SF1 = 0.7070) to RC+F (0.6651), with a particularly large drop in Pearson r, suggesting a broader context integration strategy that exploits even weakly aligned signals. These results indicate that filtering should be tuned to the target model's architecture rather than applied uniformly.

E Repeated Stratified Nested Cross-Validation

To mitigate the limitations imposed by the relatively small dataset (154 instances) and reduce the risk of overfitting, we employ repeated stratified nested cross-validation (5-fold outer, 4-fold inner, 10 repeats). Retrieval hyperparameters (top-k and similarity threshold for each scaffolding layer) are treated as discrete configuration choices: predictions are generated offline for each candidate configuration, and the inner loop selects the best-performing configuration on held-out validation folds. The outer loop then evaluates generalization on held-out test folds, ensuring that performance gains are not attributable to configuration overfitting.

F Annotation Interface Details

The annotation interface is shown in Figure 6.



Figure 6: The annotation interface showing bilingual display panel, model-predicted error spans, diagnostic explanations, editing tools, and quality score slider.

The annotation interface integrates four core functional components: (1) a bilingual display panel showing the Mandarin source and Taiwanese Hokkien translation side by side, with model-predicted error spans highlighted; (2) a diagnostic explanation panel presenting the model's error reasoning, enabling experts to assess error properties without consulting external dictionaries; (3) dynamic editing tools (resize, move, and severity selection buttons) that transform tedious character selection into intuitive button operations; and (4) a quality score slider for segment-level assessment.

Post-editing operations Each operation acts on an AI-proposed span: **No-edit** accepts a span unchanged; **Add** marks a new error span the model did not propose; **Remove** deletes an AI-proposed span (e.g., a false positive); **Resize** adjusts a span's start/end boundaries so it covers exactly the erroneous tokens, without relocating it; **Move** shifts

Top-level category	Major	Minor	Total
Accuracy	94	59	153
Fluency	11	48	59
Purity	22	14	36
Locale	2	29	31
Terminology	15	3	18
Style	1	4	5
Other	0	4	4
Total	**145**	**161**	**306**

Table 6: Error-type distribution over the 154-sentence evaluation set. Reproduced from Liu et al. (2026).

a span to a different position while keeping the span itself; **Reset** reverts a span to its original AI-proposed state, undoing manual edits; and **Severity** changes a span's label between Major and Minor.

G Retrieval Corpus Schema and Examples

Each entry, at both the word and sentence level, uses a unified JSON schema with the fields 台文 (Taiwanese Hokkien in Han script), 中文 (the Mandarin correspondence), and 台羅 (Tâi-lô romanization, included when available). A representative word-level entry is {台文: 聰明, 中文: 聰明、聰穎、聰敏、伶俐 ("clever; smart"), 台羅: tshong-bîng}, and a representative sentence-level entry is {台文: 紅嬰仔哭甲一身軀汗。, 中文: 小嬰兒哭得滿身大汗。 ("The baby cried until it was covered in sweat.")}.

H Error-Type Distribution of the Evaluation Set

Table 6 reports the error-type distribution over the 154-sentence evaluation set (306 error spans). These statistics are a property of the Liu et al. (2026) benchmark that we evaluate on; we reproduce them here for completeness.