Corpus-Grounded Reverse Verification and Abstention for Extremely Low-Resource Paiwan Translation

Yung-Yu Huang
National Sun Yat-sen University
appleb127@g-mail.nsysu.edu.tw

Abstract

Low-resource machine translation remains challenging when target-language digital resources are scarce and inference cannot rely on reference translations. We study Paiwan (ISO 639-3: pwn), a Formosan language with limited digital resources, and present a corpus-grounded framework that retrieves from a 5,923-entry Paiwan–Chinese bilingual corpus and generates candidate translations. We report its selective outputs and separately evaluate Reverse Verification (RV), a round-trip consistency classifier for candidate filtering.

We evaluate on 86 items across two diagnostic sets. On Set A (40 lexical/phrase items), the system achieves a strict normalized exact match rate of **23/40 = 57.5%** and a relaxed heuristic match rate of **30/40 = 75.0%**. On Set B (46 dialogue sentences), it produces one correct and two partial translations, 30 marker-detected refusals, and 13 pipeline failures caused by tool-call leakage; end-to-end translation coverage is **3/46 = 6.5%**, and no outputs were classified as hallucinated under the operational criteria. On the same 40 Set A inputs, RV, evaluated as a binary consistency classifier, attains **92.0%** precision (**23/25**) against automatic operational reference-match labels at a decision threshold of $\tau_{\text{rv}} = 0.35$; RV is a consistency diagnostic rather than independent evidence of translation correctness. The primary contribution is an evidence-grounded analysis of selective translation behavior and standalone consistency-based candidate filtering under severe corpus constraints, rather than high translation accuracy.

1 Introduction

Machine translation for extremely low-resource languages faces a fundamental barrier: when target-language digital resources are scarce and no reference translations exist, standard approaches fail. For Eastern Paiwan (ISO 639-3: pwn), the challenge is acute because digital resources are limited and reference translations are unavailable at inference time.

Retrieval-augmented generation (RAG) partially addresses this gap by grounding outputs in a curated bilingual corpus. However, retrieval alone cannot eliminate errors. When corpus coverage is insufficient, systems must choose between generating fluent but hallucinated translations or refusing entirely. We argue that in extremely low-resource settings, abstention is preferable to unsupported generation.

Intrinsic self-correction without external feedback can be unreliable in reasoning tasks (Huang et al., 2024). Accordingly, we do not treat model self-evaluation as independent evidence of translation correctness.

This work examines selective outputs from a corpus-grounded translator and separately tests whether reverse consistency can filter translation candidates. We propose Reverse Verification (RV), which transforms translation evaluation into a bidirectional consistency problem over shared retrieval evidence. Rather than asking "Is this translation correct?"—which cannot be directly verified without references—RV asks "Does the reverse translation return the original input?" This provides a corpus-grounded risk signal.

We evaluate on 86 items using GLM-4.5-Air (ZhipuAI, 2025). Set B yields only 3/46 = 6.5% end-to-end translation coverage. As a standalone binary classifier on Set A, RV achieves 92.0% precision against automatic operational reference-match labels. Set A and

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 180–187
November 6-7, 2026 ©2026 Association for Computational Linguistics

Set B show a marked performance disparity, but differ in both input composition and corpus overlap. Set B responses were predominantly marker-detected refusals, and none was classified as hallucinated under the operational criteria.

Scope. This work focuses exclusively on Eastern Paiwan using GLM-4.5-Air.

2 Related Work

Extremely Low-Resource MT. Low-resource language pairs remain challenging for neural machine translation (Araabi and Monz, 2020). Recent work on Formosan machine translation investigates how bilingual lexical resources can improve translation between Mandarin and 16 Formosan languages (Zheng et al., 2024). Our work aligns with constraining large language model (LLM) generation to corpus-sourced content.

Retrieval-Augmented Generation. The RAG framework (Lewis et al., 2020) combines parametric LLM knowledge with non-parametric retrieval. Retrieval augmentation has been shown to substantially reduce hallucination in knowledge-grounded dialogue (Shuster et al., 2021). In our Set B evaluation, outputs were predominantly refusals, motivating explicit analysis of coverage and abstention.

Reverse Translation and Quality Estimation. Back-translation (Sennrich et al., 2016) exploits monolingual data for neural MT, while quality estimation (QE) systems (Specia et al., 2018; Kim et al., 2017) predict translation quality without references. Our RV differs: it serves as an internal consistency diagnostic from round-trip consistency rather than a separately trained QE model.

RV measures consistency over shared retrieval evidence, not absolute correctness, and may exhibit correlated errors. We do not treat model self-evaluation or round-trip consistency as independent evidence of translation correctness.

3 Methodology

3.1 System Overview

The Layered Evaluation and Assurance Framework (LEAF) is a corpus-grounded translation framework for extremely low-resource languages. Its implementation combines exact or retrieval-augmented lookup with candidate generation in a multi-agent architecture comprising an Orchestrator, Knowledge Agent, and Quality Agent. The Quality Agent provides an optional review path when invoked. Because the frozen Set A/B artifacts do not preserve per-item Quality Agent decisions, the standalone RV analysis is evaluated separately rather than treated as the demonstrated gate for the observed Set B refusals. The implementation follows ReAct (Yao et al., 2023) with a maximum of five turns to bound execution.

3.2 Corpus and Retrieval

The archived retrieval snapshot comprises **5,923** Paiwan–Chinese bilingual entries, each containing a Paiwan string, Chinese translation, and internal source label. These project-level labels are not treated as complete external provenance. Evaluation overlap differs by set: 29/40 Set A bilingual pairs occur verbatim in the archived 5,923-entry corpus, whereas none of the 46 Set B bilingual pairs does. Because test–train overlap can confound benchmark interpretation (Lewis et al., 2021), we treat Set A as a corpus-grounded lexical/phrase diagnostic rather than an unseen-generalization test. Given input x, the system first attempts exact-match lookup in corpus dictionary $\mathcal{D}_{\text{corpus}}$; if no match exists, dense vector retrieval is performed:

$$\text{RAG}_k(x) = \text{NN}_k(x; \mathcal{D}_{\text{corpus}}) \tag{1}$$

where NN_k denotes the k nearest corpus entries under the vector index selected by the invoked retrieval route. The preserved runtime supports multiple index backends, and the frozen item artifacts do not record the index selected for each output.

3.3 Candidate Translation

Given retrieved context $\text{RAG}_k(x)$, the LLM generates candidate translation $\hat{y}$ conditioned on input x and retrieved evidence. We write $\text{RAG}(x) = \text{RAG}_{\text{LLM}}(x, d, \text{RAG}_k(x))$ for brevity:

$$\text{translate}(x, d) = \begin{cases} \text{Dict}(x) & x \in \text{dom}\, \mathcal{D}_{\text{corpus}} \\ \text{RAG}(x) & \text{otherwise} \end{cases} \tag{2}$$

where $d \in \{c2p, p2c, auto\}$ specifies translation direction and RAG_{LLM} generates translation conditioned on retrieved corpus examples.

3.4 Reverse Verification

For the standalone classifier evaluated in Table 3, a candidate Paiwan translation $\hat{y}$ of Chinese input x is translated back to Chinese:

$$x \xrightarrow{T_{\text{c2p}}} \hat{y} \xrightarrow{T_{\text{p2c}}} \hat{z} \tag{3}$$

The reverse translation uses the opposite direction:

$$\hat{z} = \text{translate}(\hat{y}, \text{p2c}) \tag{4}$$

After the evaluation script's trailing-character cleanup, $C(\cdot)$ denotes the set of characters in a string. The evaluated score is character-set Jaccard similarity:

$$s_{\text{rv}}^{\text{eval}}(x, \hat{z}) = \frac{|C(x) \cap C(\hat{z})|}{|C(x) \cup C(\hat{z})|} \tag{5}$$

Unusable candidates or reverse translations receive score zero. The standalone classifier accepts when $s_{\text{rv}}^{\text{eval}} \geq \tau_{\text{rv}} = 0.35$. This measures round-trip consistency, not linguistic correctness. For eligible generated candidates, the forward translator may independently perform an internal round-trip retry; that mechanism is separate from this classifier. Optional Quality Agent review uses a distinct RV cascade, described next.

3.5 Output Policy and Abstention

When the optional Quality Agent `review_translation` task is invoked, it computes a rule score and a separate reverse-verification score. The preserved frozen Set A/B records do not establish that every output traversed this review. Let $h = 1$ for nonempty translation content not beginning with `[不確定]`, and $h = 0$ otherwise. Let $\ell = 0.8$ when $0.2 \leq |\hat{y}|/\max(|x|, 1) \leq 5.0$, and $\ell = 0.3$ otherwise. Let $u = 1$ when none of `[不確定]`, `[unsure]`, 我無法, or 我不確定 occurs, and $u = 0.2$ otherwise; round_2 denotes rounding to two decimals:

$$s_{\text{rule}} = \text{round}_2\left(\frac{h + \ell + u}{3}\right) \tag{6}$$

The Quality Agent's own RV cascade scores normalized exact match as 1.0, fixed synonym-group match as 0.9, substring match as 0.8, and Chinese-character overlap of at least 0.6 as the overlap rounded to two decimals; otherwise it returns explicit failure with score zero. This overlap divides by the larger character-set size, not the Jaccard union in Eq. 5. If RV is unavailable, the rule score alone is used. Writing the Quality Agent score as $s_{\text{rv}}^{\text{QA}}$:

$$Q = \begin{cases} 0.2s_{\text{rule}} + 0.8s_{\text{rv}}^{\text{QA}} & \text{explicit RV failure,} \\ 0.4s_{\text{rule}} + 0.6s_{\text{rv}}^{\text{QA}} & \text{RV verified,} \\ s_{\text{rule}} & \text{RV unavailable.} \end{cases} \tag{7}$$

An explicit RV failure makes the review fail immediately. Under `mastery_first`, a non-failing $Q \geq \theta_q = 0.8$ passes directly. If $Q < 0.8$ and an LLM client is available, its review score s_{LLM} gives:

$$Q_{\text{final}} = \frac{Q + s_{\text{LLM}}}{2} \tag{8}$$

This branch passes when $Q_{\text{final}} \geq 0.8$. Without an LLM client, the below-threshold fallback passes when $Q \geq 0.6$. These are review decisions; a failed review does not itself generate user-facing refusal text.

3.6 Verification Layer

On the Orchestrator's normal reply path, a procedural check compares recognized Paiwan–Chinese pairs in candidate reply $\hat{r}$ with translation pairs parsed from available tool results. Let $P(\hat{r})$ be the recognized pairs and $\mathcal{V}$ the bidirectional lookup formed from tool-returned pairs:

$$V_{\text{bad}}(\hat{r}) = \{(w, c) \in P(\hat{r}) \mid \{w, c\} \cap \text{dom}\, \mathcal{V} = \emptyset\} \tag{9}$$

When tool-returned pairs are available, recognized unmatched pairs are flagged with a warning in the reply. This check does not universally inspect every Paiwan word or block output.

3.7 Agent Pipeline Failure Taxonomy

During bounded ReAct execution (maximum 5 turns), pipeline degradation patterns are classified with severity weights: tool-call leakage

($w = 3$, raw syntax escapes parser); JSON fragment ($w = 3$); excessive calls ($w = 2$, >5 per turn); empty reply ($w = 2$); language collapse ($w = 2$, Chinese-only output); hallucinated word ($w = 1$); uncertainty markers ($w = 1$). The failure severity score of a candidate output m is:

$$D(m) = \sum_{i=1}^{7} w_i \cdot \mathbf{1}[\mathit{indicator}_i(m)] \qquad (10)$$

where w_i are the severity weights and $\mathit{indicator}_i(m)$ detects each failure pattern.

3.8 Runtime Configuration

Runtime settings are route-dependent rather than universal across all reported analyses. GLM-4.5-Air is the recorded model backbone. The standalone RV classifier uses $\tau_{\text{rv}} = 0.35$; when optional Quality Agent review is invoked under `mastery_first`, its quality threshold is $\theta_q = 0.8$. Generation temperature and retrieval depth vary by runtime route, as summarized in Section 4.2.

4 Experimental Setup

We evaluate three research questions: (RQ1) corpus-grounded retrieval for Paiwan lexical translation in the evaluated extremely low-resource setting; (RQ2) system behavior on compositional translation, including refusals and pipeline failures; and (RQ3) standalone Reverse Verification as a binary classifier for filtering candidates.

4.1 Dataset and Corpus

The **5,923**-entry archived retrieval snapshot can be reproduced exactly, in order and content, by applying the preserved runtime cleaning and source-priority procedure to an internally merged 10,195-entry file. Complete upstream acquisition provenance is not preserved for every entry. We use two diagnostic sets: **Set A** contains 40 Paiwan–Chinese word/phrase items; 29/40 exact bilingual pairs occur verbatim in the archived 5,923-entry retrieval corpus. **Set B** contains 46 dialogue sentences from a Paiwan conversation textbook; none of the 46 exact bilingual pairs occurs verbatim in the archived corpus. We treat both as small diagnostic sets and do not claim population-level representativeness. Set A comprises 20 dictionary-style lexical entries with speaker metadata, 15 phrase items, and 5 sentence items with grammar-note/intent metadata; the Paiwan side has a median length of 1.5 whitespace-delimited tokens (range 1–12). Set B comprises 23 A-turn and 23 B-turn utterances spanning seven scenarios (morning 10, daily life 10, school 8, friendship 6, family 6, weather 4, culture 2); after removing whitespace and punctuation, the Chinese inputs have a median length of 8 characters (mean 8.0, range 3–15), and all 46 Chinese inputs and all 46 Paiwan references are unique. No formal a priori sample-size criterion was recorded for these diagnostic sets, and no explicit syntactic-difficulty annotation is available for either set; these statistics are therefore descriptive rather than validated difficulty strata.

4.2 Model Configuration

GLM-4.5-Air (ZhipuAI, 2025) is the recorded model backbone across the reported analyses. Preserved project configuration comments record a free GLM-4.5-Air token allocation and a switch intended to avoid consuming paid GLM-4-Plus quota; service comments also report exhausted Flash/Plus allocations. These documented API-resource constraints provide the practical model-selection rationale, without establishing an empirical advantage. For the frozen Set A/B evaluation, the Orchestrator/direct-LLM route requests temperature 0.3, whereas translator-service generation requests 0.2 and uses 0.1 on retry. Retrieval depth is route-dependent: Knowledge Agent retrieval defaults to top-$k = 5$, while `PaiwanTranslator.translate` defaults to top-$k = 8$; the per-item route and selected vector index are not preserved in the frozen artifact. The standalone RV analysis calls `PaiwanTranslator` with its default retrieval depth and uses a classifier threshold of 0.35. Optional Quality Agent review, when invoked under `mastery_first`, uses a quality threshold of 0.8. We therefore do not treat these settings as a single universal configuration. A single recorded backbone keeps model identity consistent within these analyses, but no controlled comparison with newer models was conducted; the results therefore do not

establish model superiority.

4.3 Evaluation Metrics

Set A. Set A is reported with a strict normalized exact-match metric (lowercasing, whitespace and punctuation removal, with the predefined i/y normalization rule) and a separate relaxed heuristic cascade including normalized exact match, substring agreement, $\geq$50% word overlap, and first-three-character matching of the first word.

Set B. Set B uses a failure taxonomy: **Correct** (reference-exact after normalization/formatting cleanup); **Partial** (not reference-exact, but manual source-assisted review found supporting evidence); **Refusal** (marker-detected refusal); **Wrong** (interpretable translation not satisfying source-assisted Partial criterion); **Hallucination** (fabricated Paiwan content with no corpus basis); **Pipeline failure** (tool-call leakage or implementation failures). Borderline non-refusal outputs underwent manual review against the single reference translation with lexical and sentence-level checks against East Paiwan materials in the Klokah Indigenous Language E-Learning Park (Indigenous Languages Research and Development Foundation, n.d.). This source-assisted review was not a blinded native-speaker evaluation.

5 Results

5.1 Lexical Translation (Set A)

A standalone raw-substring criterion matched **25/40 = 62.5%** items. Under the hierarchical relaxed cascade (Table 1), **23/40 = 57.5%** matched by normalized exact comparison, three additional items by substring comparison, zero additional by word-overlap, and four by the first-three-character heuristic, yielding **30/40 = 75.0%**.

Table 1: Set A Lexical Translation Results (N = 40)

Metric	Count	Percentage
Strict normalized exact	23/40	57.5%
Additional substring	3/40	7.5%
Additional word-overlap	0/40	0.0%
Additional 3-char heuristic	4/40	10.0%
Remaining mismatch	6/40	15.0%
Pipeline/extraction failure	4/40	10.0%
Total (relaxed)	**30/40**	**75.0%**

The six remaining automatic mismatches include alternative forms/spellings not captured by the automatic normalization rules; they were not comprehensively adjudicated by blinded native speakers.

5.2 Dialogue Translation (Set B)

On 46 dialogue sentences, the system produced **one correct**, **two partial**, **30 refusals**, and **13 pipeline failures**. Table 2 presents complete outcomes.

Table 2: Set B Outcomes and Selective Metrics ($N = 46$)

Outcome	Count	Percentage
Reference-exact	1/46	2.2%
Partial translation	2/46	4.3%
Refusal (marker-detected)	30/46	65.2%
Pipeline failure	13/46	28.3%
Total	**46/46**	**100.0%**
Selective Metrics:		
Pipeline availability	33/46	71.7%
Translation coverage	3/46	6.5%
Conditional release rate	3/33	9.1%
Conditional refusal rate	30/33	90.9%

The dominant behavior is refusal: the system explicitly states it cannot find translations in the corpus for 30/46 items. No outputs were classified as hallucinated under the operational criteria. Tool-call leakage (13 items) represents an implementation issue where internal ReAct tool-invocation syntax escaped the output parser.

Interpretation. Although 33/46 (71.7%) inputs reached interpretable outcomes, the system released translations for only 3/46 (6.5%) of all inputs, corresponding to 3/33 (9.1%) among interpretable outputs. Only 1/46 (2.2%) yielded fully correct translation. The high conditional refusal rate (30/33, 90.9%) co-occurs with very low end-to-end translation coverage (3/46, 6.5%), highlighting limited practical usefulness despite conservative output behavior.

5.3 Reverse Verification as Binary Classifier (RQ3)

We evaluate standalone Reverse Verification on the same 40 Set A inputs in a separate run, accepting candidates at $s_{\text{rv}}^{\text{eval}} \geq \tau_{\text{rv}} = 0.35$. The automatic reference-match evaluator labeled 28 candidates correct, one partial, and

11 wrong. For Table 3, Positive groups correct and partial as 29 operational positive labels; Negative comprises 11 operational negative labels. These labels are not independently adjudicated linguistic ground truth.

Table 3: Standalone Reverse Verification on Set A against automatic operational reference-match labels ($N = 40$, $\tau_{\text{rv}} = 0.35$).

	Positive	**Negative**	**Total**
RV Accept	TP = 23	FP = 2	25
RV Reject	FN = 6	TN = 9	15
Total	29	11	40

Metric	**Value**
Precision	92.0% (23/25)
Recall	79.3% (23/29)
F1	0.852
Specificity	81.8% (9/11)
Accuracy	80.0% (32/40)

Against these operational labels, 23 of 25 RV-accepted candidates received a positive reference-match label (92.0% precision). Table 4 analyzes the two false positives.

False-positive analysis. FP1 (雲 → qadupu → 雲) originated from conflicting high-priority lexical knowledge: while the main corpus contained *qerepus*→ 雲 and *qadupu*→ 書, the higher-priority core vocabulary contained conflicting *qadupu*→ 雲, creating a self-consistent but reference-inconsistent cycle.

FP2 (豬 → qacang → 豬) involved corpus-level lexical competition: both *qacang* and *dridri* were associated with 豬 in the available corpus. No dialect metadata or native-speaker adjudication is available to determine whether these forms are linguistically equivalent or regionally distinct.

These cases illustrate that RV measures consistency, not absolute correctness: shared erroneous lexical knowledge can create self-consistent incorrect cycles, and single-reference evaluation can disagree with round-trip-consistent outputs when multiple corpus-supported forms correspond to one source gloss. Some false negatives are consistent with spelling variation that falls below the character-level similarity threshold.

5.4 Exploratory Task-Completion Diagnostic

We additionally report a pre-freeze diagnostic conducted with an earlier four-agent prototype comprising an Orchestrator and Knowledge, Teaching, and Quality agents. Twenty targeted prompts (five each for knowledge, teaching, quality, and composite tasks) were evaluated in three sequential loops per configuration. The reported metric is the mean five-check task-completion rubric score over 60 prompt–loop interactions per configuration, not translation accuracy. This diagnostic predates the frozen 86-item evaluation and is therefore reported separately rather than as an ablation of the final system.

SD is the sample standard deviation across the 60 interaction-level rubric scores; Δ is computed from unrounded means. On these targeted prompts, the No Knowledge Agent configuration showed the largest observed reduction in mean rubric score relative to the Full prototype. Because this diagnostic used an earlier prototype and targeted prompts, we do not interpret these differences as causal component effects or as performance estimates for the frozen evaluation system.

5.5 Pipeline Failure Cases

Table 6 illustrates representative failure modes.

These motivate LEAF's bounded ReAct loop ($T = 5$) and verification layer for flagging unverified content.

5.6 Set B Translation Cases

b21 (Correct): 好久不見。→ *cuayanga tja sinicevunga.* (reference-exact after markdown cleanup).

b14 (Partial): 是的，我們在家常說族語。→ *ui, qivuivu amen a i tjumaq.* (reference: *ui, ru penaiwanan a men i tjumaq.*). The prediction was retained as Partial in the source-assisted adjudication; no preserved artifact independently verifies equivalence to the reference.

b15 (Partial): 今天天氣好嗎? → *bulay a kinagadavan kasauni?* (reference: *bulay a kalevelevan tucu?*). The predicted sentence is directly attested in the official East Paiwan supplementary teaching material with the cor-

Table 4: False Positive Case Analysis ($N = 2$)

Input	Ref	Output	RV	Failure Type
雲 (cloud)	qerepus	qadupu	1.0	Conflicting high-priority lexical mapping
豬 (pig)	dridri	qacang	1.0	Corpus-level lexical competition

Table 5: Exploratory pre-freeze task-completion diagnostic (20 prompts × 3 loops per configuration).

Config	Mean rubric (%)	SD (pp)	Δ (pp)
Full prototype	**99.3**	**3.6**	—
No Knowledge Agent	95.7	9.8	−3.7
No Quality Agent	98.0	6.1	−1.3
No Teaching Agent	98.3	5.6	−1.0

responding Chinese meaning (Council of Indigenous Peoples, 2014).

6 Discussion

Set A–Set B Performance Disparity. Set A and Set B differ in both input composition and corpus overlap: 29/40 exact Set A bilingual pairs occur verbatim in the archived retrieval corpus, whereas none of the 46 Set B pairs does. Set A attained 30/40 = 75.0% under the relaxed reference-match heuristic; Set B had 3/46 = 6.5% translation coverage and predominantly marker-detected refusals. The observed disparity therefore cannot be attributed solely to compositional difficulty.

Observed Output Behavior. Operational Hallucination Outcome. No Set B outputs were classified as hallucinated under the operational criteria. This result should not be interpreted as evidence that the system is hallucination-free or that the verification pipeline causally prevents hallucination.

RV Limitations. Standalone Reverse Verification achieved 92.0% precision against automatic operational reference-match labels, but cannot establish absolute correctness. When both directions use the same corpus and model, correlated errors can create false acceptances. Spelling/orthographic variation can cause false rejections. RV is a consistency check, not ground truth verification.

System Scope. LEAF is a corpus-grounded selective translation framework, not a general-purpose Paiwan translation system. In the evaluated Set B outputs, behavior is characterized by frequent refusal and very low translation coverage rather than high translation accuracy.

7 Conclusion

This work investigated translation for Eastern Paiwan in an extremely low-resource setting using GLM-4.5-Air. The results highlight a verification challenge alongside very low end-to-end translation coverage in this evaluated Eastern Paiwan setting. We study corpus-grounded translation with selective outputs and separately evaluate Reverse Verification as a candidate filter. RV transforms translation evaluation into a bidirectional consistency problem over shared retrieval evidence.

On 86 Paiwan–Chinese translation items, end-to-end translation coverage was only 3/46 = 6.5% on Set B dialogue sentences, substantially limiting practical utility. As a standalone binary classifier on Set A, RV achieves 92.0% precision against automatic operational reference-match labels while measuring consistency rather than absolute correctness. Two false-positive mechanisms illustrate evidence dependence: shared conflicting lexical mappings can create self-consistent incorrect cycles, and reference-sensitive corpus-level lexical competition can produce disagreement with round-trip-consistent outputs.

Limitations. Experiments are conducted on Eastern Paiwan only using GLM-4.5-Air. Set B outputs underwent source-assisted review against a single reference, not blinded native-speaker evaluation. The pre-freeze task-completion diagnostic is descriptive rather than causal, and we did not conduct a complete controlled comparison of Direct LLM, RAG-only, RAG+RV, and RAG+RV+Abstention. **Future work** includes establishing larger stratified test sets with native-speaker evaluation, strengthening corpus coverage, and conducting controlled

Table 6: Representative Pipeline Failure Cases

Case	Input	Expected	Failure / Type
1	孩子 (child)	kakedrekedre	Tool-call leak (Impl)
2	房子 (house)	umaq	Tool-call leak (Impl)
3	名字 (name)	ngadan	Lang. collapse (LLM)
4	Multi-turn	—	Raw JSON (Parser)

component comparisons.

References

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435. International Committee on Computational Linguistics.

Council of Indigenous Peoples. 2014. 原住民族語言學習補充教材：句型篇——東排灣語高中版. https://lokahsu.ilrdf.org.tw/web_lokahsu/Files/bookSystem/book/23_senior_book.pdf. Issued September 1, 2014.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*.

Indigenous Languages Research and Development Foundation. n.d. Klokah (族語 E 樂園): Multi-search. https://web.klokah.tw/multiSearch/. Multi-search educational resource including East Paiwan materials; accessed 2026-08-09.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. 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, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving low-resource machine translation for Formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11248–11259, Bangkok, Thailand. Association for Computational Linguistics.

ZhipuAI. 2025. GLM-4.5-Air. https://huggingface.co/zai-org/GLM-4.5-Air. Official model card; accessed 2026-09-28.