engineering proceedings

Proceeding Paper

Hybrid Dictionary–Retrieval-Augmented Generation–Large Language Model for Low-Resource Translation [†]

Reen-Cheng Wang [*], Cheng-Kai Yang, Tun-Chieh Yang and Yi-Xuan Tseng

Department of Computer Science and Information Engineering, National Taitung University, Taitung 950309, Taiwan; fk86663212@gmail.com (C.-K.Y.); fergus45065211@gmail.com (T.-C.Y.); hanstseng0312@gmail.com (Y.-X.T.)

[*] Correspondence: rcwang@nttu.edu.tw; Tel.: +886-89-517607

[†] Presented at 8th International Conference on Knowledge Innovation and Invention 2025 (ICKII 2025), Fukuoka, Japan, 22–24 August 2025.

Abstract

The rapid decline of linguistic diversity, driven by globalization and technological standardization, presents significant challenges for the preservation of endangered languages, many of which lack sufficient parallel corpora for effective machine translation. Conventional neural translation models perform poorly in such contexts, often failing to capture semantic precision, grammatical complexity, and culturally specific nuances. This study addresses these limitations by proposing a hybrid translation framework that combines dictionary-based pre-translation, retrieval-augmented generation, and large language model post-editing. The system is designed to improve translation quality for extremely low-resource languages, with a particular focus on the endangered Paiwan language in Taiwan. In the proposed approach, a handcrafted bilingual dictionary is the first to establish deterministic lexical alignments to generate a symbolically precise intermediate representation. When gaps occur due to missing vocabulary or sparse training data, a retrieval module enriches contextual understanding by dynamically sourcing semantically relevant examples from a vector database. These enriched words are then processed by an instruction-tuned large language model that reorders syntactic structures, inflects verbs appropriately, and resolves lexical ambiguities to produce fluent and culturally coherent translations. The evaluation is conducted on a 250-sentence Paiwan–Mandarin dataset, and the results demonstrate substantial performance gains across key metrics, with cosine similarity increasing from 0.210–0.236 to 0.810–0.846, BLEU scores rising from 1.7–4.4 to 40.8–51.9, and ROUGE-L F1 scores improving from 0.135–0.177 to 0.548–0.632. These results corroborate the effectiveness of the proposed hybrid pipeline in mitigating semantic drift, preserving core meaning, and enhancing linguistic alignment in low-resource settings. Beyond technical performance, the framework contributes to broader efforts in language revitalization and cultural preservation by supporting the transmission of Indigenous knowledge through accurate, contextually grounded, and accessible translations. This research demonstrates that integrating symbolic linguistic resources with retrieval-augmented large language models offers a scalable and efficient solution for endangered language translation and provides a foundation for sustainable digital heritage preservation in multilingual societies.

Keywords: low-resource machine translation; retrieval-augmented generation; lexicon-guided prompting; hybrid large language models; endangered languages

Academic Editors: Teen-Hang Meen, Chun-Yen Chang and Cheng-Fu Yang

Published: 5 February 2026

Citation: Wang, R.-C.; Yang, C.-K.; Yang, T.-C.; Tseng, Y.-X. Hybrid Dictionary–Retrieval-Augmented Generation–Large Language Model for Low-Resource Translation. *Eng. Proc.* **2025**, *120*, 52. https://doi.org/10.3390/engproc2025120052

1. Introduction

Driven by globalization and rapid technological change, linguistic diversity is declining at an unprecedented pace. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) estimated in May 2024 that about 7000 languages remain globally, about 40% of which are already classified as endangered [1]. A language disappears every two weeks on average, undermining the cultural knowledge systems encoded in these speech communities [2]. The indigenous languages of Taiwan, including the Paiwan language, face the pressures of data scarcity and shrinking speaker populations. Therefore, innovative and engaging strategies for language learning and intergenerational transmission have become an urgent priority for cultural revitalization.

Most of the existing Paiwan–Mandarin translation tools are limited to lexical lookups that cannot satisfy practical needs for sentence-level or paragraph-level translation. Word-for-word output often yields semantic dissonance and fails to convey contextual nuance or cultural specificity, hampering meaningful language acquisition. To bridge this gap, we developed a large-language-model (LLM) translation system that combines a dictionary-based pre-translation module and retrieval-augmented generation (RAG). The system moves beyond the constraints of literal translation, producing fluent and accurate sentences and thereby enhancing both learning efficiency and user engagement. The system offers a viable pathway for promoting and preserving Paiwan in contemporary society.

The remainder of this article is structured as follows. Section 2 reviews the literature on machine translation for low-resource languages. Section 3 describes the proposed methodology in detail. Section 4 reports and discusses the experimental results. Finally, Section 5 concludes the study and outlines directions for future research.

2. Literature Review

Conventional machine translation (MT) systems perform poorly on minority languages because large, high-quality parallel corpora are seldom available. Haddow et al. [3] noted that most languages lack the extensive resources required to train state-of-the-art MT models. Ranathunga et al. [4] therefore explored transfer learning, subword segmentation, data augmentation, and multilingual pre-training. While these strategies improve robustness, they remain inadequate when the parallel corpus contains fewer than 100,000 sentences. These languages are generally classified as low-resource languages (LRL), and the corresponding MT task is termed low-resource machine translation (LRMT).

Recently, research on MT has moved toward hybrid machine translation (HMT) [5] to combine the strengths of different paradigms while offsetting their respective weaknesses, thereby improving overall quality and accuracy. HMT is a composite framework that integrates rule-based (RBMT), statistical (SMT), and neural (NMT) components, often complemented by human post-editing.

LLMs have recently been widely used in low-resource translation research owing to their strong generative capacity and zero-shot or few-shot abilities. Two significant paradigms have emerged, one fine-tuning the LLM based on limited corpora and the other combining RAG with prompt engineering. Soudani et al. [6] showed that RAG consistently outperforms fine-tuned models in LRMT. RAG supplements generation with external knowledge, reducing reliance on large annotated data and decreasing computational demands, advantages that are particularly salient for Paiwan, where data are extremely scarce. By injecting retrieved evidence at inference time, RAG also mitigates over-fitting and the attendant risk of hallucinated translations. Table 1 presents a concise comparison of the principal challenges and representative methods in low-resource machine translation.

Previous studies further corroborate the effectiveness of RAG for LRMT. Shu et al. [7] adopted Cherokee, Tibetan, and Manchu texts were performed by their retrieval-based model combined with GPT-4o to perform translation. The RAG configuration markedly outperformed pure GPT-4o and Llama 3.1 405B baselines. Merx et al. [8] evaluated few-shot LLM prompting for MT of Mambai and observed a sharp disparity in Bilingual Evaluation Understudy (BLEU) scores across test sets. The score of language-manual translation excerpts achieved up to 21.2 BLEU, while those of native-speaker test sets reached only 4.4. BLEU is a metric for automatically evaluating MT text proposed by Papineni et al. in the context of the study [9].

Table 1. Evolution and challenges of mt paradigms for low-resource languages.

Model	Characteristics for LRMT	Strengths for LRMT	Weaknesses for LRMT	Data Requirement
RBMT [10]	Manual linguistic rules, dictionaries	Grammatical precision, interpretability	High development/maintenance cost, limited scalability, poor idiom handling	Low (manual rules)
SMT [11]	Probabilistic models trained on parallel data	Automated learning from data, adaptable to new domains with data	Limited fluency/grammar with scarce data, high Out-of-Vocabulary rates	Medium (parallel corpora)
NMT [12,13]	Deep neural networks (e.g., Transformers)	High fluency, contextual understanding, state of the art for high-resource modeling	Highly data-hungry, prone to overfitting with small data, and significant performance disparities for LRLs	High (large parallel corpora)
LLM-based MT [7,14]	Large pre-trained generative models	In-context learning, reasoning, adaptability, zero-shot/ few-shot potential	Performance disparities for LRLs, potential for hallucination, and computational cost	Very High (pre-training corpora)/Low (fine-tuning/ prompt engineering)

Wang et al. [15] introduced RAGtrans, a 79,000-instance benchmark whose translations were produced jointly by GPT-4o and professional translators. Using multilingual (Chinese, German, French, and Czech) Wikipedia documents as the retrieval corpus and a multitask objective for cross-lingual information completion, self-knowledge enrichment, and relevance discrimination, RAGtrans improved BLEU by 1.6–3.1 and crosslingual optimized metric for evaluation of translation (COMET) score by 1.0–2.0 over pure instruction tuning, confirming RAG's ability to improve accuracy and cross-lingual knowledge integration.

Evidence from Taiwan's Hakka illustrates similar gains. Chang et al. [16] compared dictionary lookup, ChatGPT-4, Google Gemini 2.0, and RAG-enhanced variants for Chinese-to-Hakka translation. Dictionary-only output scored about 12 BLEU, whereas an RAG+Gemini configuration achieved about 31 BLEU and delivered superior coverage of specialized terminology and cultural expressions. A two-stage approach, which combined dictionary pre-translation followed by Gemini post-editing, reached 26 BLEU, still surpassing single-method baselines.

Yet in extremely LRLs, such as Paiwan, whose orthography is primarily Romanized and whose written corpus is minimal, vector retrieval may falter because of sparse exemplars. To address this limitation, we adopted a two-phase hybrid workflow. First, a dictionary alignment phase that produces a symbolically precise intermediate representation by mapping input tokens to dictionary entries and analyzing local structure; and second, an LLM-recomposition phase that feeds the intermediate form, along with any retrievable context, into a large language model tasked with reordering and naturalizing the sentence while preserving semantic fidelity.

To mitigate the rare-word tagging problem, we used the hybrid word-character approach proposed by Luong and Manning [17], which raised BLEU by 2.1–11.4 points on the WMT'15 English to Czech translation task and reliably generated correct rare words without resorting to the <unk> token. Building on this strategy, we integrated RAG with LLM-based text synthesis, yielding a composite architecture that combines rule-based pre-processing, RAG-supported lexicon expansion, and LLM post-editing. The resulting pipeline delivers culturally coherent, high-quality translations for languages that lack both extensive corpora and standardized orthographies, thereby offering a practical pathway for language revitalization and digital preservation in extremely low-resource settings.

3. Methodology

3.1. Research Design and Architecture

We developed a hybrid translation system that integrates dictionary alignment, RAG, and LLM-based post-editing, optimized for the Paiwan–Mandarin low-resource settings. A handcrafted bilingual lexicon provides deterministic one-to-one mapping for core vocabulary. When a source item is absent from the dictionary, the system activates an RAG module that dynamically retrieves parallel sentences or related terms, thereby enriching the local context before final reconstruction by LLM. We hypothesized that this architecture outperforms an unadapted LLM in accuracy, semantic completeness, and fluency.

To test this hypothesis, we conducted quantitative analyses on three metrics: cosine similarity, BLEU, recall-oriented understudy for gisting evaluation, and longest common subsequence (ROUGE-L) F1 scores. The unadapted LLM output scores are used as the statistical baseline. Cosine similarity captures the semantic proximity between sentence embeddings (0 = dissimilar, 1 = identical). BLEU assesses n-gram overlap with the brevity penalty, while ROUGE-L F1 measures the longest common subsequence, balancing precision and recall. The combined metrics allow a comprehensive quantification of translation quality, which we subject to statistical significance tests.

3.2. Methodological Framework

The workflow of the model comprises three principal modules.

- Dictionary pre-processing: The bilingual lexicon is serialized to JSON and loaded as an in-memory hash map for constant-time look-ups. The workflow of the pre-processing hash map construction is illustrated in Figure 1.

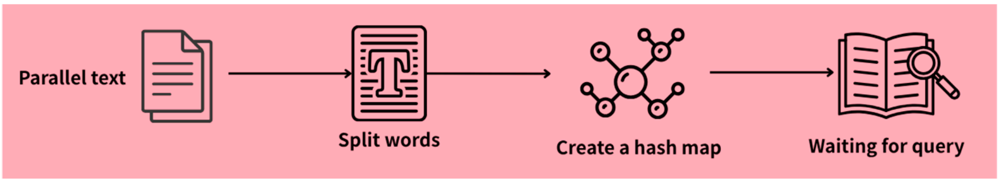

Figure 1. Pre-processing hash map construction workflow.

- Vector retrieval (RAG): Parallel corpora, including textbooks, grammars, and biblical texts, are digitized, embedded with a sentence-encoder model, and stored in a Qdrant vector database to enable rapid semantic retrieval. The pipeline of vector embedding is shown in Figure 2.

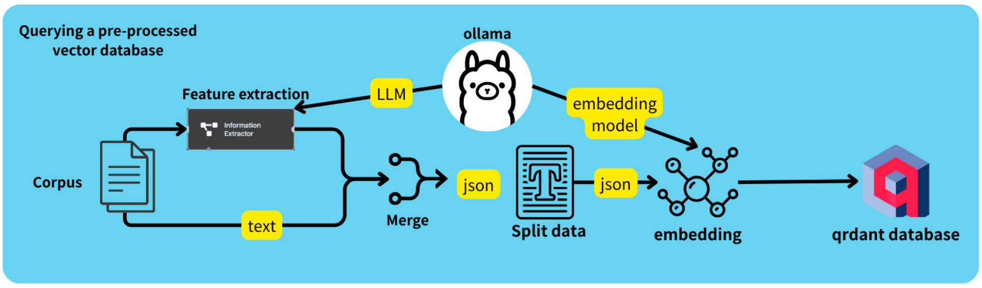

Figure 2. Pre-processed vector data database pipeline.

- LLM post-editing: The dictionary glosses and retrieved context are passed, via a syntax-aware prompt, to an instruction-tuned LLM that performs reordering, inflection, and lexical disambiguation. The overall translation workflow of what we propose is demonstrated in Figure 3.

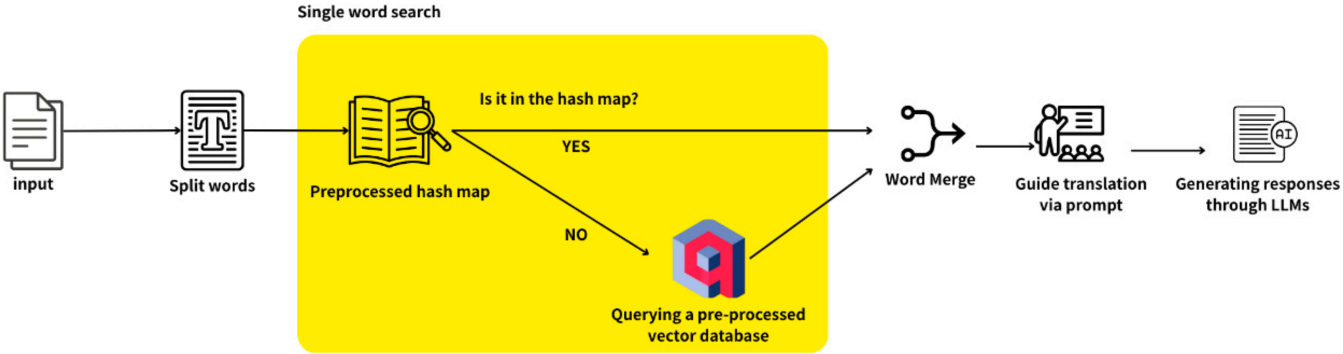
Single word search
input
Split words
Preprocessed hash map
Is it in the hash map?
YES
NO
Querying a pre-processed vector database
Word Merge
Guide translation via prompt
Generating responses through LLMs

Figure 3. Proposed overall translation workflow diagram.

3.3. Data Collection and Corpus Construction

The Paiwan corpus utilized in this study comprises four primary sources: East Paiwan Learning Handbook published by Taiwan's Ministry of Education, A Descriptive Grammar of Paiwan, the Indigenous Language Elementary Reader: Conversational Series, and the Paiwan Bible (Kai nua Cemas a Pinayuanan). A bilingual lexicon of high-frequency simplex and compound entries was compiled in both comma-separated values and JavaScript object notation formats. All parallel sentences were subsequently segmented, converted into sentence-level embeddings, and indexed in a Qdrant vector database to enable efficient semantic retrieval.

3.4. Baseline Models and Prompt Design

Baseline employs the GPT-4o mini, against which Llama 3.1-70B, Gemma 3-27B, and DeepSeek-R1-Distill-Llama-70B are compared with gauge performance and cost efficiency. Prompts explicitly describe Paiwan verbs with the initial syntax and voice morphology, provide relevant gloss and context, and instruct the model to produce fluent Mandarin.

3.5. Experiment

From the Paiwan textbooks published by the Ministry of Education, 250 sentences were randomly sampled as the test set, covering diverse syntactic constructions and discourse contexts. Each model is evaluated under two conditions, with and without our hybrid translation modules, yielding a cross-factorial design. System outputs are aligned, sentence by sentence, with the official textbook translations to compute all metrics.

4. Result and Discussion

We combined a hash-map-based dictionary pre-translation module with different LLMs to translate Paiwan into Mandarin and evaluate performance on three metrics. Semantic similarity was computed with the sup-SimCSE-bert-base-uncased encoder [18] released by the Princeton NLP Group. System outputs and reference translations were transformed into L2-normalized CLS vectors, after which their inner product was taken. Values close to 1 denote near-identical semantics.

Quantitative evaluation underscores the pivotal role of the dictionary pre-translation component. In its absence, cosine similarity values range from 0.210 to 0.236 across all LLM configurations, as shown in Figure 4. After integration, scores improve to 0.810–0.846, representing an average absolute gain of roughly 0.60. This finding indicates that the translation module markedly enhances the semantic fidelity of the output. An analogous trend emerges for the BLEU in Figure 5. Baseline scores of 1.7–4.4 rise to 40.8–51.9, an improvement of approximately 40–50. These data confirm that the translation module plays a pivotal role in maintaining both the accuracy and fluency of the output. The most pronounced enhancement occurs for ROUGE-L F1 in Figure 6, which rises from 0.135–0.177 to 0.548–0.632, an absolute increase of about 0.45. The results demonstrate that the translation module substantially improves long-sentence alignment and overall structural coherence in the translated text.

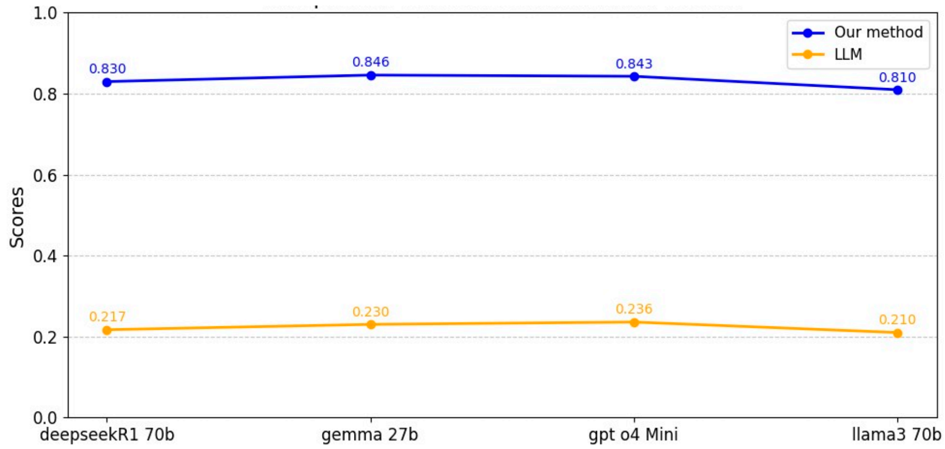

Figure 4. Comparison of cosine score.

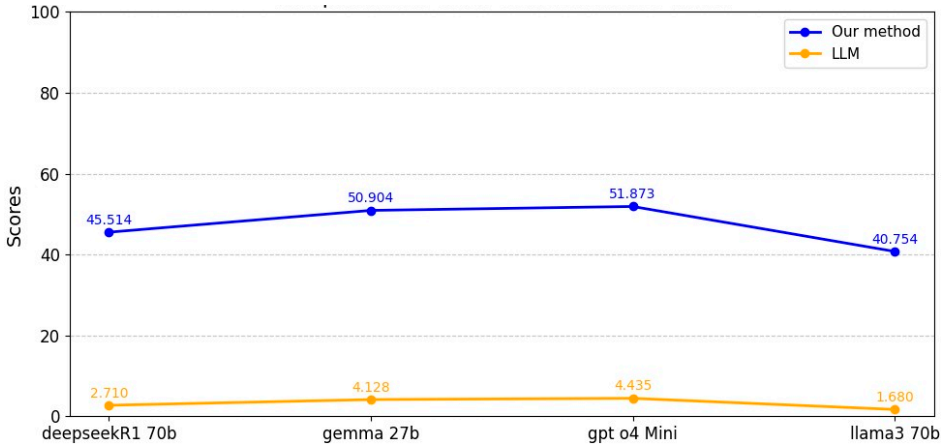

Figure 5. Comparison of BLEU score (%).

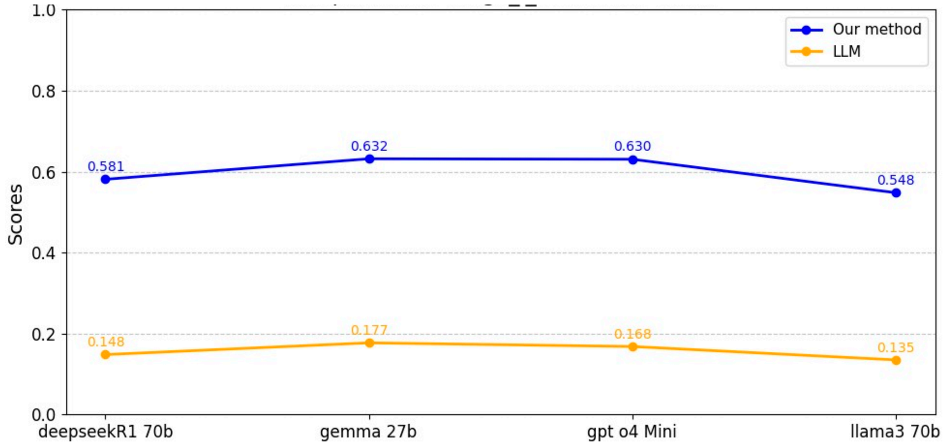

Figure 6. Comparison of ROUGE-L F1.

These results validate the quantitative gains by showing that the proposed pipeline provides stable alignments for Paiwan proper nouns and multiword expressions, thereby enabling LLM to preserve core meaning while still producing culturally appropriate, fluent Mandarin. In particular, the developed module shows a strong alignment for specialized terminology and compound forms, improving both fluency and semantic fidelity during sentence recomposition. Conversely, gaps in the dictionary lead to noticeable semantic deviations, underscoring the critical importance of comprehensive lexical resources. These observations indicate that properly embedding a dictionary within an RAG-supported, even small-scale, LLM-driven workflow constitutes an efficient and effective strategy for LRMT, one that not only achieves strong scores but also safeguards linguistic nuance and cultural detail by facilitating accurate keyword alignment and morpheme retention.

5. Conclusions

Integrating a dictionary-based pre-translation module with LLMs markedly enhances both the accuracy and semantic fidelity of Paiwan-to-Mandarin translation. Composite evaluations that use cosine similarity, BLEU, and ROUGE-L F1 confirm that the module mitigates the semantic drift commonly observed in LLM outputs for LRLs. With the module in place, cosine similarity improved by approximately 0.60, BLEU by 40–50, and ROUGE-L F1 by about 0.45. These enhancements emphasize the module's pivotal role in stabilizing lexical alignment, preserving core meaning, and improving fluency. The developed system depends heavily on dictionary coverage. When single or multi-word entries are missing, translation quality deteriorates and semantic deviations emerge. Performance is currently strongest on declarative sentences, with accuracy and fluency declining for inverted constructions and culturally laden clauses that entail complex grammatical dependencies.

Therefore, it is necessary to expand and refine the lexical database, with particular attention to proper nouns and high-frequency word and sentence collocations. More granular semantic-parsing and context-aware mechanisms, such as richer prompt tailoring and explicit grammatical cues, should further equip LLM to handle diverse sentence types and contextual complexity. Parallel engagement with native Paiwan speakers, coupled with a more comprehensive human assessment framework, also facilitates systematic tracking and continuous improvement of translation quality. This dual analysis ensures that the proposed system performs not only numerically but also meets the practical needs of end-users in real-world language-revitalization contexts. The dictionary–RAG–LLM pipeline holds substantial promise for LRMT. Continued dictionary development and architectural refinement are expected to yield a robust, scalable solution applicable to a broader range of endangered languages.

Author Contributions: Conceptualization, all authors; Methodology, all authors; Software, C.-K.Y.; Validation, C.-K.Y.; Formal analysis, R.-C.W. and T.-C.Y.; Resources, R.-C.W.; Data curation, C.-K.Y. and Y.-X.T.; Writing—original draft, Y.-X.T. and T.-C.Y.; Writing—review & editing, R.-C.W.; Visualization, T.-C.Y.; Supervision, R.-C.W.; Project administration, C.-K.Y.; Funding acquisition, R.-C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the project "Flow (mavalidi) and Morphogenesis (sivaikan): Co-creating the Conditions for Sustainable Development of the South-Link Area in Taitung County" under Grant no. NSTC 113-2420-H-143 -003 -HS3 and "An AI Conversational Assistant Based on Retrieval-Augmented Generation: A Feasibility Study for Enhancing Paiwan Language Learning" under Grant no. NSTC 114-2813-C-143-032-H of the National Taitung University, sponsored by the National Science and Technology Council, Taiwan, R.O.C.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the National Cheng Kung University Governance Framework for Human Research Ethics (protocol code 108-376-2 on 5 December 2019; 111-468-2 on 25 October 2022; and 114-0779-2 on 2 November 2025).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Acknowledgments: The authors would like to thank all colleagues and students at the Center for Innovation in Humanities and Social Practices at National Taitung University who contributed to collecting ethnographic research data for this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. UNESCO. Multilingual Education, the Bet to Preserve Indigenous Languages and Justice. Available online: https://www.unesco.org/en/articles/multilingual-education-bet-preserve-indigenous-languages-and-justice (accessed on 14 July 2025).
2. Ahmed, K. From Igbo to Angika: How to Save the World's 3,000 Endangered Languages. Available online: https://www.theguardian.com/global-development/2025/jan/07/cultural-identity-saving-worlds-endangered-languages-activists-online-tools-rohingya-igbo-angika (accessed on 14 July 2025).
3. Haddow, B.; Bawden, R.; Barone, A.V.M.; Helcl, J.; Birch, A. Survey of low-resource machine translation. *Comput. Linguist.* **2022**, *48*, 673–732. [CrossRef]
4. Ranathunga, S.; Lee, E.S.A.; Prifti Skenduli, M.; Shekhar, R.; Alam, M.; Kaur, R. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.* **2023**, *55*, 229. [CrossRef]
5. Costa-Jussa, M.R.; Fonollosa, J.A. Latest trends in hybrid machine translation and its applications. *Comput. Speech Lang.* **2015**, *32*, 3–10. [CrossRef]
6. Soudani, H.; Kanoulas, E.; Hasibi, F. Fine-tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 12–22.
7. Shu, P.; Chen, J.; Liu, Z.; Wang, H.; Wu, Z.; Zhong, T.; Li, Y.; Zhao, H.; Jiang, H.; Pan, Y.; et al. Transcending language boundaries: Harnessing llms for low-resource language translation. *arXiv* **2024**, arXiv:2411.11295. [CrossRef]
8. Merx, R.; Mahmudi, A.; Langford, K.; de Araujo, L.A.; Vylomova, E. Low-resource machine translation through retrieval-augmented llm prompting: A study on the Mambai language. *arXiv* **2024**, arXiv:2404.04809.
9. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*; Association for Computing Machinery: New York, NY, USA, 2002; pp. 311–318.
10. Shiwen, Y.; Xiaojing, B. Rule-based machine translation. In *Routledge Encyclopedia of Translation Technology*; Routledge: London, UK, 2014; pp. 186–200.
11. Lopez, A. Statistical machine translation. *ACM Comput. Surv. (CSUR)* **2008**, *40*, 8. [CrossRef]
12. Stahlberg, F. Neural machine translation: A review. *J. Artif. Intell. Res.* **2020**, *69*, 343–418. [CrossRef]
13. Enis, M.; Hopkins, M. From LLM to NMT: Advancing low-resource machine translation with Claude. *arXiv* **2024**, arXiv:2404.13813. [CrossRef]

14. Thakur, M. Towards Neural No-Resource Language Translation: A Comparative Evaluation of Approaches. *arXiv* **2024**, arXiv:2412.20584. [CrossRef]
15. Wang, J.; Meng, F.; Zhang, Y.; Zhou, J. Retrieval-augmented machine translation with unstructured knowledge. *arXiv* **2024**, arXiv:2412.04342. [CrossRef]
16. Chang, C.C.; Li, C.F.; Lee, C.H.; Lee, H.S. Enhancing Low-Resource Minority Language Translation with LLMs and Retrieval-Augmented Generation for Cultural Nuances. *arXiv* **2025**, arXiv:2505.10829.
17. Luong, M.T.; Manning, C.D. Achieving open vocabulary neural machine translation with hybrid word-character models. *arXiv* **2016**, arXiv:1604.00788. [CrossRef]
18. Gao, T.; Yao, X.; Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv* **2021**, arXiv:2104.08821.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.