From Scarcity to Scalability: Lexicon and Grammar Enhanced Amis to Mandarin Translation with GPT Models

Joseph Lin[1]   Kai-ying Lin[2]   Hung-Yu Kao[3]

[1]Hsinchu County American School, Hsinchu, Taiwan
[2]Institute of Linguistics, Academia Sinica, Taipei, Taiwan
[3]Dept. of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

lintinghaojoseph@gmail.com   limkhaiin@gate.sinica.edu.tw   hykao@cs.nthu.edu.tw

Abstract

Machine translation (MT) for low-resource languages remains constrained by extreme data scarcity, making traditional fine-tuning infeasible. This study examines Amis→Mandarin translation as a practical case, leveraging GPT-4o-mini and GPT-5-mini with dictionary integration and grammar-informed prompting. Experiments show that GPT-5-mini, supported by dictionary, achieves usable quality (BLEU-3 ~31, COMET ~78, BLEURT ~71). To address the bottleneck of incomplete dictionaries, we propose *Context-Driven Lexical Augmentation*, which infers Mandarin equivalents for unseen Amis terms from corpus context, raising BLEU-3 to 34 and establishing a stronger basis for semi-automatic corpus generation. These results demonstrate that expanding and refining dictionary provides greater benefits than parameter-intensive fine-tuning in extremely low-resource settings.

We also discuss the performance gap between Amis→Mandarin and Mandarin→Amis translation, attributing it to Amis's morphological complexity and narrower semantic coverage. Overall, our resource-driven strategy offers a scalable pathway toward high-quality MT and corpus expansion, ultimately supporting both linguistic research and language revitalization.

Keywords: Low-Resource Languages, Machine Translation, Prompt Engineering, Data Augmentation

1 Introduction

Taiwan's indigenous languages, widely recognized as the cradle of the Austronesian family (Blust, 2013), are critically endangered. Many are spoken by only a few hundred individuals, and their decline threatens both cultural continuity and linguistic diversity. Developing digital tools to support documentation and revitalization has therefore become an urgent priority. Among such tools, machine translation (MT) holds particular promise, as it can facilitate communication across language communities and accelerate the creation of linguistic resources. However, MT systems typically depend on large-scale parallel corpora, standardized orthographies, and comprehensive dictionary-conditions absent for most indigenous languages. This scarcity firmly categorizes them as low- or no-resource languages, requiring methods that can operate effectively under extreme data constraints.

In this work, we focus on Amis→Mandarin translation as a representative case of low-resource MT. Amis is the most widely spoken indigenous language in Taiwan, yet it remains critically underrepresented in digital resources, like large-scale corpora, standardized lexical tools, or annotated datasets to support NLP tasks like MT. We evaluate two large language models—GPT-4o-mini (small-scale) and GPT-5-mini (mid-tier)—in combination with existing Amis–Mandarin dictionaries (Zheng et al., 2022) and descriptive grammar resources (阿美語語法概論) (Council of Indigenous Peoples, 2017). Our experiments show that grammar-informed prompting benefits smaller models, while GPT-5-mini achieves strong performance with dictionary support alone. Error analysis further reveals that dictionary coverage, rather than syntactic complexity, is the principal bottleneck. To address this, we propose a *Context-Driven Lexical Augmentation* method that infers translations for unseen Amis words, yielding mea-

November 20-22, 2025 ©2025 Association for Computational Linguistics

surable improvements in BLEU and semantic metrics. These findings highlight that systematic enrichment of dictionary is more effective than scaling model size or parallel data alone.

We emphasize the Amis→Mandarin direction for both cultural and practical reasons. From a preservation perspective, initiating data collection from Mandarin and translating into Amis risks cultural bias, as it inevitably introduces Mandarin concepts absent from Amis traditions. Practically, Amis narratives can be translated into Mandarin with sufficient accuracy to bootstrap new corpora while keeping human validation effort low. In contrast, Mandarin encompasses a much broader semantic space, covering domains such as science, technology, and politics, whereas Amis maintains a more compact lexicon rooted in ecology, kinship, and oral tradition (Lewis et al., 2023). As a result, corpora originating in Amis are more coherent and transferable into Mandarin, while the reverse direction often requires paraphrase or approximation that BLEU score penalizes heavily. Taken together, these factors make Amis→Mandarin the most ethical and reliable pathway for semi-automatic corpus expansion.

At the same time, we acknowledge that some prior studies have reported higher BLEU scores for Mandarin→Amis, contrary to our findings. We discuss potential reasons for this discrepancy and its implications for evaluating low-resource MT. Finally, although this study focuses on Amis→Mandarin, we also outline a pathway toward robust bidirectional MT. By combining semi-automatic corpus generation, dictionary augmentation, and semantic-aware evaluation, future work can enable fine-tuning and ultimately achieve high-quality Amis–Mandarin translation in both directions, advancing both NLP research and the revitalization of Taiwan's indigenous languages.

Our contributions are threefold:

- We conduct the first systematic evaluation of Amis→Mandarin translation with GPT-based models, showing that mid-tier LLMs achieve usable quality when supported by dictionary resources.
- We further introduce *Context-Driven Lexical Augmentation*, a lightweight method for expanding dictionary coverage by inferring translations for unseen Amis words, directly improving BLEU and semantic scores.
- We establish Amis→Mandarin as a practical direction for semi-automatic corpus generation and outline a pathway toward high-quality bidirectional MT through future fine-tuning on expanded corpora.

2 Related Work

Low-resource machine translation (MT) faces a fundamental challenge: parallel corpora are too small to support stable fine-tuning of large models. With billions of parameters but only a few thousand sentence pairs, gradient updates are weak, training quickly overfits to idiosyncratic examples, and generalization suffers (Haddow et al., 2022). Empirical studies suggest that with only a few to at most tens of thousands of pairs, full fine-tuning of large models tends to be unstable and prone to overfitting, making parameter-efficient alternatives preferable (Gu et al., 2018). Since the Amis–Mandarin corpus contains only ~5,000 pairs, our setting falls well below this threshold, motivating approaches that leverage external resources such as bilingual dictionaries and grammatical descriptions rather than relying solely on parallel data.

A range of alternatives to full fine-tuning has been explored. Prompt-based methods and in-context learning reduce dependence on large datasets but often deliver inconsistent results. For example, retrieval-augmented prompting with dictionary support reached BLEU ~21 for English–Mambai in one domain but dropped to ~4 in another, revealing limited robustness (Merx et al., 2024). Prompt tuning can exploit structural cues (Schucher et al., 2022), yet its success is highly sensitive to template design and it often struggles to enforce lexical fidelity. Liao et al. (Liao et al., 2024) examined error-feedback prompting for Mandarin→Amis translation, showing that iterative correction brought modest improvements. By contrast, our work centers on Amis→Mandarin, integrating dictionary and grammar resources into prompting and extending coverage through automated lexical augmentation, producing more reliable gains.

In the Formosan and Austronesian context, resources remain sparse but are slowly expanding. Zheng et al. (Zheng et al., 2022) introduced the first Amis–Mandarin parallel corpus and dictionary, demonstrating that dictionary augmentation benefits fine-tuned mBART models. Their experiments reported higher BLEU for Mandarin→Amis (15–19) than for Amis→Mandarin (<7), suggesting directional asymmetry. Yet other research points the other way: Zhang et al. (Zhang et al., 2024) showed that Mandarin→Zhuang achieved much lower BLEU (~16) than Zhuang→Mandarin (~32). Taken together, these studies indicate that directionality may be influenced by morphology, semantic coverage, dictionary completeness, and modeling strategy (fine-tuning vs. prompting).

Lin et al. (Lin et al., 2025) advanced this line of work by releasing FormosanBench, a benchmark spanning Amis, Atayal, and Paiwan across several NLP tasks, including MT. Their evaluation revealed persistent performance gaps relative to high-resource languages, underscoring the importance of approaches tailored to the typological and lexical characteristics of Formosan languages rather than relying exclusively on transfer from unrelated high-resource settings.

Building on this foundation(Lin, 2025), our work proposes a dictionary- and grammar-driven framework for Amis→Mandarin translation with GPT models. Unlike earlier prompting studies that relied on static dictionary , we introduce *Context-Driven Lexical Augmentation*, a proof-of-concept method that infers Mandarin equivalents for unseen Amis words from corpus context. This augmentation improved BLEU-3 from ~31 to 34 and raised semantic scores, surpassing the modest gains reported for prior prompting strategies (Merx et al., 2024; Liao et al., 2024). More broadly, our findings suggest that lexical expansion and semantic-aware evaluation are more scalable and effective than parameter-intensive fine-tuning in extremely low-resource conditions, while also shedding new light on the role of directionality in Amis–Mandarin MT.

3 Translation Framework and Evaluation Metrics

We present an Amis–Mandarin translation framework that integrates mid-tier large language models (LLMs) with lexical and grammatical resources. The system combines dictionary pre-searching and grammar-informed prompting with an iterative auto-prompting procedure, which refines outputs by dynamically adjusting prompts across evaluation rounds. This design offers a practical and scalable strategy for machine translation in low-resource Austronesian languages.

3.1 Models and Data

We evaluate two large language models: GPT-4o-mini, a smaller-scale model, and GPT-5-mini, a mid-tier model. The dataset comprises 5,751 Amis–Mandarin sentence pairs (Zheng et al., 2022), partitioned into 576 for training (used exclusively in the auto-prompt setting), 575 for validation, and 4,600 for testing across all prompt strategies. To maximize evaluation coverage, 80% of the data is allocated to testing, reflecting the fact that prompt engineering does not rely on training sets. In addition, we utilize a bilingual glossary containing 7,927 Amis–Mandarin entries (Zheng et al., 2022), implemented as a Pandas DataFrame to facilitate efficient search and retrieval. Collectively, these resources serve as the most comprehensive Amis–Mandarin parallel dataset currently available.

3.2 Preprocessing

All sentence pairs were standardized before translation. Preprocessing involved removing extra spaces, newline markers, and punctuation. Amis tokens were lowercased while retaining apostrophes, which carry morphological information. Sentences were processed in batches of 20 to balance efficiency with model context length. For each sentence, tokens were normalized and matched against the glossary using RapidFuzz similarity, retrieving up to three candidate translations, or a single match when similarity exceeded 95%.

3.3 Prompting Strategies

We evaluate three prompting strategies, each incorporating glossary-based lexical hints:

1. **Baseline Prompting** (Figure 1(a)): Prompts incorporate detailed formatting instructions and a comprehensive glossary look-up table to guide initial translation efforts. For each word, a fuzzy matching algorithm is employed to retrieve up to three candidate translations, prioritizing those exceeding 80% similarity to ensure high relevance. Sentences are processed efficiently in batches of 20 to optimize computational resources and maintain consistency across translations.

2. **Grammar-Rule Prompting** (Figure 1(b)): Builds on the baseline by appending a curated set of rules from the Amis grammar book (Council of Indigenous Peoples, 2017) on word order, affixation, and case markers, using a similar batching process. Instead of embedding the full 177-page 《秀姑巒阿美語—語法概論》 into each prompt, we use GPT-5 to distill it into a ~3-page "core pack" of high-impact rules (e.g., clause structure, linker *a*, case/voice morphology, negation, relative clauses). An ablation study with 500 randomly selected sentence pairs showed that appending the full dictionary offered no benefit in improving scores. This pack is *frozen* and prepended to Amis→Mandarin prompts. This extract-then-inject approach may mitigate long-context issues like "lost in the middle" (Liu et al., 2024). The compact pack lowers latency/cost, freeing tokens for lexicon snippets and enhancing controllability.

3. **Auto-Prompt Training** (Figure 1(c)): An iterative refinement cycle designed to enhance translation quality through systematic feedback, comprising:

   (a) **Batch Translation**: Process 20-sentence batches using the baseline setup, ensuring consistent input handling and initial translation generation across the corpus.

   (b) **Error Analysis**: Conduct a detailed comparison of translated outputs against reference texts, identifying systematic errors such as lexical mismatches, syntactic deviations, or semantic inaccuracies to pinpoint areas for improvement.

   (c) **Prompt Update**: Revise prompt instructions to address identified issues, incorporating targeted adjustments —e.g., clarifying ambiguous rules or adding contextual cues—based on error patterns observed.

   (d) **Iteration**: Apply the updated prompt to subsequent batches, iteratively refining the process across the training dataset to progressively enhance translation fidelity and coherence.

   Auto-Prompt Training can be conceptualized as a dynamic process wherein the large language model (LLM) implicitly derives grammar rules and linguistic patterns through iterative error analysis and correction. This self-adaptive mechanism leverages accumulated insights to evolve the prompt, with prompts automatically generated by GPT based on the difference between prediction and ground truth reference. The resulting optimized prompt, enriched with corrections and contextual understanding, is then deployed across the test dataset.

3.4 Evaluation Metrics

We assess translation quality using three complementary metrics—BLEU-3, BLEURT, and COMET—each normalized to a 0—100 scale for unified comparison.

BLEU-3, which measures up to 3-gram overlap, is better suited than full BLEU for low-resource MT because shorter n-grams are more reliably captured and impose fewer penalties on valid paraphrases (Liao et al., 2024). The BLEU-3 score is calculated as:

$$\text{BLEU-3} = \text{BP} \cdot \exp\left(\sum_{n=1}^{3} \frac{1}{3} \log p_n\right),$$

where p_n is the precision for each n-gram order and BP is the brevity penalty.

For all BLEU calculations, we apply the `method1` smoothing function from NLTK's `SmoothingFunction`. Under this method, when an n-gram precision would otherwise be zero, it is replaced with a very small constant

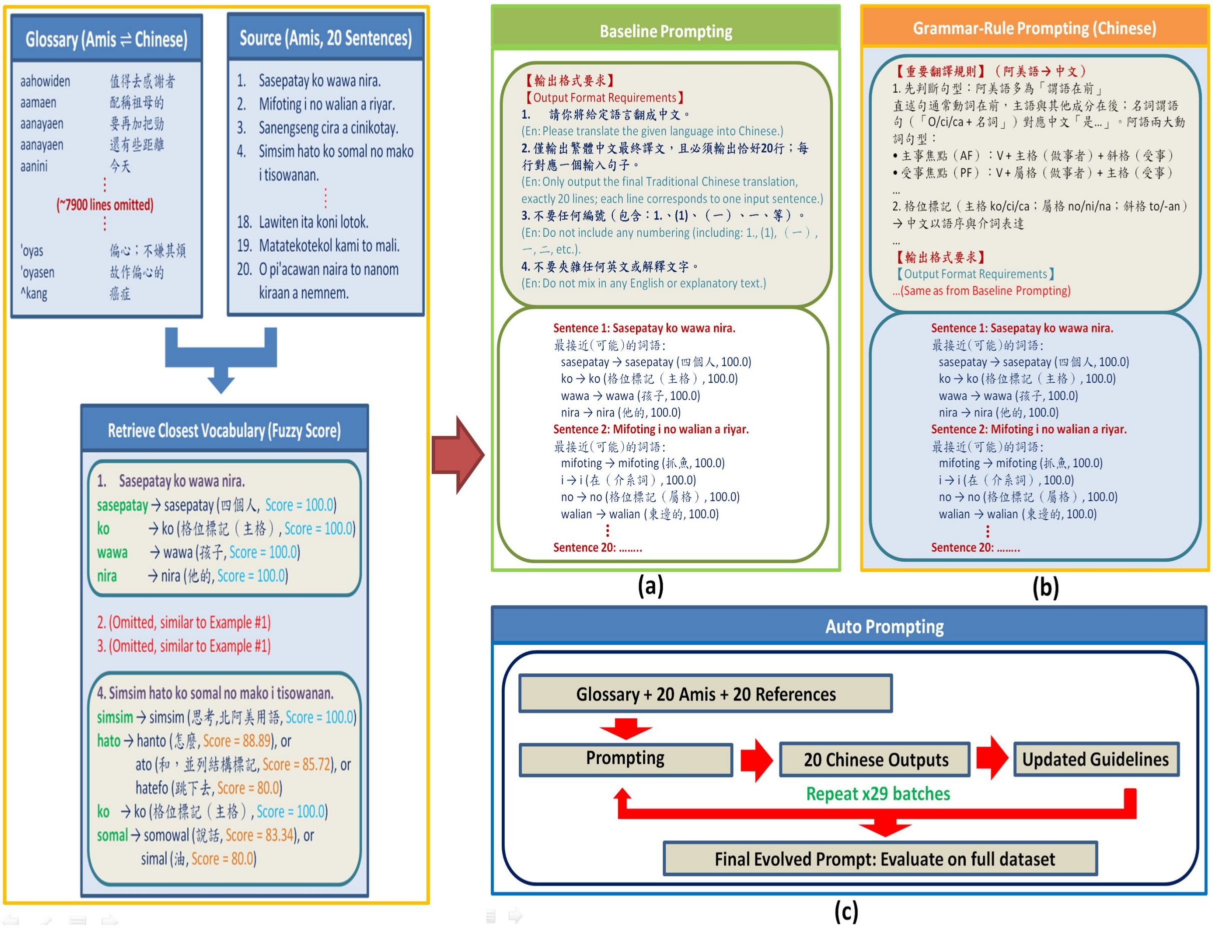

Glossary (Amis ⇌ Chinese)
aahowiden 值得去感謝者
aamaen 配稱祖母的
aanayaen 要再加把勁
aanayaen 還有些距離
aanini 今天
(~7900 lines omitted)
'oyas 偏心；不嫌其煩
'oyasen 故作偏心的
^kang 癌症
Source (Amis, 20 Sentences)
1. Sasepatay ko wawa nira.
2. Mifoting i no walian a riyar.
3. Sanengseng cira a cinikotay.
4. Simsim hato ko somal no mako i tisowanan.
18. Lawiten ita koni lotok.
19. Matatekotekol kami to mali.
20. O pi'acawan naira to nanom kiraan a nemnem.
Retrieve Closest Vocabulary (Fuzzy Score)
1. Sasepatay ko wawa nira.
sasepatay → sasepatay (四個人, Score = 100.0)
ko → ko (格位標記（主格）, Score = 100.0)
wawa → wawa (孩子, Score = 100.0)
nira → nira (他的, Score = 100.0)
2. (Omitted, similar to Example #1)
3. (Omitted, similar to Example #1)
4. Simsim hato ko somal no mako i tisowanan.
simsim → simsim (思考,北阿美用語, Score = 100.0)
hato → hanto (怎麼, Score = 88.89), or
ato (和，並列結構標記, Score = 85.72), or
hatefo (跳下去, Score = 80.0)
ko → ko (格位標記（主格）, Score = 100.0)
somal → somowal (說話, Score = 83.34), or
simal (油, Score = 80.0)
Baseline Prompting
【輸出格式要求】
【Output Format Requirements】
1. 請你將給定語言翻成中文。
(En: Please translate the given language into Chinese.)
2. 僅輸出繁體中文最終譯文，且必須輸出恰好20行；每行對應一個輸入句子。
(En: Only output the final Traditional Chinese translation, exactly 20 lines; each line corresponds to one input sentence.)
3. 不要任何編號（包含：1.、(1)、（一）、一、等）。
(En: Do not include any numbering (including: 1., (1), （一）, 一, 二, etc.).
4. 不要夾雜任何英文或解釋文字。
(En: Do not mix in any English or explanatory text.)
Sentence 1: Sasepatay ko wawa nira.
最接近(可能)的詞語:
sasepatay → sasepatay (四個人, 100.0)
ko → ko (格位標記（主格）, 100.0)
wawa → wawa (孩子, 100.0)
nira → nira (他的, 100.0)
Sentence 2: Mifoting i no walian a riyar.
最接近(可能)的詞語:
mifoting → mifoting (抓魚, 100.0)
i → i (在（介系詞）, 100.0)
no → no (格位標記（屬格）, 100.0)
walian → walian (東邊的, 100.0)
Sentence 20: ....
(a)
Grammar-Rule Prompting (Chinese)
【重要翻譯規則】（阿美語→中文）
1. 先判斷句型：阿美語多為「謂語在前」
直述句通常動詞在前，主語與其他成分在後；名詞謂語句（「O/ci/ca + 名詞」）對應中文「是…」。阿語兩大動詞句型：
• 主事焦點（AF）：V + 主格（做事者）+ 斜格（受事）
• 受事焦點（PF）：V + 屬格（做事者）+ 主格（受事）
...
2. 格位標記（主格 ko/ci/ca；屬格 no/ni/na；斜格 to/-an）→ 中文以語序與介詞表達
...
【輸出格式要求】
【Output Format Requirements】
...(Same as from Baseline Prompting)
Sentence 1: Sasepatay ko wawa nira.
最接近(可能)的詞語:
sasepatay → sasepatay (四個人, 100.0)
ko → ko (格位標記（主格）, 100.0)
wawa → wawa (孩子, 100.0)
nira → nira (他的, 100.0)
Sentence 2: Mifoting i no walian a riyar.
最接近(可能)的詞語:
mifoting → mifoting (抓魚, 100.0)
i → i (在（介系詞）, 100.0)
no → no (格位標記（屬格）, 100.0)
walian → walian (東邊的, 100.0)
Sentence 20: ....
(b)
Auto Prompting
Glossary + 20 Amis + 20 References
Prompting
20 Chinese Outputs
Updated Guidelines
Repeat x29 batches
Final Evolved Prompt: Evaluate on full dataset
(c)

Figure 1: Overview of the prompting framework for Amis→Mandarin MT. Each Amis sentence is first processed through fuzzy matching to retrieve up to three glossary candidates with corresponding Mandarin translations. These lexical hints are then integrated into either (a) **Baseline Prompting**, which applies only formatting requirements, or (b) **Grammar-Rule Prompting**, which supplements the baseline with explicit grammatical rules. (c) **Auto-Generated Prompting** further refines the prompt iteratively by comparing translations against references, analyzing errors, and updating guidelines before final evaluation on the full dataset.

ϵ instead. This avoids BLEU scores collapsing to zero, which is particularly important in our setting where sentences are short and higher-order matches are often sparse. In cases where matches do exist, smoothing has no effect.

To evaluate semantic quality beyond surface overlap, we also include BLEURT and COMET. BLEURT uses pretrained language models fine-tuned on human ratings, while COMET leverages multilingual contextual embeddings; both metrics correlate strongly with human judgments of translation quality. As with BLEU, we normalize BLEURT and COMET to a 0—100 range for consistent comparison.

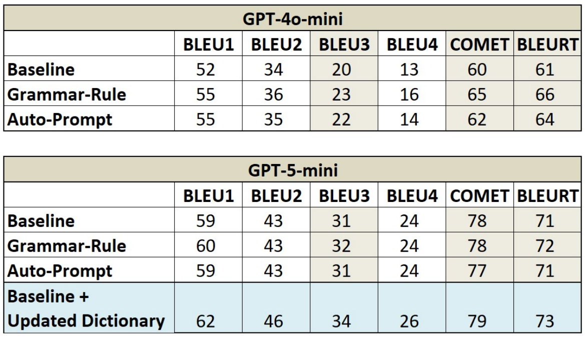
By combining n-gram precision (BLEU-3 with smoothing) and semantic adequacy (BLEURT, COMET), our evaluation framework balances literal accuracy with meaning preservation—an essential requirement for low-resource MT.

4 Results and Analysis

4.1 Evaluation with GPT Models

Figure 2 presents the performance of GPT-4o-mini and GPT-5-mini under three prompting strategies—Baseline, Grammar-Rule, and Auto-Prompt—plus an additional condition for GPT-5-mini with an augmented dictionary. Evaluation metrics include BLEU1—4, BLEU-3 (our primary n-gram metric for low-resource MT), and the semantic measures COMET and BLEURT, all normalized to a 0—100 scale. Figure 3 plots BLEU-3, COMET, and

BLEURT scores across models and prompting methods for easy comparison.



GPT-4o-mini						
	BLEU1	BLEU2	BLEU3	BLEU4	COMET	BLEURT
Baseline	52	34	20	13	60	61
Grammar-Rule	55	36	23	16	65	66
Auto-Prompt	55	35	22	14	62	64

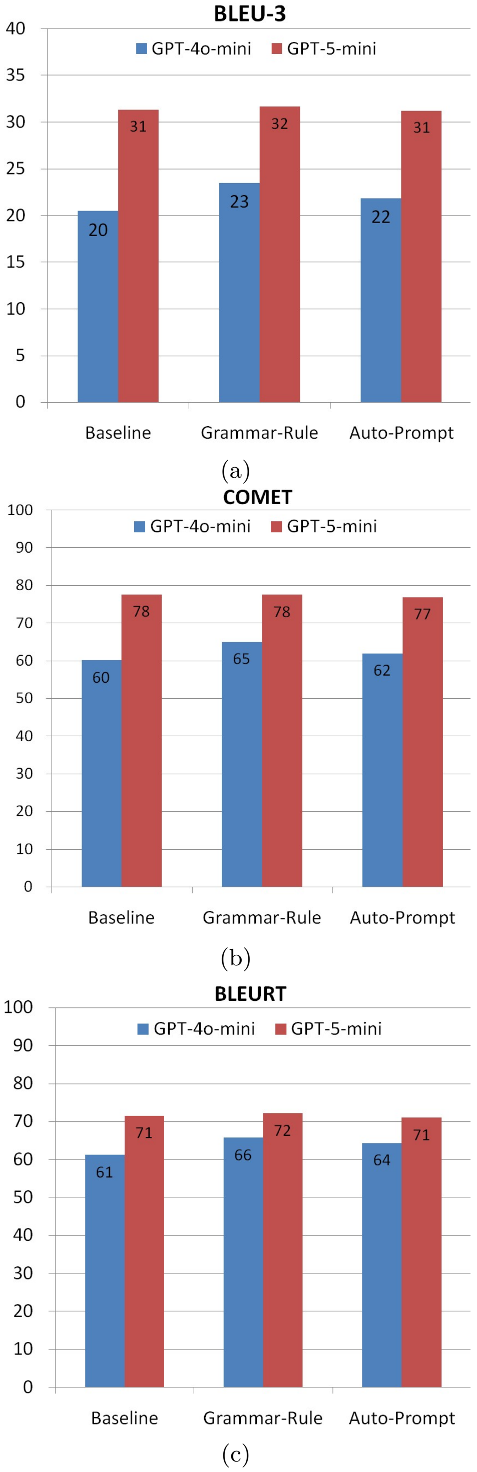
GPT-5-mini						
	BLEU1	BLEU2	BLEU3	BLEU4	COMET	BLEURT
Baseline	59	43	31	24	78	71
Grammar-Rule	60	43	32	24	78	72
Auto-Prompt	59	43	31	24	77	71
Baseline + Updated Dictionary	62	46	34	26	79	73

Figure 2: Comparison of GPT-4o-mini and GPT-5-mini performance with BLEU1-4, COMET, and BLEURT. Grammar-Rule prompting benefits GPT-4o-mini (BLEU-3: 20→23, COMET: 60→65, BLEURT: 61→66). For GPT-5-mini, dictionary augmentation delivers the largest improvement (BLEU-3: 31→34, COMET: 78→79, BLEURT: 71→73), whereas grammar rules and auto-prompting provide only marginal gains.

For GPT-4o-mini, Grammar-Rule prompting yielded consistent gains over both Baseline and Auto-Prompt: BLEU-3 rose from 20 to 23, COMET from 60 to 65, and BLEURT from 61 to 66. Auto-Prompt achieved a BLEU-3 of 22 but lagged on semantic metrics, indicating that explicit grammatical guidance is particularly valuable for smaller models that struggle with morphosyntactic variation.

GPT-5-mini, by contrast, performed strongly across all conditions, showing that it can already exploit dictionary support without extensive prompting. Baseline results reached BLEU-3 ~31, COMET ~78, and BLEURT ~71. Grammar-Rule and Auto-Prompt strategies offered only marginal gains (BLEU-3 at 32, with semantic scores differing by at most one point), suggesting that larger models are less dependent on handcrafted grammatical cues and generalize robustly from lexical hints alone.

The most notable improvement for GPT-5-mini came from context-driven lexical augmentation (detailed in the next subsection). Incorporating inferred dictionary entries for out-of-vocabulary terms increased BLEU-3 from 31 to 34, COMET from 78 to 79, and BLEURT from 71 to 73. This pattern indicates that dictionary completeness, rather than prompt complexity, is the decisive factor in improving



(a)

(b)

(c)

Figure 3: Comparison of GPT-4o-mini and GPT-5-mini performance **before dictionary augmentation**: (a) BLEU-3, (b) COMET, and (c) BLEURT.

translation quality. While additional prompting yields diminishing returns, enriching lexical coverage directly addresses the core bottleneck of low-resource MT.

In sum, GPT-5-mini consistently outperformed GPT-4o-mini on both surface overlap and semantic adequacy. Grammar rules provide clear benefits for smaller models, but for stronger LLMs, the greatest gains derive from expanding dictionary rather than layering increasingly complex prompts.

4.2 Context-Driven Lexical Augmentation

Table 1 demonstrates the effect of out-of-vocabulary (OOV) tokens on translation quality for specific cases. In the baseline system, sentences containing unknown terms such as *fitaol* achieved only BLEU-3 scores of 8—11. To address this, we applied a context-driven lexical augmentation strategy: candidate Mandarin equivalents were inferred from parallel corpus contexts (e.g., *fitaol* → 蛋殼, *pilipayan* → 禮拜天, *atolay* → 地震), and new entries were added to the Amis—Mandarin dictionary, which originally contained ~8000 words, with ~1200 additional OOV words incorporated. For OOV items occurring in multiple sentences, all plausible Mandarin interpretations were appended under the same Amis entry.

We then re-evaluated translation performance with GPT constrained to the augmented dictionary and extracted grammar rules, excluding access to reference translations to preserve evaluation integrity. This approach substantially improved BLEU-3 scores for sentences containing unseen Amis words, with some cases reaching 100 (Table 1). These results highlight the pivotal role of enriched dictionary in improving low-resource MT. While the inferred mappings require further validation by native speakers, the findings underscore the importance of systematic lexical development as a foundation for advancing Amis→Mandarin translation. Future ablation studies could isolate the impact of dictionary size versus prompt complexity to further refine these gains.

Amis **(Reference below)**	**Baseline** **(BLEU-3)**	**Updated** **Dictionary** **(BLEU-3)**
Mifitelak to fitaol ko ciwciw。 *Reference:* 小雞破殼而出。	小雞把蛋弄破了。 (11)	小雞破殼而出。 (100)
O pilipayan i nacila。 *Reference:* 昨天是禮拜天。	那是昨天發生的事。 (9)	昨天是禮拜天。 (100)
Mangernnger no atolay ko loma'。 *Reference:* 房屋被地震震動。	我家的南邊被震動了。 (8)	房屋被地震震動。 (100)

Table 1: Impact of context-driven lexical augmentation on Amis–Mandarin translation. Augmentation resolves OOV terms (e.g., *fitaol* → 蛋殼, *pilipayan* → 禮拜天, *atolay* → 地震), improving BLEU-3 from 8—11 to up to 100 for specific cases.

4.3 Limitations of BLEU for Semantic Evaluation

Table 2 illustrates cases where BLEU underestimates translation quality. GPT-5-mini Auto-Prompt outputs for some sentences receive very low BLEU scores (4-8) despite being semantically accurate, as reflected by much higher COMET scores (72—90). For instance, the Amis sentence *"Narikoran no faliyos matomes ko sota' i lalan."* is translated as " 颱風過後，路上滿是泥巴。" (BLEU-3 = 8, COMET = 90). While lexically divergent from the reference, the meaning is preserved.

Similar discrepancies appear in other examples, where paraphrasing reduces BLEU but COMET captures semantic fidelity. These results highlight BLEU's limitations in low-resource MT, particularly for languages where flexible phrasing is common. Semantic metrics such as COMET provide better alignment with human judgment and should complement BLEU in evaluation frameworks for endangered and low-resource languages. Future evaluations might explore large language models for direct scoring to further reduce bias.

Amis	**Reference /** **(GPT-5-mini)**	**COMET** **(BLEU-3)**
Narikoran no faliyos matomes ko sota' i lalan.	颱風之後馬路塡滿了污泥。 (颱風過後，路上滿是泥巴。)	90 (8)
Aka pahacikay a miparakat to tosiya.	不可開快車。 (不要把車開得太快。)	88 (5)
Do^do han ko rakat ako!	請跟從我的腳步！ (跟著我走！)	84 (4)
Ma' adangen kako to ngiha' no dadacdac.	我覺得蟬叫聲很吵。 (我被蟬的聲音吵到。)	72 (5)

Table 2: Examples where BLEU-3 penalizes paraphrasing despite high semantic fidelity, as reflected by COMET.

5 Challenges and Pathways for Amis–Mandarin MT

Prior work has reported that Mandarin→Amis translation can achieve higher BLEU than the reverse. In contrast, our experiments consistently find the opposite: Amis→Mandarin yields stronger performance. We attribute this to the fact that Amis is morphologically rich, with many surface forms for the same concept, which leads to frequent mismatches under BLEU. For example, a single Mandarin word may correspond to several Amis forms depending on context, and without explicit disambiguation, models often choose the wrong variant, resulting in lower BLEU score.

These challenges highlight why dictionary expansion with contextual metadata are crucial for future progress. Our system already achieves sufficient accuracy in the Amis→Mandarin direction to enable large-scale semi-automatic corpus generation, easing the burden on human validators and accelerating resource development. By complementing this with automatic dictionary augmentation, we can steadily improve lexical coverage and translation fidelity.

Looking ahead, we see a clear pathway: use automatic Amis→Mandarin translation to bootstrap corpora from elder narratives, refine outputs through lightweight human feedback, and progressively enrich the dictionary with contextual information. In the long term, integrating direct speech-to-text translation will further reduce barriers to language documentation and revitalization, while offering a generalizable framework for other low-resource, morphologically complex languages.

6 Conclusion

This study examined Amis→Mandarin translation as a practical case of low-resource MT, focusing on strategies that enable scalable corpus expansion despite the limited parallel data available. Our experiments show that mid-tier LLMs, particularly GPT-5-mini, can achieve usable quality when paired with dictionary support (BLEU-3 ~31, COMET ~78, BLEURT ~71). The proposed framework is applicable to any large language model comparable to or exceeding the capabilities of GPT-5-mini. While grammar-informed prompting benefits smaller models, dictionary coverage emerged as the decisive factor.

We further demonstrated that augmenting the glossary with context-inferred entries improves translation quality and establishes a threshold where large-scale semi-automatic data generation becomes feasible. This approach allows Amis narratives to be translated into Mandarin with sufficient accuracy for bootstrapping new corpora, reducing human effort to lightweight validation.

In summary, our contribution lies in reframing low-resource MT for endangered languages: progress is driven less by parameter-intensive fine-tuning and more by systematic lexical expansion, context-sensitive dictionary design, and semantic-aware evaluation. Crucially, the resulting expanded corpora will make future fine-tuning feasible, enabling higher-quality Amis–Mandarin bidirectional MT and providing a sustainable foundation for language preservation.

Acknowledgements

This research is supported by the Institute of Linguistics, Academia Sinica, Taiwan, and the AI Advanced Talent Cultivation Program of Taiwan's Ministry of Education.

References

Robert Blust. 2013. *The Austronesian Languages*. Asia-Pacific Linguistics.

Council of Indigenous Peoples. 2017. *臺灣南島語言叢書 _1. 阿美語語法概論*. Council of Indigenous Peoples, Taipei, Taiwan.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of NAACL-HLT*, pages 344–354.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Molly Lewis, Aoife Cahill, Nitin Madnani, and James Evans. 2023. Local similarity and global variability characterize the semantic space of human languages. *Proceedings of the National Academy of Sciences*, 120(51):e2300986120.

You Cheng Liao, Chen-Jui Yu, Chi-Yi Lin, He-Feng Yun, Yen-Hsiang Wang, Hsiao-Min Li, and

Yao-Chung Fan. 2024. Learning-from-mistakes prompting for indigenous language translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 146–158, Bangkok, Thailand. Association for Computational Linguistics.

Joseph Lin. 2025. Tackling data scarcity: A practical framework for amis-to-mandarin machine translation. The Fourth Taiwan High School Linguistics Science Fair, National Taiwan Normal University, Taipei, Taiwan.

K. K. Lin, H. Chen, and H. Zhang. 2025. Formosanbench: Benchmarking low-resource austronesian languages in the era of large language models. *arXiv preprint arXiv:2506.21563*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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.

Nathan Schucher, Siva Reddy, and Harm de Vries. 2022. The power of prompt tuning for low-resource semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 148–156, Dublin, Ireland. Association for Computational Linguistics.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800, Bangkok, Thailand. Association for Computational Linguistics.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2022. A parallel corpus and dictionary for Amis-Mandarin translation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 79–84, Taipei, Taiwan. Association for Computational Linguistics.