Can Large Language Models Translate Spoken-Only Languages through International Phonetic Transcription?

Jiale Chen[1], Xuelian Dong[1], Qihao Yang[1], Wenxiu Xie[2], Tianyong Hao[1]*
[1]School of Computer Science, South China Normal University, China
[2]School of Computer Science, Guangdong Polytechnic Normal University, China
Correspondence: jlchen@m.scnu.edu.cn

Abstract

Spoken-only languages are languages without a writing system. They remain excluded from modern Natural Language Processing (NLP) advancements like Large Language Models (LLMs) due to their lack of textual data. Existing NLP research focuses primarily on high-resource or written low-resource languages, leaving spoken-only languages critically underexplored. As a popular NLP paradigm, LLMs have demonstrated strong few-shot and cross-lingual generalization abilities, making them a promising solution for understanding and translating spoken-only languages. In this paper, we investigate how LLMs can translate spoken-only languages into high-resource languages by leveraging international phonetic transcription as an intermediate representation. We propose UNILANG, a unified language understanding framework that learns to translate spoken-only languages via in-context learning. Through automatic dictionary construction and knowledge retrieval, UNILANG equips LLMs with more fine-grained knowledge for improving word-level semantic alignment. To support this study, we introduce the SOLAN dataset, which consists of Bai (a spoken-only language) and its corresponding translations in a high-resource language. A series of experiments demonstrates the effectiveness of UNILANG in translating spoken-only languages, potentially contributing to the preservation of linguistic and cultural diversity. Our dataset and code will be publicly released[1].

1 Introduction

It is estimated that there are approximately 7,000 languages worldwide, the majority of which are considered low-resource. Many of these are spoken-only languages, meaning they have no writing system and rely entirely on oral transmission and communication (Evans and Levinson, 2009; Yang et al., 2023). Moreover, the vast majority of spoken-only languages are used by relatively small populations and fewer than one-tenth of the these languages are spoken by over one million people (Besacier et al., 2014). This presents a significant challenge for preserving linguistic diversity, as spoken-only languages are among the most vulnerable to extinction.

In recent years, text-based Natural Language Processing (NLP) methods have flourished, with Large Language Models (LLMs) emerging as the dominant paradigm (Chang et al., 2024). However, most NLP research like LLM has focused on high-resource languages such as English and Chinese (Ahuja et al., 2023), while studies on low-resource languages tend to explore the written language (Yang et al., 2025). As a result, spoken-only languages are systematically excluded from the advances brought by modern NLP technologies. Extending the benefits of NLP technologies to spoken-only languages remains a fundamental challenge for the field. To the best of our knowledge, we are the first to investigate how spoken-only languages can directly leverage existing NLP technologies.

Due to extreme scarcity of research on spoken-only languages, one viable idea to enable their use in modern NLP systems is to translate them into a written language. Inspired by advance in machine translation, a straightforward method is to adopt a sequence-to-sequence architecture. For example, using Wav2Vec (Baevski et al., 2020) as an encoder and a decoder-only model such as GPT-2 to generate translations. However, such methods suffer from the same limitations faced by other methods targeting low-resource written languages: the lack of large-scale language-paired data leads to overfitting and poor generalization (Yong et al., 2023).

Given these limitations, alternative methods that do not rely on large-scale training data are needed. LLMs have demonstrated strong reasoning capabilities in many NLP tasks (Ghazvininejad et al.,

[*] Corresponding author

[1] https://github.com/Libv-Team/UNI-SO

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23420–23435
November 4-9, 2025 ©2025 Association for Computational Linguistics

2023), and recent study shows promising results in translating low-resource written languages into high-resource ones (Zhang et al., 2024a). These observations suggest that it is possible to leverage LLMs for spoken-only languages by first converting them into an intermediate textual representation. This is particularly important because most LLMs are designed to process text or images, and even speech-enabled models are typically not trained on the audio signals of low-resource languages (Zhang et al., 2024b). As a result, LLMs struggle to interpret such inputs directly.

One potential solution is to transcribe speech into the International Phonetic Alphabet (IPA), which provides a standardized, language-agnostic representation of spoken sounds (Taguchi et al., 2023). However, it remains unclear whether LLMs can effectively learn from such phonetic representations. This poses a challenge for LLMs, as the international phonetic transcription of a spoken-only language represents an unseen input distribution for which the model lacks any prior knowledge.

In this paper, our aim is to explore the feasibility of enabling LLMs to translate spoken-only languages via international phonetic transcription. We introduce a unified language understanding framework UNILANG to efficiently translate a low-resource language to another high resource language via in-context learning. Through automatic dictionary construction and knowledge retrieval, UNILANG provides fine-grained linguistic cues that help LLMs align word-level semantics more accurately during translation. Additionally, we introduce the SOLAN dataset, which contains approximately 2,800 sentences in the spoken-only language Bai (ISO 639-3: bfs) and their corresponding English and Chinese translation. Experiment results on SOLAN benchmark and another public low-resource benchmark ZHUANG indicate our UNILANG achieves state-of-the-art performance on them, providing a potential resolution of preservation for linguistic and cultural diversity. In summary, our main contributions are as follows:

- A unified language understanding framework named UNILANG that enables spoken-only language translation by LLMs. It includes automatic dictionary construction and knowledge retrieval two components, which have proven effective on two benchmarks.
- The first dataset for spoken-only language translation named SOLAN is proposed. It contains IPA-transcribed Bai language sentences and corresponding high-resource language translation.
- A series of experiments demonstrate the feasibility of UNILANG in understanding and translating spoken-only languages, highlighting its value in efforts to preserve and revitalize spoken-only languages.

2 Related Work

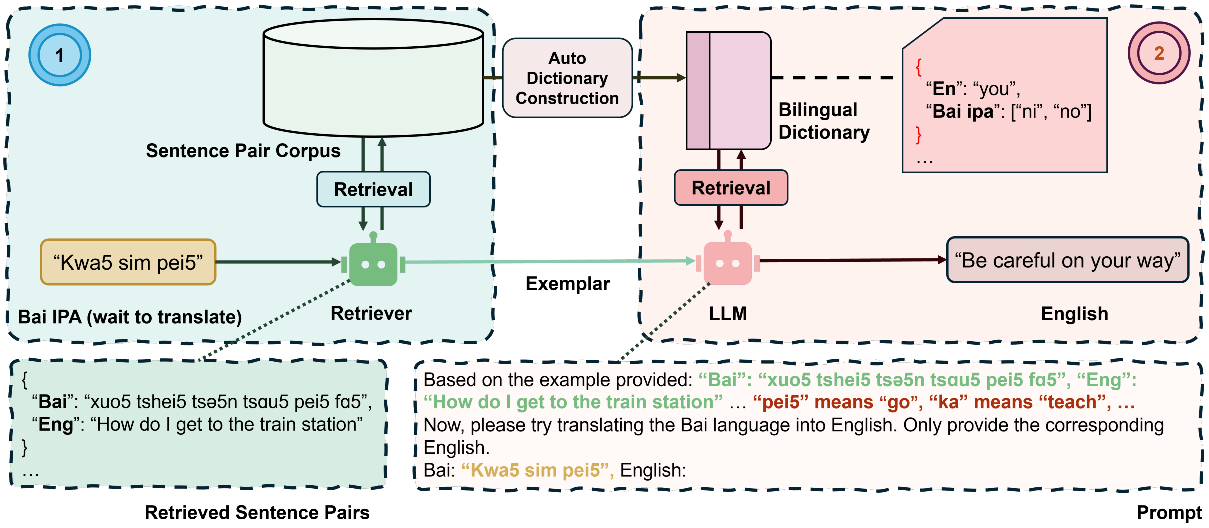
Due to the scarcity of research specifically targeting spoken-only languages, we draw insights from related studies in low-resource language processing and machine translation.

Recent research has shown that LLMs can achieve competitive results in machine translation (MT) tasks, especially when provided with several in-context examples (Zhang et al., 2023). However, such successes are primarily observed in high-high resource language translation.

In contrast, low-resource languages particularly those without sufficient parallel corpora pose challenges for LLM-based translation. To address these challenges, researchers explore various approaches that can be broadly categorized into two types.

The first type approach typically involves supervised fine-tuning of LLMs with cross-lingual instructions. For instance, the UROMAN tool (Hermjakob et al., 2018) was employed for romanization, thereby supporting LLM-based named entity recognition in unseen scripts (Purkayastha et al., 2023). BigTranslate (Yang et al., 2023) proposed a multi-stage optimization pipeline for adapting LLaMA to over 100 languages. The Aya model (Üstün et al., 2024), based on mT5 (Xue et al., 2021), was introduced to support multilingual instruction tuning. Cheng et al. (2024) adapted compact MT models to better support low-resource scenarios. However, these methods are not effective when the training dataset is extremely small.

The second type approach aims to steer LLMs toward better performance without extensive retraining. For example, the PLUG framework (Zhang et al., 2024d) leveraged a high-resource pivot language to refine prompts before translating into the low-resource target language. Jiao et al. (2023) explored pivot-based translation pipelines using high-resource languages as intermediaries. (Elsner and Needle, 2023) explored using GPT-3 with dictionary definitions (instead of parallel corpora) to



```
{
  "En": "you",
  "Bai ipa": ["ni", "no"]
}
...
```

```
{
  "Bai": "xuo5 tshei5 tsə5n tsɑu5 pei5 fɑ5",
  "Eng": "How do I get to the train station"
}
...
```

Based on the example provided: **"Bai": "xuo5 tshei5 tsə5n tsɑu5 pei5 fɑ5", "Eng": "How do I get to the train station"** ... **"pei5" means "go", "ka" means "teach", ...**
Now, please try translating the Bai language into English. Only provide the corresponding English.
Bai: **"Kwa5 sim pei5",** English:

Figure 1: The overall illustration of UNILANG.

translate the polysynthetic language Inuktitut. Similarly, DIPMT (Ghazvininejad et al., 2023) was proposed to incorporate bilingual dictionaries to guide translations, and DIPMT++ (Zhang et al., 2024a) enhanced translation quality through sentence-level supplements. More recently, LingoLLM (Zhang et al., 2024c), a training-free approach was proposed to exploit dictionaries and grammar books to improve LLM translation for endangered languages. Although these studies do not directly address spoken-only languages, the approaches they propose for low-resource written languages offer valuable insights for our work.

3 Method

3.1 Task Definition

We define the task of spoken-only language translation as translating an IPA sentence from a spoken-only low-resource language into a written high-resource language.

Formally, given a bilingual sentence-level parallel corpus $\mathcal{C} = \{(L_i, H_i)\}_{i=1}^{N}$, where L_i is a spoken-only language sentence represented in IPA and H_i is its corresponding translation in a high-resource written language (e.g., English, Chinese), the goal is to learn a mapping ($f : L \rightarrow H$) that generalizes to unseen sentences in the spoken-only language. Here, L_i is represented as a sequence of IPA phonemes: $L_i = [l_i^1, l_i^2, \ldots, l_i^N]$. H_i is a sequence of textual words in the high-resource language: $H_i = [w_i^1, w_i^2, \ldots, w_i^M]$.

3.2 The UNILANG Framework

We propose UNILANG, a unified language understanding framework designed to adapt LLMs to translate spoken-only language to high-resource written language efficiently. It provides a baseline for the machine translation task in the SOLAN benchmark.

The overall illustration of UNILANG is shown in Figure 1. There are two key components in the UNILANG framework, automatic dictionary construction and knowledge retrieval. Given a parallel corpus $\mathcal{C}$, UNILANG first automatically constructs a bilingual dictionary. This step leverages the contextual reasoning capabilities of LLMs to align source-language words with their counterparts in the target language, providing a fine-grained lexical grounding to support subsequent translation. Next, UNILANG retrieve the relevant bilingual sentence pairs and dictionary pairs based on the retrieval strategy and the source sentence. Finally, LLMs generate the target language translation base on the given knowledge from UNILANG.

3.2.1 Automatic Dictionary Construction

Constructing a bilingual dictionary for spoken-only languages is extremely challenging due to the lack of standardized orthography and written resources. To address this, we propose using LLMs to automatically summarize cross-lingual dictionaries. The full process is illustrated in Figure 2.

Written Language Word to Spoken-Only Sentence Given a bilingual sentence-level corpus $\mathcal{C} = \{(H_1, L_1), (H_2, L_2), ..., (H_i, L_i)\}$, we first

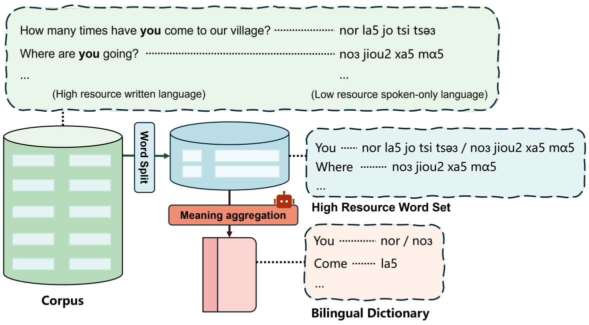

How many times have you come to our village? nor la5 jo tsi tsəɜ
Where are you going? noɜ jiou2 xa5 mɑ5
...
...
(High resource written language)
(Low resource spoken-only language)
Corpus
Word Split
You nor la5 jo tsi tsəɜ / noɜ jiou2 xa5 mɑ5
Where noɜ jiou2 xa5 mɑ5
...
High Resource Word Set
Meaning aggregation
You nor / noɜ
Come la5
...
Bilingual Dictionary

Figure 2: The illustration of automatic dictionary construction.

extract individual words w_i^j from all high-resource written language sentence and form a word set $\mathcal{W}$. Here, w_i^j is indicated that it is the j-th word of the i-th sentence from the corpus $\mathcal{C}$.

For each word w_i^j in this word set $\mathcal{W}$, we maintain the original mapping relationship to the low-resource sentence L_i and construct the word to sentence pair (w_i^j, L_i).

However, the mapping from each w_i^j to a lexical item in the high-resource language is not necessarily one-to-one. For instance, both w_1^2 and w_2^3 may correspond to the same high-resource word, such as "you". To account for this, we apply a merge operation that groups together all instances referring to the same word, resulting in a one-to-many mapping of the form:

$$(w^*, \{L_i \mid w_i^j = w^*\}) \tag{1}$$

where w^* denotes a unique high-resource word, and the set contains all low-resource sentences L_i where any instance w_i^j of w^* occurs. $\mathcal{W}^*$ denotes the set of all such unique high-resource words. This aggregation allows us to associate each high-resource word with a collection of corresponding low-resource sentences, which serves as input to the next stage of "sentence to word".

Spoken-Only Sentence to Spoken-Only Word In the second stage, we aim to construct the final word-level dictionary for the low-resource spoken-only language by leveraging the high-resource constructed in the previous step. Specifically, for each unique high-resource word $w^* \in \mathcal{W}^*$ and its associated set of low-resource sentences $\{L_i \mid w_i^j = w^*\}$, we design a prompt to query a LLM. The prompt is crafted to ask the LLM to infer the most probable spoken-only word(s) or morphemes in the low-resource language that correspond to w^*, based on the surface forms and distributional patterns observed across the associated sentences.

For example, given $w^* =$ *"you"* and its associated low-resource sentence set $\{L_1, L_2, \ldots\}$, the LLM might infer that the most likely spoken forms corresponding to "*you*" are *“nor”* or *“noə”*. As a result, we obtain a set of candidate dictionary entries of the form:

$$(w^*, \{\ell_1, \ell_2, \ldots\}) \tag{2}$$

where each ℓ_k is a hypothesized low-resource word or morpheme aligned to the high-resource word w^*, inferred by the LLM.

By following these two steps, we can construct a bilingual dictionary automatically, without relying on manual annotations for the low-resource language.

3.3 Knowledge Retrieval

Given a spoken-only sentence L_s represented in IPA, our goal is to retrieve relevant sentence pairs from a parallel corpus $\mathcal{C} = \{(L_i, H_i)\}_{i=1}^N$ to construct few-shot exemplars for prompting the LLM. To this end, UNILANG dynamically selects semantically or structurally related exemplars that help the LLM infer lexical and syntactic mappings from context.

While high-resource written languages benefit from a wide range of word-level processing tools, spoken-only languages lack such infrastructure entirely. To address this disparity, we explore two retrieval strategies in the UNILANG framework:

Word Match-Based Retrieval Given the availability of robust tokenization and matching tools for high-resource written languages, we adapt a word-overlap-based retrieval method using the written (high-resource) side of the parallel corpus. For each high-resource translation H_s associated with a spoken-only sentence L_s, we compute token-level overlap with the high-resource side of each corpus pair (H_i, L_i). Sentence pairs with the highest overlap scores are selected as exemplars. This method is simple, fast, and effective when surface-level lexical similarity is a good proxy for underlying semantic similarity, and serves as a strong baseline in low-resource scenarios.

Embedding Similarity-Based Retrieval While training a robust embedding-based retriever is infeasible under extremely low-resource conditions,

Statistic	Value
Avg. IPA per sentence	35.21
Total sentence-pairs	2,803
Contained languages	3
Domain / Topic coverage	Life
Translation directions	Zh-Bai, Eng-Bai

Table 1: Statistics of the SOLAN dataset. "Zh" refers to Chinese, "Eng" to English, and "Bai" to the Bai language, a spoken-only language.

it still can capture coarse-grained semantic similarities across languages and provide a useful semantic signal than lexical retrieval methods such as BM25 (Robertson et al., 2009).

Specifically, we utilize the train dataset to train an encoder by contrast learning. For each spoken-only sentence L_s, we first obtain its embeddings and its corresponding high-resource translation embedding as the positive sample, and other high-resource translation embedding within the same training batch as the negative samples. The overall training description is detailed in Appendix A.

In the validated stage, given the source spoken-only sentence L, we also obtain its embedding H_s using this frozen embedding model, and compute its cosine similarity with all target sentences H_i in the corpus. The top-k most similar pairs (L_i, H_i) are then selected as few-shot exemplars for LLM to generate the final translation, where k is empirically set to 25.

3.4 The SOLAN Dataset

To investigate the capability of LLMs in translating a low-resource spoken-only language, we construct a new dataset named SOLAN, focusing on the Bai language. Bai[2] is a spoken-only, low-resource language. To collect data, we conducted in-person interviews with native speakers.

We recruited six native Bai speakers, aged between 18 and 55, to participate in the data collection process. Since all participants were literate in a high-resource language, we presented sentence prompts in that language and instructed them to orally translate each sentence into Bai and record their speech. Each speaker recorded approximately 500 sentences in a quiet environment, resulting in a total of 2,803 sentence-level audio samples. To facilitate efficient and standardized collection, we developed a custom mobile application to record and export the audio (see Appendix C for implementation details). All recordings were saved in WAV format for further processing.

For phonetic transcription, we first applied an automatic speech-to-IPA conversion model (Xu et al., 2022) to produce an initial phonetic transcription for each spoken-only language sentence. The reason to use this model is that it was trained on multiple languages and their corresponding IPA representations, giving it strong cross-linguistic generalization capabilities, which is an essential feature when dealing with spoken-only languages that lack standardized phonetic resources. These transcriptions were subsequently reviewed and corrected by a linguistic expert[3] to ensure precision.

The SOLAN dataset provides Bai IPA and corresponding high resource language Chinese and English. We select Chinese and English as the target high-resource languages for two main reasons. First, both languages are among the most widely studied and supported in modern NLP research, making them ideal benchmarks for evaluating cross-linguistic performance. Second, Chinese and Bai both belong to the Sino-Tibetan language family, which may offer structural or lexical similarities that facilitate translation (Wang, 2015). In contrast, English belongs to the Indo-European language family, enabling us to assess the challenges of translating between linguistically distant language families. Studying Bai-to-English translation, therefore, provides insights that may generalize to other spoken-only languages from typologically distant families. Table 1 summarizes key statistics of the collected Bai speech data in the SOLAN dataset.

4 Experiments

4.1 Experiment Setup

Dataset The effectiveness of the proposed UNILANG framework is validated by SOLAN and a publicly available low-resource written language dataset ZHUANG (Zhang et al., 2024a). For SOLAN, the data is further divided into training, validation, and test sets following an 8:1:1 ratio. For ZHUANG, we follow the official split of original dataset[4].

Backbone In this work, we utilize both open-source and close-source LLMs as the backbone

[2] The detailed introduction can be found in Appendix B.

[3] Education background can be found in Appendix D.

[4] More information about the setting of datasets can be found in Appendix E.

Model	Chinese → Bai			Bai → Chinese				English → Bai			Bai → English			
	BLEU	chrF++	chrF	BLEU	chrF++	chrF	SBERT-score	BLEU	chrF++	chrF	BLEU	chrF++	chrF	SBERT-score
Zero Shot														
Qwen-2.5-7B	0.00	0.27	0.18	0.26	2.10	1.41	27.36	0.24	3.26	2.18	0.06	7.45	9.26	19.88
Llama-3.1-8B	0.19	1.76	1.18	0.10	0.91	0.61	28.54	0.28	5.08	3.40	0.04	4.34	4.96	19.58
Gemma-3-12B	0.08	1.44	0.96	0.30	1.80	1.20	32.52	0.07	2.23	1.49	0.32	6.89	8.28	27.88
Few Shot														
Qwen-2.5-7B	22.36	27.63	22.45	1.23	3.64	2.74	28.19	17.14	24.66	19.69	0.60	11.40	13.57	25.74
Llama-3.1-8B	20.35	23.02	18.54	1.32	2.83	2.08	33.62	1.01	5.09	3.87	0.44	9.51	11.08	27.85
Gemma-3-12B	19.09	24.47	19.51	0.53	2.55	1.71	34.52	18.60	23.98	18.97	1.49	11.37	13.15	30.38
DeepSeek-V3	25.51	32.85	27.59	2.25	5.33	4.12	34.04	2.68	12.15	9.34	0.94	11.31	13.20	31.76
DIPMT++														
Qwen-2.5-7B	23.14	27.33	22.49	2.79	4.50	3.37	34.24	19.40	25.42	20.46	0.13	9.55	11.58	26.21
Llama-3.1-8B	25.67	27.28	22.74	4.22	5.43	4.23	33.62	10.35	20.18	15.82	0.21	10.69	12.54	30.16
Gemma-3-12B	21.63	27.37	22.02	5.60	7.15	5.72	38.94	19.70	25.13	20.40	0.15	8.29	9.94	25.90
GPT-4o	27.83	32.08	26.77	9.07	10.83	9.27	39.57	19.08	25.03	20.19	0.15	7.96	9.82	28.86
DeepSeek-V3	25.51	31.74	26.60	7.29	11.30	9.45	41.55	3.21	13.55	10.62	0.15	8.34	10.14	31.77
Claude	26.65	30.74	25.56	2.75	7.58	6.19	34.43	11.56	22.61	18.34	0.23	8.31	9.8	28.75
Gemini	27.07	30.06	24.78	11.35	12.14	10.06	43.89	17.02	24.17	19.83	0.94	10.09	11.92	32.46
UNILANG (Ours)														
Qwen-2.5-7B	24.23	29.32	24.44	5.61	8.82	7.10	37.35	20.55	26.10	21.18	2.51	14.10	16.10	33.97
Llama-3.1-8B	28.25	29.10	24.50	6.37	7.16	5.82	39.46	16.57	22.30	18.61	0.77	12.51	14.60	28.89
Gemma-3-12B	24.45	29.74	24.97	5.29	6.74	5.12	39.17	22.12	24.04	19.57	1.14	10.37	12.04	30.27
GPT-4o	27.15	32.10	26.85	8.75	10.05	8.82	38.49	18.67	25.15	20.35	0.63	11.94	13.79	31.11
DeepSeek-V3	24.75	**33.88**	**28.71**	8.24	14.22	**11.90**	46.04	3.02	13.14	10.19	0.95	11.97	13.46	33.23
Claude	28.80	30.69	25.72	11.17	12.15	10.02	44.78	12.24	24.02	19.52	1.71	14.63	16.16	36.53
Gemini	**29.71**	31.93	26.96	**13.28**	**14.30**	11.79	**49.29**	**26.52**	**28.14**	**23.38**	**4.11**	**16.78**	**18.44**	**39.08**

Table 2: Translation performance between Chinese, English, and Bai language on the SOLAN test set. **Boldface** denotes the highest score and underlining denotes the second-highest score.

of the proposed UNILANG framework and other baseline methods: (1) **Open-Source Models** Like Llama-3.1 series, (Grattafiori et al., 2024) Qwen-2.5 (Yang et al., 2024), DeepSeek-V3 (Liu et al., 2024), and Gemma (Team et al., 2025). (2) **Close-Source Models** Claude, GPT, Gemini.

This integration of both open and commercial LLMs allows UNILANG to flexibly adapt to spoken-only language translation tasks while optimizing performance across high-resource and low-resource spoken-only language directions.

Baselines We adapt several baseline methods for comparison. (1) Zero-shot: This method directly ask LLMs to perform translations without any examples from corpus. This method can reflect whether the LLMs already know the language. (2) Few-shot: Given the relevant sentence pairs from corpus, this method provides some in-context learning samples for LLMs. (3) DIPMT++ (Zhang et al., 2024a): This method proposes an efficient strategy for retrieving exemplars from the corpus to support in-context learning for LLMs.

Evaluation We use BLEU (Papineni et al., 2002), chrF (including chrF++) (Popović, 2015) and SBERT-score (Zhang et al., 2019). for evaluation. BLEU measures the n-gram precision between machine-generated and reference translations. chrF computes the F-score over character n-grams. SBERT-score measures semantic similarity by computing the cosine similarity of sentence embeddings produced by a pretrained Sentence-BERT model. All metrics are standard tools for assessing translation quality in machine translation.

4.2 Main Results

We report the comparison results on the SOLAN dataset in Table 2. Across all methods, we observe a consistent trend: translation from high-resource languages to the spoken-only language (Zh→Bai, Eng→Bai) yields better performance, while the reverse direction (Bai→Zh, Bai→Eng) remains comparatively weaker. This highlights the greater challenge of generating fluent high-resource language from spoken-only language. Compared to vanilla sentence-pairs methods (including Few-Shot and DIPMT++), our UNILANG framework automatically constructs a dictionary from the sentence pair corpus and utilizes the dictionary to provide fine-grained language knowledge. In the results[5], the UNILANG framework achieves the best performance in all evaluation metrics. Most LLMs within

[5] The output samples of all methods are presented in Appendix F.

Model	**Chinese → Zhuang**		**Zhuang → Chinese**	
	BLEU	chrF	BLEU	chrF
		Few Shot		
Qwen2.5-7B	1.46	23.24	0.14	1.10
Llama-3-8B	2.23	24.32	0.19	1.22
Gemma-3-12B	2.09	24.17	0.34	1.41
		DIPMT++		
Qwen2.5-7B	2.49	25.37	6.88	7.36
Llama-3-8B	2.80	22.94	4.01	5.85
Gemma-3-12B	3.04	24.81	6.43	6.75
		UNILANG (Ours)		
Qwen2.5-7B	3.93	<u>32.91</u>	**9.01**	<u>9.32</u>
Llama-3-8B	<u>5.40</u>	30.70	6.21	8.55
Gemma-3-12B	**7.27**	**35.00**	<u>8.73</u>	**9.62**

Table 3: Translation results in ZHUANG. **Boldface** denotes the highest score and <u>underlining</u> denotes the second-highest score.

our UNILANG framework achieve better performance compared to the baselines. For the Chinese → Bai task, Gemini within UNILANG obtains a 9.8% improvement in BLEU over DIPMT++. In the hardest task, Bai → English, Gemini with UNILANG also obtains the best score 4.11 in BLEU, while Gemini with DIPMT++ only achieves 0.94 BLEU score. These results highlight the effectiveness of UNILANG in improving LLM translation quality for spoken-only languages, especially in low-resource and typologically diverse scenarios.

To further validate the generalizability of our framework UNILANG, we conduct additional experiments on a different low-resource language dataset. Due to the limited availability of publicly accessible spoken-only language resources, we select ZHUANG, a low-resource written language, for this study. The experiment results are presented in Table 3. These results show that UNILANG achieves the best performance on this written language dataset, further supporting the effectiveness and applicability of our framework.

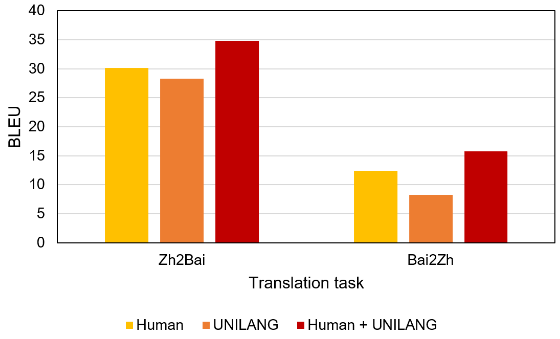


Figure 3: The performance of human translation comparison. The UNILANG uses Qwen-2.5-7B as backbone LLM.

4.3 Ablation Study

To investigate which components of our UNILANG framework contribute the most to its performance, we conduct an ablation study using Qwen-2.5-7B as the backbone model. The evaluation is carried out on SOLAN dataset with two translation tasks: English → Bai and Bai → English. As shown in Table 4, the automatic dictionary construction component plays a critical role in enhancing translation quality[6], particularly in the high-resource to low-resource direction (English → Bai), where it provides fine-grained lexical knowledge. In contrast, in the low-resource to high-resource direction (Bai → English), both dictionary construction and knowledge retrieval components are essential to achieve strong performance. Interestingly, we observe a slight increase in chrF and chrF++ scores after removing the dictionary module. We attribute this to the nature of the IPA, which uses Latin characters: even random or meaningless sequences tend to yield non-zero character-level scores, potentially inflating chrF metrics despite a drop in true translation quality.

4.4 Human Translation Comparison

To evaluate the effectiveness of our framework, we conduct an experiment comparing the performance of our UNILANG, human[7], and a combination of both. In this setup, humans are provided with 5 Zh-Bai audio pairs as reference examples to perform the translation task. In contrast, LLM receives the corresponding 5 IPA-Chinese text pairs as few-shot exemplars. The combined setting allows humans to revise translations generated by the LLMs. The results, presented in Figure 3. In these two tasks, human perform well in this few-shot learning scenario, while LLM with UNILANG is relatively weaker. However, when LLM with UNILANG generate reference translations to assist human, the overall performance improves, achieving a 2.65% increase in BLEU score. These findings highlight the complementary strengths of human intuition

[6] To further results that compare this LLM-based method with statistic method can be found in Appendix G.

[7] The human evaluators are native Chinese speakers with no prior exposure to the Bai language.

Model	English → Bai			Bai → English		
	BLEU	chrF++	chrF	BLEU	chrF++	chrF
UNILANG	20.55	26.10	21.18	2.51	14.10	16.10
w/o Auto Dictionary Construction	19.40 (1.15↓)	25.42 (0.68↓)	20.46 (0.72↓)	1.85 (0.66↓)	14.16 (0.06↑)	16.13 (0.03↑)
w/o Knowledge Retrieval	20.45 (0.10↓)	24.69 (1.41↓)	20.20 (0.98↓)	1.67 (0.84↓)	12.99 (1.11↓)	15.05 (1.05↓)

Table 4: Ablation study on the SOLAN dataset for both English→Bai and Bai→English translations.

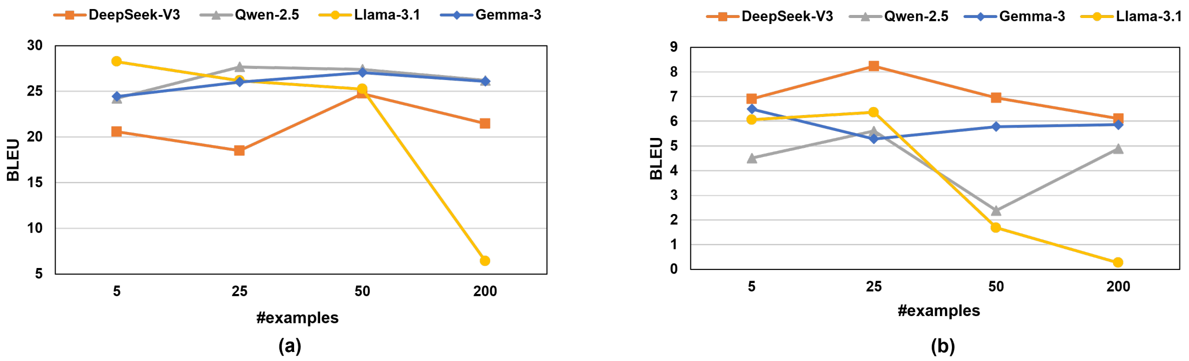


Figure 4: The results of impact experiments for lingual corpus size. (a) is Chinese → Bai, (b) is Bai → Chinese.

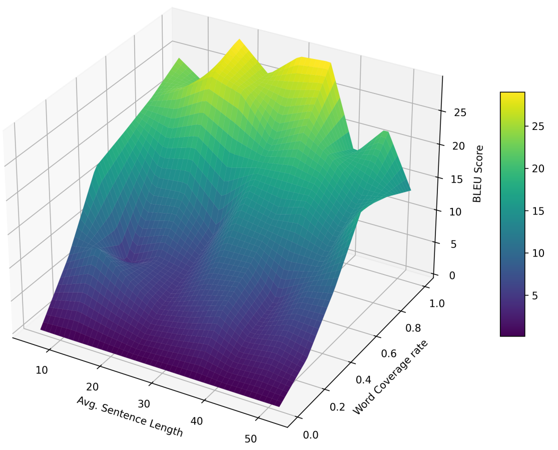


Figure 5: The experiment result of exploring two effect dimensions.

and machine-generated assistance, suggesting that collaborative approaches may be especially promising for low-resource spoken-only language translation tasks.

4.5 The Impact of the Size of Lingual Corpus

To balance translation performance and computational cost, we examine how input size affects LLM reasoning. The results are shown in Figure 4. These results indicate that input size influences the performance of LLMs within UNILANG. In particular, using 25 exemplars yields the best performance, suggesting this quantity strikes an effective balance between providing sufficient context and staying within input length constraints. When 200 examples were input to Gemma-3, the performance dropped sharply compared to smaller input sizes. We hypothesize that when the number of examples is large, many of them are irrelevant to the translation task, thereby negatively impacting performance. In summary, huge examples as input are heavily depended on the reasoning ability of LLMs. The more huge size of input examples are, the more noisy example may be induced. If the reasoning ability of LLMs is weak, it may have a bad performance.

4.6 Influence of Sentence Complexity on Translation

By comparing Table 3 and Table 2, we find that translating from Chinese to low-resource languages yields lower performance in Table 3, despite both targeting low-resource outputs—colloquial text vs. IPA transcription. To explore this discrepancy, we conduct controlled experiments on the Chinese-Bai bi-directional task using Qwen2.5-7B, as shown in Figure 5. The result suggests that for languages unseen during LLM training, the final translation performance is primarily influenced by two factors: (1) word coverage, and (2) sentence length of word-level translations. As shown in the figure, when word coverage is zero—i.e., none of the words in the target sentence are present in the LLM's vocabulary—translation quality remains consistently poor, regardless of the sentence length. This indicates that word coverage plays a critical role in enabling effective transfer to unseen languages.

In addition, sentence length also plays a key role. Very short sentences often fail to provide enough contextual information for the LLM to perform reliable reasoning, leading to suboptimal translations. On the other hand, overly long sentences may exceed the LLM's capacity for maintaining coherence and managing dependencies over extended input, thus degrading performance. Empirically, we find that sentence-level translations with an average word length between 20 and 30 yield the best results, striking a balance between sufficient context and model capacity.

5 Conclusion

In this paper, we explore the potential of LLMs to translate spoken-only languages using international phonetic transcription. To address the challenges of low-resource and unseen spoken-only language translation, we propose UNILANG, a unified language understanding framework for on-the-fly language learning. Additionally, we introduce SOLAN, a new benchmark comprising the Bai language, where speech data is transcribed into IPA to enable textual processing. Experimental results demonstrate that UNILANG consistently outperforms baselines on both translation directions in SOLAN, and its strong generalizability is further validated on Zhuang, an unseen low-resource written language. These findings underscore the potential of LLM-based frameworks in bridging the gap between spoken-only languages and modern NLP and provide a potential method for the research and preservation of spoken-only languages.

Acknowledgements

The work is supported by grants from National Natural Science Foundation of China (No. 62372189) and the Scientific Research Innovation Project of Graduate School of South China Normal University (2025KYLX092).

Limitations

Due to budget constraints, this study was only able to collect a commonly used corpus for a single spoken-only language (Bai). Additionally, since this is the first attempt to annotate Bai using the IPA, some transcription inaccuracies may exist. While using IPA as an intermediate representation makes it possible to interface with LLMs, it remains uncertain whether this approach can be generalized to all spoken-only languages, and further empirical validation is required. Finally, although our proposed automatic dictionary construction method provides more fine-grained lexical knowledge to LLMs, it may also suffer from potential error propagation throughout the pipeline.

Ethics Statement

The data of SOLAN dataset was collected from everyday, non-sensitive language contexts and does not contain personally identifiable information.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, et al. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Xin Cheng, Xun Wang, Tao Ge, Si-Qing Chen, Furu Wei, Dongyan Zhao, and Rui Yan. 2024. Scale: Synergized collaboration of asymmetric language translation engines. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15903–15918.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Micha Elsner and Jordan Needle. 2023. Translating a low-resource language using gpt-3 and a human-readable dictionary. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–13.

Nicholas Evans and Stephen C Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(5):429–448.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal romanization tool uroman. In *Proceedings of ACL 2018, system demonstrations*, pages 13–18.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

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.

Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. Romanization-based large-scale adaptation of multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. 2023. Universal automatic phonetic transcription into the international phonetic alphabet. In *Proc. Interspeech 2023*, pages 2548–2552.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939.

Feng Wang. 2015. Language contact between tibeto-burman languages and chinese. *The Oxford Handbook of Chinese Linguistics*, page 248.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and effective zero-shot cross-lingual phoneme recognition. In *Proc. Interspeech 2022*, pages 2113–2117.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. 2025. Nüshurescue: Reviving the endangered nüshu language with ai. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7020–7034.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, et al. 2023. Bloom+ 1: Adding language support to bloom for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, pages 41092–41110.

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

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024b. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12401–12430.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024c. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024d. Plug: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046.

A Contrastive Learning for Embedding Similarity-Based Retrieval

To align English text and its corresponding phonetic representation of the Bai language, we adopted a contrastive learning framework based on a shared Transformer backbone. Specifically, we utilized a pretrained BERT model from HuggingFace[8] as the base encoder for English text and Bai IPA input. The training objective is to map the semantically and phonetically corresponding pairs into a shared embedding space.

We used the SOLAN dataset for training. During training, both inputs are independently tokenized and encoded through the same Transformer encoder. To obtain fixed-size sentence embeddings, we applied mean pooling over the last hidden states of the model, followed by L2 normalization.

Our loss function is a symmetrical contrastive loss based on the InfoNCE principle, treating all other samples in the batch as negatives (i.e., Multiple Negatives Ranking Loss). For a given batch of size N, we compute a similarity matrix $S \in \mathbb{R}^{N \times N}$ between the normalized English embeddings $\{\mathbf{e}_i\}$ and IPA embeddings $\{\mathbf{z}_j\}$ as follows:

$$S_{ij} = \frac{\mathbf{e}_i \cdot \mathbf{z}_j}{\tau}, \tag{3}$$

where τ is a temperature hyperparameter. The contrastive loss is defined as the average of the bidirectional InfoNCE losses:

$$\mathcal{L}_{\text{en}\to\text{ipa}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{N} \exp(S_{ij})}, \tag{4}$$

$$\mathcal{L}_{\text{ipa}\to\text{en}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{N} \exp(S_{ji})}, \tag{5}$$

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2}\left(\mathcal{L}_{\text{en}\to\text{ipa}} + \mathcal{L}_{\text{ipa}\to\text{en}}\right). \tag{6}$$

This objective encourages aligned pairs to have high cosine similarity while pushing apart mismatched pairs within the same batch.

Finally, We trained the model using the AdamW optimizer with a learning rate of 2e-5 and a batch size of 32, for a total of 64 epochs. To evaluate alignment quality, we used retrieval accuracy as the primary validation metric: for each English query in the validation set, we identified the most similar IPA vector and measured top-1 accuracy. The model checkpoint with the highest validation accuracy was saved.

This method effectively learns a joint embedding space where phonetically similar text pairs in different scripts are closely aligned.

B The Bai Language

Bai is a Sino-Tibetan language spoken in China, primarily in Yunnan Province, by the Bai people. The language has over a million speakers and is divided into three main dialects, which may actually be distinct languages: Jianchuan (Central), Dali (Southern), and Bijiang (Northern). The dataset used in this study was primarily collected from speakers of the Jianchuan dialect, which serves as a central and representative variant of the Bai language.

C App Design

To facilitate data collection, we have developed a WAV voice recording application specifically designed for capturing spoken-only utterances. The software is installable on mobile devices and was developed using the Dart programming language and the Flutter framework[9]. It supports multiple platforms, including iOS, Android, and Windows. The screenshot of our App is shown in Figure 6.

D Background of Annotation Participants and Linguistic Expertise

The SOLAN dataset was constructed with the help of six native Bai speakers whose educational backgrounds ranged from secondary school to postgraduate degrees. The linguistic expert responsible for reviewing the transcriptions holds a Ph.D. in linguistics with a specialization in English.

E Settings of Dataset

For fair evaluation on the SOLAN dataset, we ensure that the vocabulary of the test set is a subset of that of the training set. Additionally, we maintain a similar distribution of sentence lengths between the training and test subsets to control for potential confounding factors. For other publicly available datasets, such as ZHUANG, we follow the original experimental settings established in prior work. Notably, in this experiment, ZHUANG includes only

[8] https://huggingface.co/

[9] https://flutter.dev/

a bilingual sentence corpus, allowing us to better simulate the challenges inherent in low-resource language scenarios.

F Samples of Main Results

Translation examples from Bai to Chinese using different systems are shown in Table 5, including both reference (Golden) and system-generated outputs (DMIPT++ and UNILANG).

G Comparison with Statistical Word Alignment

We further compare the LLM-based automatic dictionary construction method with a statistical word alignment baseline implemented using the IBM model–based approach (Dyer et al., 2013). Specifically, we extract word correspondences via the statistical method and apply them to UNILANG under two settings: (i) using the extracted dictionary directly for translation, and (ii) replacing the LLM-inferred correspondences with the statistical ones. As shown in Figure 7, the statistical method provides a reasonable baseline, but overall underperforms compared to LLM-based dictionary construction. For instance, in Bai → Chinese translation, Gemma-3-12B with statistical correspondences achieves BLEU 6.72, slightly higher than its LLM-based counterpart (5.29), whereas for Qwen2.5-7B and Llama-3-8B the LLM-based correspondences consistently lead to better BLEU and chrF scores. The advantage of LLM-based alignment is more evident in Bai → English translation, where the LLM approach clearly surpasses the statistical baseline. These results indicate that while statistical alignment captures certain word-level consistencies, UNILANG benefits more from the semantically richer and context-sensitive correspondences provided by LLMs.

H Instructions Given To Participants

Participants were instructed to read aloud the on-screen content in the Bai language and to record their speech accordingly. The primary objective of this process was to obtain audio samples suitable for transcription into IPA representations specific to the Bai language. Prior to their participation, all individuals provided informed consent, explicitly agreeing to the use of non-personally identifiable data for research and dataset development. Importantly, the audio recordings will not be released in any public dataset. We are committed to safeguarding participant privacy, including the protection of biometric voice features such as voiceprints.

Bai: xɑu5 wɑ5

Chinese (Golden): 好啊 (Okay)

Chinese (DMIPT++): 好啊 (Okay) ✓

Chinese (UNILANG): 好呀 (Okay) ✓

Bai: u kʉ ti te

Chinese (Golden): 现在几点了 (What time is it now?)

Chinese (DMIPT++): 去哪里 (Where are you going?) ✗

Chinese (UNILANG): 现在几点了 (What time is it now?) ✓

Bai: si5 si5n na5

Chinese (Golden): 谢谢您 (Thank you)

Chinese (DMIPT++): 现在下雨 (It is raining now) ✗

Chinese (UNILANG): 谢谢您 (Thank you) ✓

Bai: puo5 si5ŋ də1

Chinese (Golden): 不行的 (No way)

Chinese (DMIPT++): 下雨了 (It is raining) ✗

Chinese (UNILANG): 行不通 (That’s not feasible) ✓

Bai: tso5 tə1 jiɜ ŋ ka5n pu5 si5x wa5ŋ tsha5ŋ kou5

Chinese (Golden): 有的人不喜欢唱歌 (Some people don’t like singing)

Chinese (DMIPT++): 还有的人不喜欢唱歌 (There are also people who don’t like singing.) ✓

Chinese (UNILANG): 还有的人不喜欢唱歌 (There are also people who don’t like singing.) ✓

Bai: nong5 jiə5 kuoɜ tai5 pu5 tshuo5

Chinese (Golden): 你也过得不错 (You’re doing well too)

Chinese (DMIPT++): 农家锅太不粗 (**A sentence without meaning cannot be translated**) ✗

Chinese (UNILANG): 我一句过得不错 (I’m doing pretty well in a word.) ✗

Bai: non5 gu5 pai5n tai5 kha5n sou5 si5

Chinese (Golden): 我来办贷款手续 (I’m here to handle the loan procedures)

Chinese (DMIPT++): 农民在看电视剧 (The farmer is watching a television series) ✗

Chinese (UNILANG): 你陪我来贷款 (You came with me to apply for a loan) ✗

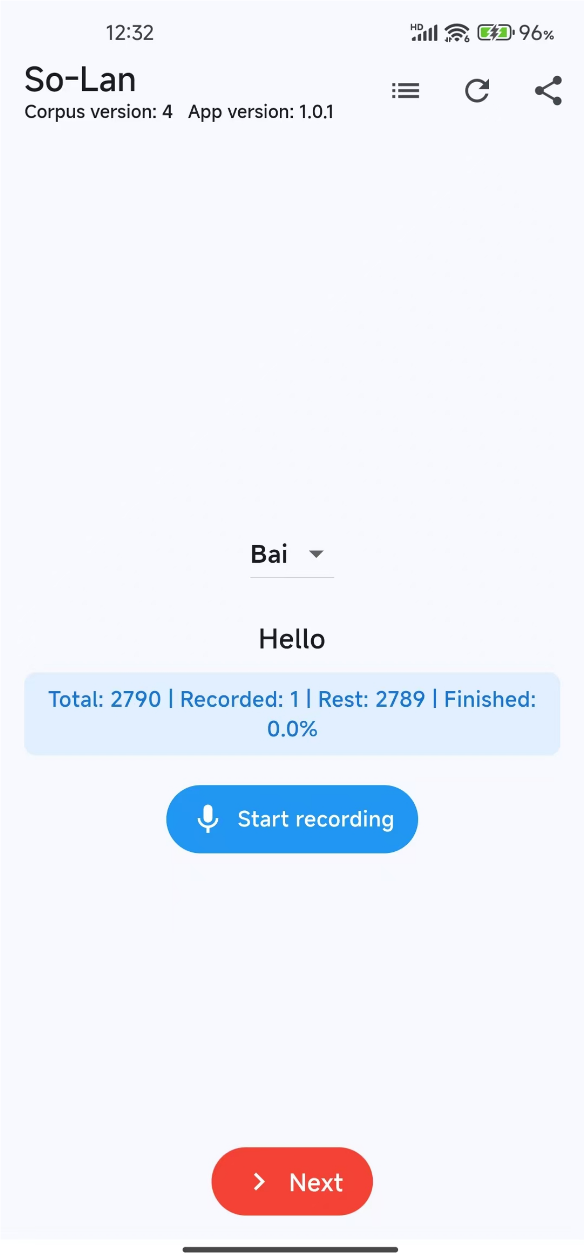
Bai: ni meː rʌm mʌn nʌ

Chinese (Golden): 你叫什么名字 (What’s your name)

Chinese (DMIPT++): 你没有钱 (You have no money.) ✗

Chinese (UNILANG): 你有多少个 (How many do you have?) ✗

Table 5: Translation examples from Bai to Chinese using different systems. The reference translations (Golden) are compared against outputs from DMIPT++ and UNILANG.



12:32 96%

So-Lan

Corpus version: 4 App version: 1.0.1

Bai

Hello

Total: 2790 | Recorded: 1 | Rest: 2789 | Finished: 0.0%

Figure 6: The screenshot of our App.

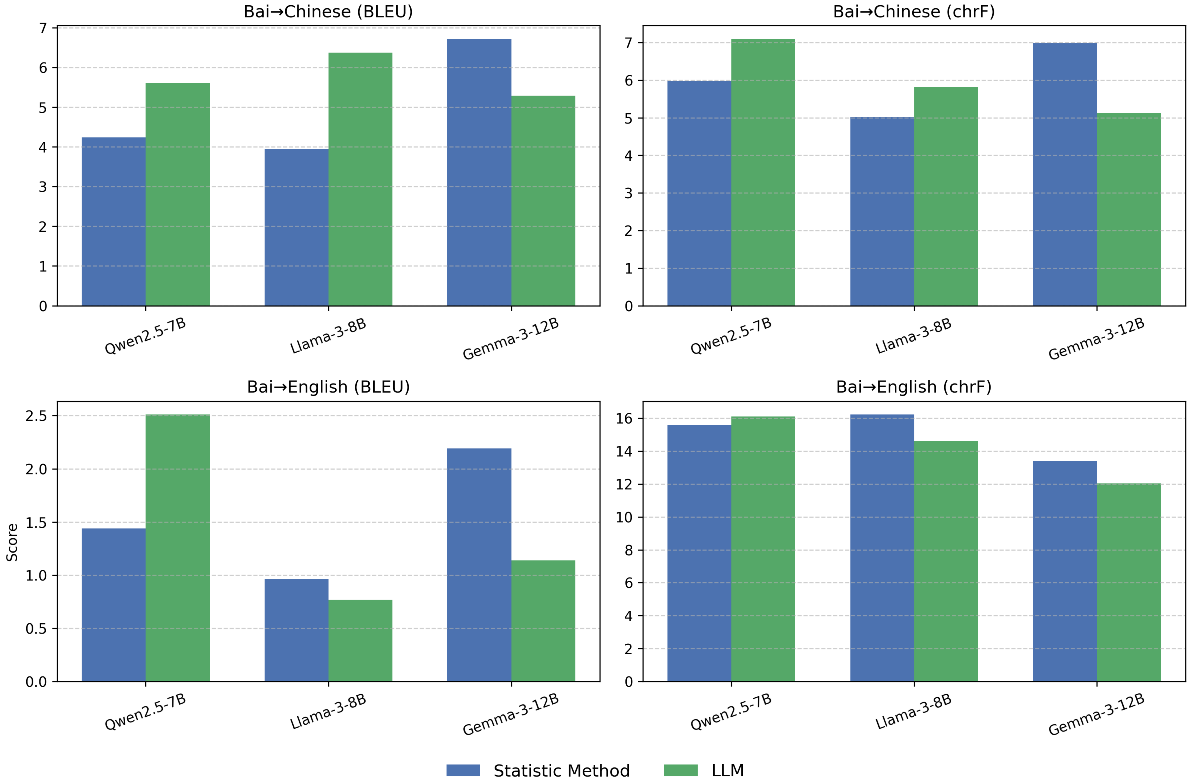


Figure 7: The results of comparison with statistical word alignment.