A Unified Framework for Low-Resource Sawndip–Chinese Bidirectional Translation and Bilingual Dialogue Modeling

Anonymous ACL submission

Abstract

The decline of proficient practitioners poses an existential threat to numerous ancient writing systems, risking the loss of the cultural heritage they encode. Among these systems is Sawndip, a logographic script whose computational preservation faces severe challenges due to the scarcity of standardized corpora. In this paper, we present the first systematic framework for low-resource Sawndip-Chinese translation and bilingual conversational modeling. We introduce a newly curated dataset comprising sentence-level Sawndip-Chinese parallel data and character-level lexical definitions. Building upon this foundation, we propose a knowledge-enhanced supervised fine-tuning paradigm that incorporates lexical knowledge during training. To address the additional challenges of Chinese-to-Sawndip generation, we further develop a manually curated template-based substructure substitution strategy for synthesizing reverse-direction data. Bilingual conversational data is then constructed through a dimension-conditioned dialogue generation and multi-candidate back-translation pipeline, enabling controllable and diverse dialogue synthesis. Experimental results show that our final unified model effectively supports both Sawndip-Chinese translation and bilingual conversational abilities. Ultimately, this work provides a foundational reference for the digital preservation of Sawndip, offering broader insights into the computational documentation of globally-endangered languages.

1 Introduction

The preservation of endangered languages is a critical global imperative, as these languages serve as important carriers of cultural diversity and historical knowledge. Estimates suggest that over 60% of the world's roughly 7,000 languages are currently endangered (Moseley, 2010). Among these threatened systems is **Sawndip** (Old Zhuang Script), a traditional logographic writing system used by the Zhuang People, an ethnic group primarily residing in southern China. Historical studies suggest that Sawndip can be traced back at least to the Tang dynasty (7th century CE), giving it a documented history of more than 1,300 years (Diller et al., 2004). As an essential carrier of Zhuang culture, Sawndip holds significant value for historical linguistics and ethnographic research. However, with the promotion of the modern Zhuang romanization system and shifting sociolinguistic environments, its active usage has steadily declined. Today, Sawndip survives primarily in a limited number of historical documents and a dwindling circle of individual practitioners, while digitally annotated resources remain extremely scarce. Consequently, Sawndip represents both a low-resource and endangered writing system, highlighting the urgency of systematic computational documentation to safeguard this component of global cultural heritage.

While Large Language Models (LLMs) have demonstrated remarkable capabilities in machine translation (Costa-Jussà et al., 2022; Vilar et al., 2023) and instruction following (Ouyang et al., 2022; Dubey et al., 2024), Sawndip presents unique challenges that render standard fine-tuning ineffective. The script possesses an *immense character inventory* with many local and historical variants, most of which are not covered by Unicode. In addition, the system exhibits *severe semantic ambiguity*, characterized by extensive one-to-many glyph-to-meaning mappings. Finally, the limited historical corpora suffer from *substantial orthographic noise*, complicating model training and generation. Consequently, frontier models such as GPT-5.5 (OpenAI, 2025), Claude Opus 4.7 (Anthropic, 2026), and even Chinese-focused systems like DeepSeek-v4-Pro (DeepSeek AI, 2026) cannot reliably translate or generate Sawndip based on their native knowledge.

To address these challenges, we present the first unified framework for Sawndip-Chinese bidirec-

tional translation and bilingual dialogue modeling. We first construct a curated dataset, consisting of Sawndip-Chinese parallel sentences and Sawndip character-level lexical definitions collected from lexical databases and historical materials. Based on this dataset, we investigate machine translation and propose a knowledge-enhanced supervised fine-tuning (KE-SFT) strategy that incorporates character-level lexical information during training to improve Sawndip-to-Chinese translation. For Chinese-to-Sawndip translation, where reverse lexical resources are unavailable, we construct manual templates and further expand them via LLM-guided substructure substitution for synthetic data augmentation to alleviate data scarcity.

Beyond translation, we argue that preserving an endangered writing system requires not only sentence-level translation, but also **interactive bilingual generation** capable of conveying cultural knowledge and historical context. Motivated by this goal, we further extend the translation framework into a bilingual dialogue system. To support dialogue training under noisy low-resource conditions, we propose a dimension-conditioned dialogue synthesis pipeline for generating dialogues aligned with Sawndip cultural and linguistic contexts, together with a multi-candidate back-translation filtering strategy for constructing high-quality bilingual dialogue data. Finally, we jointly train the model on both translation and dialogue resources, enabling a unified system for both bidirectional translation and bilingual interaction under extreme low-resource conditions.

To the best of our knowledge, this is the first study to explore Sawndip through a systematic, data-driven computational framework. We hope that this work can serve as an initial step towards the computational study of Sawndip and, more broadly, raise awareness of the importance of collecting, documenting, and preserving endangered, historically grounded languages globally. Our core contributions are summarized as follows:

- We curated the first Sawndip–Chinese computational resource, including authentic parallel sentence pairs, lexical definitions, and synthetic bilingual dialogue data, and introduce Sawndip-Chinese machine translation and bilingual dialogue benchmarks.

- We propose a KE-SFT framework for Sawndip-to-Chinese translation and an LLM-guided substructure substitution method with manually designed templates for synthetic Chinese-to-Sawndip data augmentation under extreme low-resource conditions.

- We introduce a dimension-conditioned dialogue synthesis pipeline together with multi-candidate back-translation filtering to construct culturally coherent bilingual dialogue data for Sawndip.

- Extensive experiments demonstrate that our framework significantly improves bidirectional translation quality, outperforming strong commercial LLM baselines, while also enabling coherent bilingual dialogue generation.

2 Related Work

Computational Studies of Low-Resource and Endangered Languages The preservation of linguistic diversity has motivated increasing Natural Language Processing (NLP) research on low-resource and endangered languages. While foundational work remains a cornerstone of the field, recent studies also emphasize corpus construction, benchmark design and computational modeling for under-explored languages, scripts, and dialects. For instance, Zheng and Yu (2025) introduced the first Tangut-Chinese parallel corpus, utilizing LLM-based translation with lexicon-aligned prompting. In the domain of dialectology, Le and Tuan (2023) presented a parallel corpus for central-northern Vietnamese dialect transfer and evaluating monolingual and multilingual language models on this task. Beyond Asian contexts, datasets for indigenous languages like Mapudungun (Pendas et al., 2023), supported by initiatives like the AmericasNLP Workshop (Mager et al., 2024), further advance research on cross-lingual transfer. These studies highlight growing interest in computational modeling of under-explored languages and scripts. Building on this momentum, our work introduces the first systematic framework for Sawndip, a historically-grounded but digitally neglected writing system.

Low-Resource Neural Machine Translation Low-resource machine translation remains challenging due to the scarcity of parallel corpora or linguistic resources required for extensive pre-training. Prior work has explored a variety of strategies in neural machine translation (NMT) settings (Ren et al., 2021; Haddow et al., 2022; Ranathunga et al., 2023). Traditional approaches leverage data

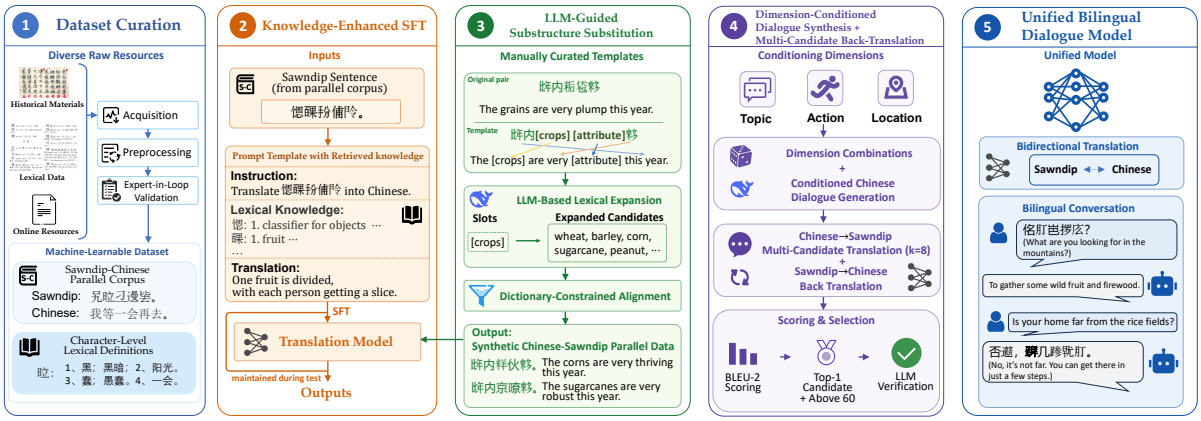

Figure 1: Overview of our unified framework for low-resource Chinese-Sawndip bidirectional translation and bilingual dialogue modeling. All data and supervision are in Chinese-Sawndip format. Our framework consists of five stages: (i) data construction and curation from diverse sources with expert evaluation; (ii) knowledge-enhanced SFT for Sawndip-to-Chinese translation with character-level lexical grounding; (iii) manually designed templates combined with LLM-guided substructure substitution for synthetic Chinese-to-Sawndip data augmentation; (iv) dimension-conditioned dialogue synthesis combined with multi-candidate back-translation filtering for constructing high-quality bilingual dialogue data; and (v) a unified model jointly trained on translation and dialogue data, enabling bidirectional translation and bilingual interaction under extremely low-resource conditions.

augmentation techniques such as back-translation (Sennrich et al., 2016; Fadaee and Monz, 2018), which synthesize pseudo-parallel samples to bootstrap translation models. Other paradigms focus on knowledge transfer, including pivot-based translation (Cheng, 2019; Kim et al., 2019), which utilizes an intermediate language for indirect alignment, or transfer learning (Zoph et al., 2016; Dabre et al., 2017; Neubig and Hu, 2018), where models pretrained on high-resource language pairs are adapted to low-resource targets. Furthermore, Multilingual NMT (Arivazhagan et al., 2019; Fan et al., 2021) leverages shared representations across languages to improve performance on tail languages. While effective, these methods often struggle with scripts like Sawndip, as its highly complex writing system hinders reliable cross-lingual mapping.

Low-Resource LLM-Based Machine Translation Recent LLMs demonstrate emergent translation capabilities via In-Context Learning (ICL) in low resource settings (Brown et al., 2020; Lin et al., 2022) without task-specific training. However, for rare languages like Sawndip, performance degrades significantly due to limited pretraining exposure, leading to hallucinations and unstable generation (Ghazvininejad et al., 2023). To address this, recent research emphasizes the integration of external knowledge sources. Strategies include incorporating dictionary entries and parallel sentence examples (Ghazvininejad et al., 2023; Zhang et al., 2024; Zheng et al., 2024), grammar references (Tanzer et al., 2023; Hus and Anastasopoulos, 2024; Zhang et al., 2025), and morphological analyzers (Zhang et al., 2024) during fine-tuning or inference.Pei et al. (2025) further show that lexical resources often provide stronger gains than explicit grammatical supervision. Consistent with these findings, we propose a knowledge-enhanced SFT approach for Sawndip-to-Chinese translation. Building upon this, we further extend our framework to support Chinese-to-Sawndip translation and bilingual dialogue modeling, enabling a unified system for both translation and interactive generation under extreme low-resource conditions.

3 Method

In this section, we present a unified framework for low-resource Sawndip-Chinese bidirectional translation and bilingual dialogue modeling. As shown in Figure 1, the proposed system integrates dataset construction, knowledge-enhanced translation, synthetic data augmentation, and dialogue synthesis into a single pipeline. In the following, we describe each component in detail, emphasizing how they jointly form a unified framework that couples translation and dialogue modeling to address the data scarcity and structural complexity of Sawndip.

3.1 Dataset Curation

We present a machine-learnable Sawndip–Chinese dataset, comprising both sentence-level parallel

translations and character-level definitions. The core data is derived from a lexical database developed and continuously curated over the past two decades by domain experts at collaborating institutes, supplemented with online resources and historical materials.

To transform these diverse sources into a machine-learnable dataset, we conducted a multi-stage curation pipeline over several months:

- **Data Acquisition:** We extracted all accessible Sawndip-Chinese sentence pairs and character-level lexical entries from the source database. Since many Sawndip logograms are not yet covered by the Unicode standard, we adopt a specialized Sawndip TTF font provided by our collaborators as the unified encoding standard throughout the pipeline.

- **Preprocessing and Normalization:** We conducted a systematic refinement process to ensure cross-source coherence, including mapping variant codes for the same glyph to a single standard, converting raw dictionary entries into a standardized key-value format, and removing duplicate or corrupted entries.

- **Expert-in-Loop Validation:** After automated cleaning, linguistics specialists performed iterative manual reviews. In each iteration, approximately 5% of the corpus was randomly sampled for expert validation. When inconsistencies were identified, programmatic corrections were applied globally to the entire dataset. This iterative process continued until the corpus reached a stable quality level.

The final dataset consists of **5,805 sentence pairs and 11,573 character-level definition entries**. Detailed statistics are summarized in Table 1, highlighting the extreme sparsity, long-tail character distribution, and semantic ambiguity of Sawndip.

Category	Statistic	Value
Corpus	Sentence Pairs	5,805
	Avg. Sawndip Length	4.17
	Avg. Chinese Length	5.65
	Type-Token Ratio	0.249
Characters	Unique Characters	6,008
	Single-occurrence Characters	3,122
	Rare Characters ($f \leq 5$)	5,215
Lexicon	Lexical Entries	11,573
	Avg. Definitions per Character	2.65
	Characters with ≥ 3 Definitions	3,748
	Characters with ≥ 5 Definitions	851

Table 1: Statistics of the original Sawndip-Chinese resources.

3.2 Knowledge-Enhanced Sawndip-to-Chinese Translation

We adopt a knowledge-enhanced SFT paradigm tailored for extremely low-resource scenarios. While recent works utilize dictionary-guided inference-time prompting for LLM-based translation (Ghazvininejad et al., 2023) or investigate the synergy between domain-specific fine-tuning and dictionary injection (Zheng et al., 2024), we shift the focus to the **fundamental character-level lexical unit**, where we explicitly incorporate character-level lexical knowledge into the fine-tuning phase. This is particularly critical for Sawndip, where the model's pre-trained parametric knowledge of its unique logographic system is nearly non-existent.

To implement this, we formalize the Sawndip-to-Chinese (S2C) translation task into a unified instruction-response format. Character definitions are retrieved through exact string matching against the lexical entries in our dataset. Following the format of Ghazvininejad et al. (2023), each training instance structurally consists of **(i)** a translation instruction with the source Sawndip sentence, **(ii)** retrieved character-level lexical definitions, **(iii)** a prompt guiding the model to generate the translation, and **(iv)** the target Chinese sentence. During inference, we maintain the same structure while removing the target answer portion. Providing definitions in a structured and explicit manner enables the model to access lexical information and resolve ambiguities within a highly contextualized translation objective. For controlled comparison, we also consider a dictionary-free variant where the model is trained and tested solely on sentence-level parallel data, serving as a rigorous ablation baseline.

3.3 LLM-Guided Substructure Substitution for Chinese-to-Sawndip Translation

While the knowledge-enhanced approach substantially improves S2C translation, establishing the reverse Chinese-to-Sawndip (C2S) translation capability presents a **distinct set of challenges**: **(i)** extremely limited parallel data that is insufficient for the model to generate target-side morphosyntax; **(ii)** the absence of a curated reverse dictionary; and **(iii)** the severe vocabulary long-tailness where the unique unrepeated logograms are exceptionally high despite the small total sentence count.

To overcome these limitations, we introduce an LLM-guided substructure substitution framework tailored for C2S data augmentation. Inspired by Shi et al. (2021), we reformulate our augmentation as a localized, grammar-preserving LLM-guided replacement mechanism within manually curated sentence templates.

The pipeline of our proposed framework operates as follows:

- **Manual Template Curation:** We manually curate 25 representative semantic templates derived from authentic sentences in our dataset, as exemplified in Appendix A. These master templates are carefully engineered to span a wide variety of scenarios and domains, thereby establishing a reliable grammatical framework that accommodates rich lexical variations.

- **Generative Lexical Expansion:** Within each template, specific semantic slots are designated for replacement. We prompt a high-capacity LLM to perform open-ended vocabulary expansion, generating diverse, contextually appropriate Chinese word variants for each slot. This maximizes vocabulary diversity while strictly preserving the overall sentence fluency and grammatical correctness compared with fully free-form generation.

- **Dictionary-Constrained Alignment:** Generated Chinese phrases are mapped back to Sawndip via an inverted variant of the character-level definitions simplified from our dataset. Crucially, this inverted version consists solely of isolated, decontextualized lexical pairs and lacks target-side syntactic rules or particles. Consequently, it cannot be directly integrated into the translation architecture as in Section 3.2, as rich Chinese synonyms and contextual paraphrases inevitably trigger matching omissions or structural misalignments during decoding. Instead, it acts as a post-expansion filter: only Chinese phrases with completely unambiguous mappings are retained to construct the synthetic parallel pairs to prevent semantic drift.

Through this controlled pipeline, we successfully synthesize **7,491 high-quality, syntactically fluent parallel sentence pairs** for C2S training. These synthetic samples are combined with the original corpus to jointly fine-tune a unified Bidirectional Sawndip-Chinese Translation Model.

3.4 Dimension-Conditioned Dialogue Synthesis and Multi-Candidate Back-Translation Filtering

While the bidirectional translation model demonstrates competent cross-lingual alignment, extending it into a functional conversational model requires high-quality bilingual chat data. However, generic Chinese dialogue corpora are heavily saturated with contemporary slang and Internet expressions, which are poorly aligned with the historical and cultural context associated with Sawndip.

A naive solution is to prompt a high-capacity LLM to generate historical-style dialogues. However, our preliminary experiments reveal that such unconstrained generation suffers from **repetitive phrasing and limited lexical diversity**. To address this issue, we introduce a dimension-conditioned dialogue synthesis strategy. Specifically, we first prompt the LLM to generate a rich pool of specific values across three daily-life dimensions: *(i) Topic* (e.g., farming, trading), *(ii) Action* (e.g., arguing, inviting) and *(iii) Location* (e.g., marketplace, farmland). One element from each dimension is randomly sampled and combined into a structured generation constraint, forcing the LLM to synthesize dialogues under concrete and unique scenarios. Examples are provided in Appendix B.

To project the synthesized Chinese dialogues into Sawndip while maintaining semantic equivalence, we implement the multi-candidate back translation filtering mechanism. For each generated sentence, we utilize our fine-tuned translation model to generate $k = 8$ Sawndip candidates, which are subsequently back-translated into Chinese. Each reconstructed candidate is evaluated against the original Chinese seed using BLEU-2. For each generation pool, only the top-scoring candidate surpassing a predefined threshold ($\tau = 60$) is retained, and is then verified by a high-capacity LLM to ensure absolute cross-lingual consistency.

Through this joint pipeline, we ultimately curate a bilingual conversational corpus comprising **11,629 highly aligned dialogue turns.** To arm the model with bidirectional conversational capability, the verified parallel pairs are compiled into two single-turn formats: Chinese-Query to Sawndip-Response, and vice-versa. The resulting dialogue corpus, along with all translation data, is used to fine-tune our final Bidirectional Sawndip-Chinese Dialogue Model, bridging the gap between raw text translation and authentic conversational interaction.

4 Experiments

4.1 Experimental Setup

Baselines and Tasks We evaluate our framework under two settings: **Machine Translation (MT)** and **Bilingual Dialogue (BD)**. Due to the extreme low-resource nature of Sawndip, general-purpose LLMs exhibit very limited capability in generating valid Sawndip text, and no publicly available domain-specific baselines currently exist. Therefore, baseline comparisons are primarily conducted on the S2C MT task, while the C2S MT and BD settings are used to evaluate the effectiveness of our proposed framework and the bilingual conversational capability of the resulting model. We adopt Qwen3-8B (Yang et al., 2025) as the backbone model for all specialized variants. For the S2C MT task, we compare against DeepSeek-v4-Flash (DeepSeek AI, 2026), DeepSeek-v4-Pro, GPT-5.5 (Singh et al., 2025) and Qwen-3.6-27B (Qwen Team, 2026) under a zero-shot prompting setting.

Datasets We organize our localized resources into two task-specific configurations:

- **Machine Translation:** The MT benchmark is built from the authentic corpus introduced in Section 3.1, consisting of 5,805 Sawndip-Chinese sentence pairs. We split the corpus into 5,205 pairs for training and 300 pairs each for validation and testing. To improve C2S generation, we further augment the training set with 7,491 synthesized pairs developed via the method described in Section 3.3. These synthetic samples are used solely during training and are excluded from validation and testing to guarantee unbiased benchmarking.
- **Bilingual Dialogue:** The BD benchmark is constructed from the synthesized dialogue corpus described in Section 3.4, containing 11,629 single-turn dialogue pairs. For balanced bidirectional evaluation, we reserve 500 pairs for validation and 1,000 pairs for testing, both evenly divided between the C2S and S2C directions. For the final BD training stage, all MT training resources are combined with the dialogue training split to jointly optimize a robust and multi-task Sawndip-specific model.

Implementation Details We fine-tune Qwen3-8B using the Swift (Zhao et al., 2024) framework with full-parameter SFT. Detailed implementation settings are provided in Appendix E.

Evaluation Evaluating Sawndip generation remains particularly challenging due to the lack of standardized evaluation resources and the limited Sawndip ability of existing LLMs. Therefore, for all tasks involving Sawndip outputs, we employ a fixed external back-translation model to translate the generated Sawndip text back into Chinese before downstream evaluation, enabling both automatic metrics and LLM-based evaluation to operate in a stable and reproducible semantic space. Specifically, DeepSeek-v4-Flash is prompted under the same zero-shot knowledge-enhanced translation setting used in our S2C task, where retrieved lexical definitions are provided during inference. We intentionally avoid using our own fine-tuned translation model in order to reduce self-evaluation bias and ensure consistent comparison across all systems, thereby providing a relatively fair semantic comparison in the absence of standardized Sawndip evaluation benchmarks.

For the MT task, we evaluate translation quality using BLEU-2 (Papineni et al., 2002) and chrF++ (Popović, 2017). Details of automatic metrics are provided in Appendix E. We further conduct LLM-as-a-Judge evaluation using DeepSeek-v4-Flash, where the judge model compares the generated or reconstructed Chinese sentence against the original reference under a fixed evaluation prompt and assigns a quality score from 1 to 5.

For the BD task, overlap-based metrics are less reliable due to the open-ended nature of conversation generation. Therefore, we employ DeepSeek-v4-Flash as a dialogue judge to evaluate responses from three perspectives: *(i) logical coherence*, evaluating whether the response is contextually appropriate to the query; *(ii) interaction quality*, evaluating the informativeness and engagement level of the response compared with the reference answer; and *(iii) factual faithfulness*, evaluating whether the response contains meaningful content without hallucinated, repetitive, or empty generation.

4.2 Machine Translation Results

Table 2 presents the machine translation results on both the S2C and C2S tasks. Vanilla SFT denotes the configuration where dictionary knowledge is absent during both training and inference, while all other configurations incorporate retrieved dictionary knowledge during training, inference or both.

Effects of Knowledge-Enhanced Training For the S2C task, knowledge-enhanced SFT (KE-SFT)

Model	S2C			C2S		
	BLEU-2 ↑	chrF++ ↑	LLM-Judge ↑	BLEU-2 ↑	chrF++ ↑	LLM-Judge ↑
DeepSeek-v4-Flash	42.21	21.71	3.19	–	–	–
DeepSeek-v4-Pro	40.25	21.29	3.20	–	–	–
GPT-5.5	43.54	**23.26**	3.33	–	–	–
Qwen-3.6-27B	40.48	21.00	3.15	–	–	–
Qwen3-8B	27.60	13.28	2.42	–	–	–
Vanilla SFT	19.50	8.84	1.06	20.37	8.61	1.34
Knowledge-Enhanced SFT	**44.53**	22.61	**3.49**	20.65	8.72	1.37
+ Synthetic Data	43.21	22.26	3.25	**28.62**	**12.03**	**1.88**

Table 2: Machine translation results on the Sawndip-Chinese benchmark. C2S outputs are back-translated using DeepSeek-v4-Flash before evaluation. General-purpose LLM baselines are not reported for C2S due to their inability to consistently generate valid Sawndip outputs under the zero-shot setting.

substantially outperforms the Vanilla SFT baseline across all evaluation metrics, with BLEU-2 improving dramatically from 19.50 to 44.53 and the average LLM-Judge score increasing from 1.06 to 3.49. These results indicate that sentence-level parallel supervision alone is insufficient for extremely low-resource logographic scripts such as Sawndip. By explicitly incorporating character-level lexical definitions, the model can better integrate lexical semantics with contextual information and reduces translation ambiguity.

Despite being trained on substantially larger multilingual corpora, high-capability LLMs including DeepSeek-v4-Flash, DeepSeek-v4-Pro, GPT-5.5, and Qwen-3.6-27B fail to consistently outperform the proposed KE-SFT framework on the S2C task. Our KE-SFT model achieves the highest BLEU-2 and LLM-judge score among all evaluated models.

Notably, the original Qwen3-8B backbone exhibits very limited Sawndip understanding capability under the same dictionary-augmented prompting format during inference, suggesting that inference-time lexical retrieval alone is also insufficient for robust Sawndip translation in the extremely low-resource setting. These results highlight the importance of combining retrival-grounded lexical augmentation with domain-specific supervised adaption.

Effectiveness of Synthetic Augmentation for C2S Generation For the C2S task, Vanilla SFT and KE-SFT exhibit similarly weak performance, indicating that the limited volume of authentic parallel data is insufficient for stable target-side generation. This is expected because reverse generation requires the model to actively synthesize valid Sawndip sequences under severe vocabulary sparsity and long-tail character distributions.

Metric	Value
Seen-in-Train Character Rate	98.20%
Lexicon Coverage Rate	97.92%
Full Sentence Lexicon Coverage Rate	93.33%

Table 3: Target-side lexical validity statistics for generated C2S outputs on the evaluation set.

After introducing our synthetic augmentation framework, the model achieves significant improvements across all C2S metrics with only a negligible trade-off on S2C. BLEU-2 increases from 20.65 to 28.62, and the LLM-judge score rises from 1.37 to 1.88. These results show that the proposed LLM-guided substructure substitution framework effectively provides the contextual diversity necessary to improve the controllability and stability of Sawndip generation in the C2S direction.

Target-side Lexical Validity Analysis To further validate the structural plausibility of generated Sawndip outputs beyond back-translation based evaluation, we analyze the lexical validity of generated C2S outputs on the 300-sentence evaluation set. As shown in Table 3, most of the generated character appear in the training corpus and lexical dictionary. This suggest that the model rarely produces hallucinated glyphs and largely remains within the observed Sawndip lexical space.

4.3 Bilingual Dialogue Results

Table 4 and Figure 2 present the performance of the final dialogue model on MT and BD benchmarks.

Preserved Translation Capability in Joint Fine-tuning A major concern of adding conversational data under extremely low-resource settings is that the model may overfit to dialogue-style generation and consequently degrade its original translation capability. However, the result in Table 4 shows

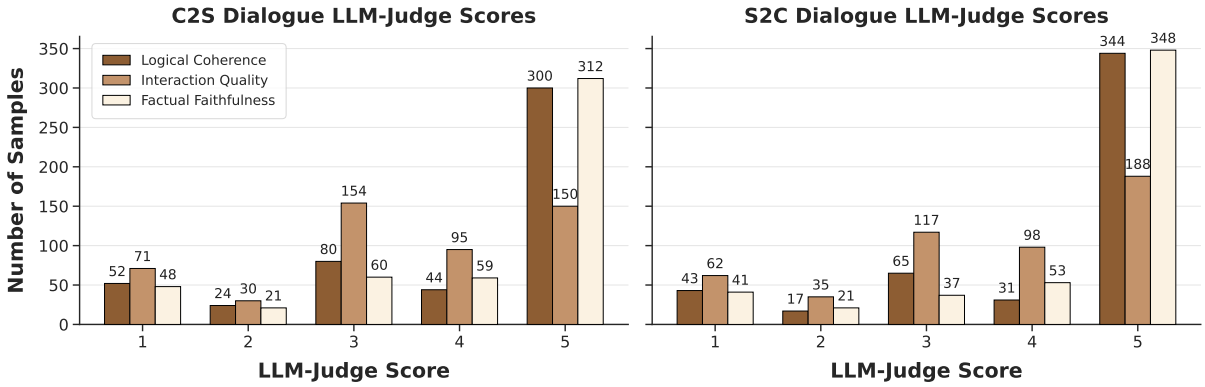

Figure 2: Distribution of LLM-judge scores on the bilingual dialogue benchmark for the S2C and C2S directions across three evaluation dimensions: logical coherence, interaction quality, and factual faithfulness.

	Metric	KE-SFT + Syn.	Dialogue Model
S2C	BLEU-2	43.21	**44.81**
	chrF++	22.26	**22.87**
	LLM-Judge	3.25	**3.44**
C2S	BLEU-2	**28.62**	28.22
	chrF++	12.03	**12.32**
	LLM-Judge	1.88	**2.10**

Table 4: Comparison of the KE-SFT + synthetic data and the final dialogue model on the MT benchmark.

that the final dialogue model exhibits no major degradation on the MT benchmarks.

Notably, on the more challenging C2S MT task, the dialogue model further improves the LLM-Judge score from 1.88 to 2.10. This suggests that conversational fine-tuning further enhances the semantic adequacy and fluency of generated Sawndip responses, even without explicitly optimizing for translation quality during the dialogue training stage. These results indicate that the proposed bilingual dialogue synthesis framework activates conversational capability while preserving the model's original translation performance.

Dialogue Quality Analysis Figure 2 illustrates the score distribution of the DeepSeek-v4-Flash dialogue evaluation for both the S2C and C2S directions. Most generated responses receive scores of 4 or 5 in logical coherence and factual faithfulness, indicating that the proposed framework successfully enables semantically consistent bilingual understanding and dialogue generation capability under the extremely low-resource setting.

Compared with other dimensions, interaction quality exhibits a broader distribution with more score-3 responses. This is partly because the reference responses are generated by a strong LLM and therefore contain richer conversational content, and partly due to the inherent difficulty of generating engaging bilingual dialogues in an extremely low-resource Sawndip setting. In addition, the C2S direction remains slightly weaker than S2C, suggesting that target-side Sawndip generation still leaves room for further improvement. We believe that collecting larger-scale authentic corpora, scaling higher-quality conversational resources and improving target-side generative supervision remain important directions for future work.

5 Conclusion

In this paper, we present a unified computational framework for bidirectional machine translation and bilingual dialogue modeling of Sawndip. To mitigate the severe scarcity of digitized resources, we curated a machine-learnable parallel corpus and a character-level lexical database under expert validation. Addressing the unique lexical ambiguities and non-Unicode orthographic noise of Sawndip, we introduced a KE-SFT paradigm for S2C translation. We further introduce an LLM-guided substructure substitution strategy to support reverse data augmentation based on manually curated templates. Furthermore, we advance the framework to interactive bilingual communication via a dimension-conditioned dialogue synthesis and multi-candidate back-translation filtering pipeline. Experiments demonstrate that our specialized models substantially outperform strong, general-purpose commercial LLM baselines in S2C translation quality, while also exhibiting solid capability in interactive conversations. Ultimately, this work serves as a foundation for the digital recovery of Sawndip and advocates for the synergistic integration of data collection and language modeling to safeguard global linguistic heritage.

Limitations

Although our framework demonstrates promising results for Sawndip–Chinese translation and bilingual dialogue generation, several limitations remain.

First, due to the scarcity of authentic Sawndip resources, a substantial portion of the reverse translation and dialogue data is inevitable synthetically constructed. While this helps strengthen model training under a low-resource setting, it may still introduce distributional bias and limit linguistic diversity.

Second, the evaluation of Sawndip generation remains challenging due to the lack of large-scale standardized benchmarks or native-speaker evaluation protocols. Our C2S evaluation therefore partially relies on back-translation-based assessment, which provides an indirect, though imperfect, yet useful proxy for fluency and structural validity.

Finally, the current dialogue setting focuses mainly on single-turn interactions under constrained daily-life scenarios, and does not yet capture more complex multi-turn conversational dynamics.

We hope future work can benefit from expanded authentic corpora, the incorporation of expert human evaluation, and further exploration of more comprehensive preservation-oriented modeling for endangered historical scripts.

References

Anthropic. 2026. Introducing claude 4. https://www.anthropic.com/news/claude-4.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, and 1 others. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yong Cheng. 2019. Joint training for pivot-based neural machine translation. In *Joint training for neural machine translation*, pages 41–54. Springer.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia conference on language, information and computation*, pages 282–286.

DeepSeek AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro.

Anthony Diller, Jerry Edmondson, and Yongxian Luo. 2004. *The Tai-Kadai Languages*. Routledge.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. *arXiv preprint arXiv:1808.09006*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

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*.

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.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-english languages. *arXiv preprint arXiv:1909.09524*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Proceedings of the 29th Symposium on Operating Systems Principles*.

Thang Le and Luu Anh Tuan. 2023. A parallel corpus for vietnamese central-northern dialect text transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13839–13855.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and 1 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9019–9052.

Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina Von Der Wense. 2024. Proceedings of the 4th workshop on natural language processing for indigenous languages of the americas (americasnlp 2024). In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

OpenAI. 2025. *GPT-5 System Card*. OpenAI. Accessed: 2026-01-04.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schütze. 2025. Understanding in-context machine translation for low-resource languages: A case study on manchu. *arXiv preprint arXiv:2502.11862*.

Begoña Pendas, Andrés Carvallo, and Carlos Aspillaga. 2023. Neural machine translation through active learning on low-resource languages: The case of spanish to mapudungun. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 6–11.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on machine translation: Research papers*, pages 186–191.

Qwen Team. 2026. Qwen3.6-27B: Flagship-level coding in a 27b dense model.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. 2021. A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys (CSUR)*, 54(4):1–34.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 86–96.

Freda Shi, Karen Livescu, and Kevin Gimpel. 2021. Substructure substitution: Structured data augmentation for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3494–3508.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book. *arXiv preprint arXiv:2309.16575*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chen Zhang, Jiuheng Lin, Xiao Liu, Zekai Zhang, and Yansong Feng. 2025. Read it in two steps: Translating extremely low-resource languages with code-augmented grammar books. *arXiv preprint arXiv:2506.01796*.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. 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.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. Swift:a scalable lightweight infrastructure for fine-tuning. *Preprint*, arXiv:2408.05517.

Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *arXiv preprint arXiv:2402.15061*.

Yuxi Zheng and Jingsong Yu. 2025. Incorporating lexicon-aligned prompting in large language model for tangut–chinese translation. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 127–136.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

A Cases of Manually Curated Substructure Substitution Template

Original Sawndip:
剼芓庲廸杯
Original Chinese:
割青草来喂牛
English Translation for Original Chinese:
Cut grass to feed the cows
Sawndip Template:
剼[植物]庲廸[动物]
Chinese Template:
割[植物]来喂[动物]
English Translation for Template:
Cut [kind of plant] to feed the [kind of animal]

Original Sawndip:
批歪堆造苒
Original Chinese:
去坡上做屋
English Translation for Original Chinese:
Go up the slope to build a house
Sawndip Template:
批[方位词][地点/地形][建造动作][建筑]
Chinese Template:
去[地点/地形][方位词][建造动作][建筑]
English Translation for Template:
Go [direction] the [location] to [action of building] a [building]

Original Sawndip:
歐奉毕膨碟
Original Chinese:
拿木棍去打果子
English Translation for Original Chinese:
Use a stick to knock down the fruit
Sawndip Template:
歐[工具]毕[动作][果实/目标]
Chinese Template:
拿[工具]去[动作][果实/目标]
English Translation for Template:
Use a [tool] to [action of hitting] [fruit/target]

B Examples for Dimension-Conditioned Dialogue Generation

Topic: 饮食(food) 农作(agriculture) 天气(weather) 健康(health) 买卖(trade) 节庆(festival) 教育(education) ...

Action: 请求(request) 陈述(statement) 抱怨(complaint) 询问(inquiry) 叮嘱(reminder) 赞美(praise) 安慰(comfort) 拒绝(refusal) ...

Location: 厨房(kitchen) 卧室(bedroom) 田地(field) 菜园(vegetable garden) 山上(mountain) 竹林(bamboo forest) 河边(riverside) ...

Examples:

饮食-请求-厨房(food-request-kitchen): -锅里有什么吃的？-还有一碗稀饭。(-What's in the pot to eat? -There's only a bowl of porridge.)

农作-陈述-卧室(agriculture-statement-bedroom): -床底下的麻袋里有黄豆。-拿出来晒晒吧。(-There are soybeans in the sack under the bed. -Take them out and dry them in the sun.)

天气-抱怨-田地(weather-complaint-field): -天老不下雨，苗都黄了。-再不来雨，今年就白忙活了。(-It hasn't rained for so long, the crops are turning yellow. -If it still doesn't rain, this year will be wasted.)

.....

C Ablation Study: Synthetic Data Scaling for C2S Generation

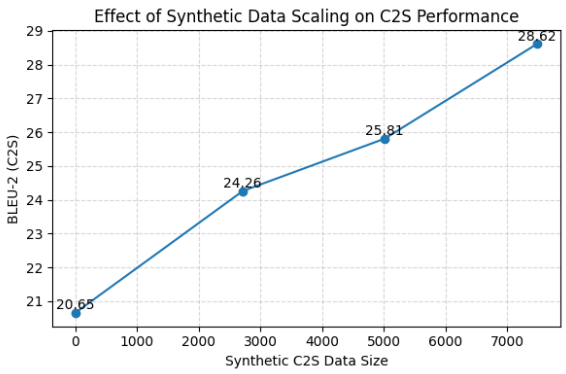

Figure 3: Effect of synthetic data scale on C2S translation performance. The x-axis represents the number of synthetic Chinese-to-Sawndip training pairs, and the y-axis reports BLEU-2 scores on the C2S test set. The results show a consistent improvement with increasing data scale.

As shown in Figure 3, we observe that BLEU-2 scores consistently increase as the amount of synthetic training data grows, indicating that scaling synthetic data is effective for improving C2S translation performance.

D Examples of LLM-as-a-Judge Back-Translation Evaluation

Original Sentence
→Selected Sawndip for back-translation
→Back-Translated Sentence
→LLM Judge(pass or fail)

还有一碗稀饭。(There is still a bowl of porridge.)
→里眉碗餂澈。
→还剩一碗稀饭。 (There is still one bowl of porridge left.)
→ T

舀点水给我喝。(Scoop some water for me to drink.)
→愕淰擀兄呻。
→舀水给我喝。 (Scoop water for me to drink.)
→ T

成，你要几个？ (OK, how many do you want?)
→貧，佫歐几憁？
→成，你拿几个？ (OK, how many do you take?)
→ T

每家都能领一份。(Each household can receive one copy.)
→峦峦糺口份。
→家家都能做一份。(Each household can produce one copy.)
→ F

水槽里有水吗？ (Is there water in the sink?)
→唔潧眉淰咪？
→潭里有水吗？ (Is there water in the pond?)
→ F

渡船码头有卖鱼汤的吗？ (Is fish soup sold at the ferry pier?)
→朔瞙眉艚糎？
→码头有卖米汤的吗？ (Is rice porridge sold at the pier?)
→ F

E Implementation and Evaluation Details

We fine-tune the Qwen3-8B model using the Swift framework with full-parameter SFT on 8 NVIDIA A100 (40GB) GPUs. The model is optimized using the AdamW optimizer with a learning rate of 1×10^{-5}, a cosine learning rate scheduler, a weight decay of 0.1 and a warmup ratio of 0.03 for a maximum of 4 epochs. The global batch size is set to 32 and the maximum sequence length is set to 1024 tokens. To prevent overfitting, we employ an early stopping strategy based on validation loss.

We adopt the vLLM (Kwon et al., 2023) framework during inference. To ensure reproducibility, all decoding processes are performed with temperature set to 0 and a fixed random seed.

BLEU-2 and chrF++ are computed using SacreBLEU (Post, 2018). For BLEU-2 evaluation, we use SacreBLEU with the internal zh tokenization, and employ the default exponential smoothing. We adopt BLEU-2 instead of the commonly used BLEU-4 because the Sawndip-Chinese corpus mainly contains relatively short sentences where higher n-gram overlaps are extremely sparse.

We use the original Qwen3-8B tokenizer without any modification. The tokenizer follows the standard Qwen byte-level BPE scheme. For Sawndip codepoints that are not included in the tokenizer vocabulary, they are decomposed and encoded through the tokenizer's byte-level fallback mechanism, ensuring that all unseen characters are represented as sequences of bytes.

F System Prompt for Qwen3-8B Translation SFT

你是一名“古文字和古语智能助手”，专长于古壮字与汉语之间的双向翻译任务。

为保证输出稳定可评估，请严格遵守以下规则：

1. 语言与字符：根据指令要求确定输出语言。若源语言为古壮字，则结果使用简体中文作答；若源语言为简体中文，则输出对应的古壮字文本。除翻译所必需的转换外，不得对原文中的字符形态进行改写、替换或规范化。

2. 输出结构：不要寒暄或复述题目，仅输出最终翻译答案文本，不得包含解释、注解、背景说明或任何非翻译内容。

3. 参考说明：若输入中附带了字、词或句的释义、对照或提示信息，这些内容可以作为翻译时的参考依据，在翻译时应结合上下文进行判断，并给出你认为最合理的翻译结果。

G System Prompt for Qwen3-8B Dialogue SFT

你是一名“古文字和古语智能助手”，专长于古壮字与汉语之间的双向翻译和双语对话任务。

为保证系统输出的稳定性和评估的高效性，请严格遵守以下规则：

1. 任务模式判定与应答逻辑：

- 【翻译模式】：若输入包含翻译指令（如“翻译”、“译作”等），系统仅输出最终的目标语言翻译文本，不得有任何寒暄、题目复述、解释、注解、背景说明或非翻译内容。

- 【对话模式】：若输入为日常交谈、问候或场景对话，系统应扮演地道的对话伙伴进行自然应答。允许输出符合情境的完整句子，但依然不进行技术性的自我解释。

2. 语言与字符交互规则：

- 当需要翻译的文本或用户输入为古壮字时，系统必须使用简体中文回答。

- 当需要翻译的文本或用户输入为简体中文时，系统必须使用古壮字回答。

- 无论处于哪种模式，除翻译所必需的转换外，不得对原文中的字符形态进行改写、替换或规范化。

3. 参考说明：

- 若输入中附带了字、词或句的释义、对照或提示信息，这些内容可以作为翻译时的参考依据，在翻译时应结合上下文进行判断，并给出你认为最合理的翻译结果。

H System Prompt for DeepSeek-v4-Flash Chinese Judge

你是一名翻译评测专家。

你的任务是：

对“标准答案”和“翻译结果”的语义一致性进行打分。

评分标准：

- 5分：语义完全一致，翻译自然准确，没有遗漏错误

- 4分：基本正确，存在轻微措辞或细节偏差

- 3分：大体能表达原意，但存在明显遗漏/错误

- 2分：仅部分相关，大量信息缺失或错误

- 1分：基本无法对应原意

- 0分：完全错误、无关、拒答、空白

要求：

1. 只输出一个0-5 的整数

2. 不要输出解释

3. 不要输出其他内容

I System Prompt for DeepSeek-v4-Flash Dialogue Judge

你是一名严谨的“多轮对话评测专家”。你的任务是根据提供的一组中文对话交互数据，在纯语义和对话逻辑层面对大模型的回复质量进行多维度综合裁决。

你会得到三句话，其中Query 是人类的提问。Response 是模型的回答。Reference 是可供参考的人类回答。

请从以下三个核心维度，根据提供的Query ，对模型的Response 进行评分（每个维度满分5 分）：

1. 语境连贯性(Contextual Coherence, 1-5分):

- 评估目标：只评价Response 对Query 的逻辑承接度。这个评分和Reference无关。

- 评分标准：

* 5分：Response 极其自然地承接了Query 的话题，回答完全契合历史文化/日常交际语境，合情合理。

* 3分：Response 与Query 大体相关，能勉强聊下去，但存在轻微的答非所问或生硬的语义转折。

* 1分：完全答非所问，逻辑彻底断层，对话完全崩溃。

2. 标准参照下的语义对齐与意图合理度(Reference-Anchored Semantic & Intent Alignment, 1-5分):

- 评估目标：参考Reference的语义与行为边界，评估模型回复在逻辑上的“合理性”，以及在信息量和交互深度上是否与标准回复具有同等的水准。

- 评分标准：

* 5分：【逻辑与水准完美对齐】模型的回复在逻辑上极其合理。无论它选择了与标准回复相同还是相反的立场（如标准回复是拒绝，模型是接受），其体现出的对前文Query 的理解深度、上下文契合度、以及言语行

为的得体程度，都达到或超越了Reference的水准。

* 3分：【立场合理但深度不足】模型的回复核心意图大体说得通，但在语义深度或信息密度上明显弱于Reference（例如：标准回复给出了具体的拒绝理由，而模型只回了一句简单的“我不想去”），或者语气与古人历史文化背景有轻微的违和感。

* 1分：【彻底偏离标准约束】模型的回复在逻辑上完全无法对齐。其交互意图产生了严重的社会学或常识性坍塌（例如：标准回复是在解答问题，模型回复却在无理质问；或者标准回复表现出恭敬，模型回复表现出极度轻浮），完全丧失了标准回复所具备的跨语言交互水准。

3. 信息密度与反幻觉度(Informativeness & Anti-Hallucination, 1-5分):

- 评估目标：Response 的语言质量，是否存在低级的废话、复读机现象、车轱辘话或无意义的幻觉编造。

- 评分标准：

* 5分：结构精炼干脆，信息密度高。

* 3分：包含有效核心信息，但伴随有轻微的语义重复、复读机倾向，或者吐出了少量不影响大局的废话。

* 1分：产生严重的语义坍塌，疯狂复读某一词汇，或者通篇全是不着边际的幻觉信息与口水废话。

[输出格式]

为了方便上游Python 脚本精准解析，请严格按照以下JSON 格式输出你的裁决结果，不要包含任何前导词（如"Here is the result:"）、Markdown 格式标记（如“‘json）或任何额外的解释：{ "coherence_score": 0.0, "intent_score": 0.0, "informativeness_score": 0.0, }