NUOSUBENCH: REVEALING THE LIMITS OF MULTILINGUAL LARGE LANGUAGE MODELS IN STANDARDIZED YI

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual large language models (LLMs) exhibit strong cross-lingual transfer across many languages, but it remains unclear whether this capability extends to languages written in severely underrepresented scripts. We investigate this question through Standardized Yi (Nuosu), a language spoken by millions of people in southwestern China but largely absent from existing LLM training and evaluation ecosystems. We introduce **Nuosu-Datasets** and **Nuosu-Benchmark**, the first large-scale resource suite for training and systematically evaluating Standardized Yi language models. Nuosu-Datasets contains 78.6M tokens for continual pre-training and 166M tokens comprising 1.1M instruction instances for supervised fine-tuning and cross-lingual alignment. The data are collected from authoritative web and expert-provided sources through OCR-based digitization, multi-stage filtering, and expert validation. Nuosu-Benchmark contains 110,513 instances across five sub-benchmarks that evaluate lexical knowledge, sentence-level understanding, culturally grounded knowledge, administrative language, and educational comprehension using 12 complementary metrics. Evaluating 14 representative LLMs reveals a substantial gap between general multilingual capability and competence in Standardized Yi. All evaluated general-purpose models achieve aggregate scores below 40%, indicating that model scale and broad multilingual training alone do not ensure effective transfer to an underrepresented script. In contrast, BIMO-8B, an 8B model adapted using Nuosu-Datasets, achieves 50.1%, outperforming the strongest evaluated general-purpose model by 10.2 percentage points. Further analyses show that language-specific training produces improvements across multiple levels of linguistic competence, while culturally grounded and knowledge-intensive tasks remain particularly challenging. Our resources and findings establish a reproducible foundation for studying cross-lingual transfer, data-efficient adaptation, and evaluation in extremely low-resource language modeling.[1]

1 INTRODUCTION

Large language models (LLMs) have opened new frontiers for language technologies, but their benefits remain unevenly distributed (Costa-Jussà et al., 2022). Most progress has primarily benefited a small group of high-resource languages, leaving the majority of the world's 7,000+ languages underrepresented in current LLM ecosystems (Costa-Jussà et al., 2022; Zhou et al., 2024). This disparity is largely driven by a resource gap: effective LLM adaptation and evaluation require language-specific training resources and reliable benchmarks, whose development is especially valuable for low-resource languages (Chang et al., 2024; Huang et al., 2023; Muennighoff et al., 2022). The gap is especially severe for culturally rich low-resource languages, where language-specific resources can capture distinctive structures and cultural knowledge (Duncan et al., 2024).

Standardized Yi-script, also known as Nuosu, exemplifies this challenge. It is one of the world's six major ancient writing systems and a carrier of culturally specific knowledge across literature, education, governance, and oral traditions. Although standardized into 819 standard phonograms and digitally encoded after the 1980 *Yi-script Standardization Scheme* (Qian, 1990; Kong, 2004), it

[1] Resources:`https://anonymous.4open.science/r/NuosuBench-3D83`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

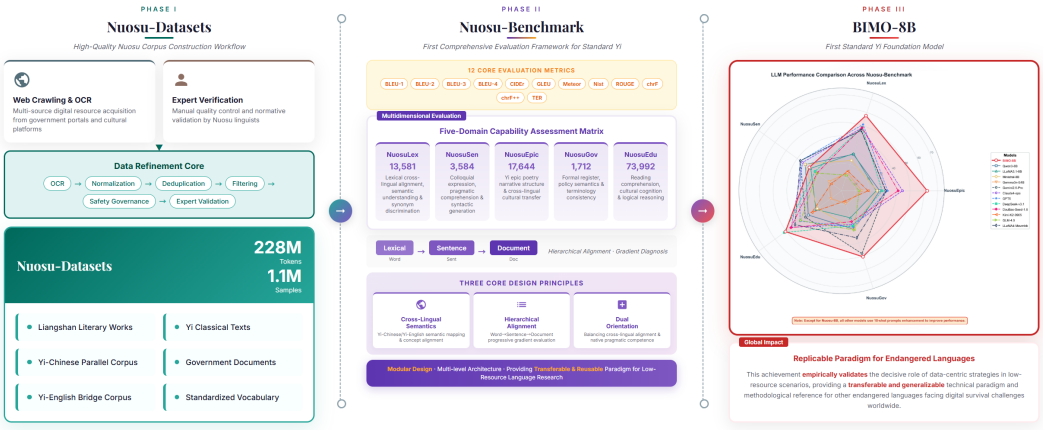


Figure 1: Overview. *Left*: Nuosu-Datasets, yielding approximately 244.6M tokens across monolingual and cross-lingual alignment instruction data. *Center*: Nuosu-Benchmark, covering five sub-benchmarks with 12 complementary metrics. *Right*: BIMO-8B, trained on Nuosu-Datasets, outperforms evaluated general-purpose LLMs on Yi-script tasks.

remains largely absent from the current LLM ecosystem. Existing work mainly focuses on OCR, handwriting recognition, and document-level visual processing (Sun et al., 2024; Shang et al., 2024; Sun et al., 2025; Liu et al., 2024b; Chen et al., 2022; Bi et al., 2025; Xiao et al., 2024), providing a foundation for training resources and benchmarks for LLM-based semantic understanding, reasoning, and generation.

We address this gap by constructing, to our knowledge, the first large-scale training-and-evaluation resource suite for standardized Yi-script LLM research. Our goal is not only to collect more data, but also to provide language-specific training resources and a multi-level benchmark that can diagnose model capabilities across lexical, syntactic, culturally grounded, administrative, and comprehensive evaluation settings. Our contributions are:

- **Nuosu-Datasets**: We construct a quality-centric training resource for standardized Yi-script LLM development from both authoritative web sources and expert-provided materials, including Yi–Chinese–English aligned texts and scanned Yi documents. The pipeline integrates OCR-based digitization, text preparation, and alignment filtering, yielding 78.6M tokens for continual pretraining and 166M tokens of cross-lingual alignment instruction data for supervised fine-tuning.
- **Nuosu-Benchmark**: We introduce a script-aware, multi-level diagnostic benchmark with 110,513 instances across five sub-benchmarks. It evaluates lexical alignment, sentence-level semantic integration, culturally grounded interpretation, domain-specific register understanding, and comprehensive reasoning using 12 complementary metrics under a unified 10-shot protocol.
- **Comprehensive Evaluation**: We evaluate 14 LLMs and characterize distinct lexical, sentence-level, and culturally grounded performance profiles. BIMO-8B trained on Nuosu-Datasets achieves 50.11 points, outperforming the strongest evaluated general-purpose model by 10.24 points.

The datasets, benchmarks, and accompanying code support reproducible research on Standard Yi-script and related low-resource languages.

2 Related Work

Standard Yi-script. Standard Yi-script is the official writing system of the Yi people in southwestern China and an important carrier of Yi cultural heritage. The 1980 *Yi-script Standardization Scheme* consolidated thousands of regional variants into a set of 819 standard phonograms, enabling digital encoding and Unicode support (Kong, 2004; Huang, 2009; Qian, 1990). Existing technical work has mainly focused on visual-language processing, including OCR, handwriting recognition, document detection, and segmentation (Liu et al., 2024b; Bi et al., 2025; Xiao et al., 2024; Sun et al., 2024; Shang et al., 2024; Sun et al., 2025; Chen et al., 2022). Large-scale text resources and standardized

evaluation protocols extend this foundation to semantic understanding, reasoning, and generation in Standard Yi-script.

Evaluation benchmarks for low-resource languages. As LLMs improve, standard benchmarks have begun to saturate. MMLU-Pro (Wang et al., 2024) and LiveBench (White et al., 2024) address this through expanded reasoning tasks and contamination-resistant evaluation design, but they remain primarily oriented toward high-resource languages. Recent benchmarks have begun to target minority and low-resource languages: MiLiC-Eval (Zhang et al., 2025) covers Tibetan, Uyghur, Kazakh, and Mongolian in a cross-lingual parallel framework, while CMHG (Xu et al., 2025) focuses on headline generation for Tibetan, Uyghur, and Mongolian. These works demonstrate the importance of language-specific evaluation. We extend this direction to Standard Yi-script by introducing NUOSU-BENCHMARK, covering five sub-benchmarks and 12 complementary metrics.

Modeling for low-resource languages. Recent low-resource language development follows two main paradigms. The *data-driven* paradigm, exemplified by NLLB (Costa-Jussà et al., 2022), builds multilingual corpora and evaluation resources at scale. The *model-transfer* paradigm adapts strong open-source LLMs to target languages through continual pretraining and instruction tuning (Zhang et al., 2023; 2024; Lu et al., 2024; Ji et al., 2024). TiLamb (TiLamb Team, 2024) provides a relevant example for Tibetan by combining a strong backbone with language-specific corpora and domain-adaptive continual pretraining. These studies show that low-resource language adaptation depends on both target-language training data and reliable evaluation standards. These complementary goals motivate our construction of Nuosu-Datasets and Nuosu-Benchmark for Standard Yi-script.

3 NUOSU-DATASETS

3.1 DATA COLLECTION AND QUALITY CONTROL

Collecting high-quality Yi-script corpora is essential for advancing LLM development in this under-resourced language. Authoritative published materials provide the basis for our training corpus and benchmark construction.

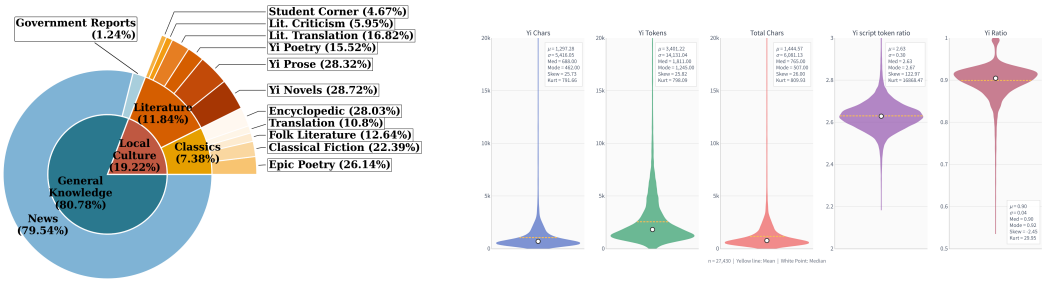
Our data sources consist of two categories. (1) **Web-sourced resources**: we extract Yi text from authoritative platforms, including *Liangshan Daily* and *People's Daily*, alongside Yi poetry, prose, and articles. (2) **Expert-provided resources**: domain experts provide existing Yi–Chinese–English aligned texts and Yi-related scanned PDF materials, including textbooks, biographies and classical novels, etc. For scanned materials, we fine-tune PaddleOCRv5 (Cui et al., 2025) to build Nuosu-OCR (97% accuracy on synthetic text-line images; Appendix E) for extracting texts from scanned materials. Experts contribute source materials. Overall, these authoritative sources cover diverse domains and preserve local Yi culture, ethnic knowledge, and textual traditions, providing a solid data foundation.

CPT uses HTML text extraction, formatting standardization, and article-level hash deduplication (Appendix C.2). Aligned resources additionally undergo: (1) **Normalization** unifies Unicode encoding and removes format-related noise; (2) **Deduplication** removes repeated aligned pairs using MD5 hashing; (3) **Filtering** removes short, language-mixed, low Yi-script-ratio, or misaligned samples; (4) **Safety governance** removes sensitive information and non-releasable content.

After collection and quality control, we obtain 27,430 Yi-script monolingual samples, totaling 78.6M tokens, and 218,522 Yi–Chinese and Yi–English parallel samples, totaling 4.2M tokens. The monolingual corpus is used for CPT, while the alignment data are converted into instruction-format data for SFT and benchmark construction.

3.2 MONOLINGUAL CORPUS FOR CPT

The cleaned monolingual Yi-script corpus is used for CPT to develop linguistic grounding by adapting the base model's representations to Yi-script lexical, syntactic, and discourse patterns (Zheng et al., 2024b). Collected from the Yi editions of *Liangshan Daily* and *People's Daily*, plus Yi poetry, prose, and articles, the corpus spans diverse genres, including news (79.54%), literature (11.84%), Yi classical texts (7.37%), and government reports (1.24%), covering both general-domain knowledge



(a) **CPT corpus distribution.** (b) **CPT corpus statistics.**

Figure 2: CPT corpus composition and distributional analysis. *(a)* Distribution of the 78.6M-token CPT corpus across news (79.54%), literature (11.84%), Yi classical texts (7.37%), and government reports (1.24%), covering both general-domain and Yi local cultural knowledge. *(b)* Statistics of 27,430 monolingual samples, showing long-tailed length distributions, tokenizer characteristics, and Yi-script proportion.

and Yi cultural traditions (Figure 2a). CPT follows the collected genre distribution; Appendix C.1 gives the subject breakdown within news.

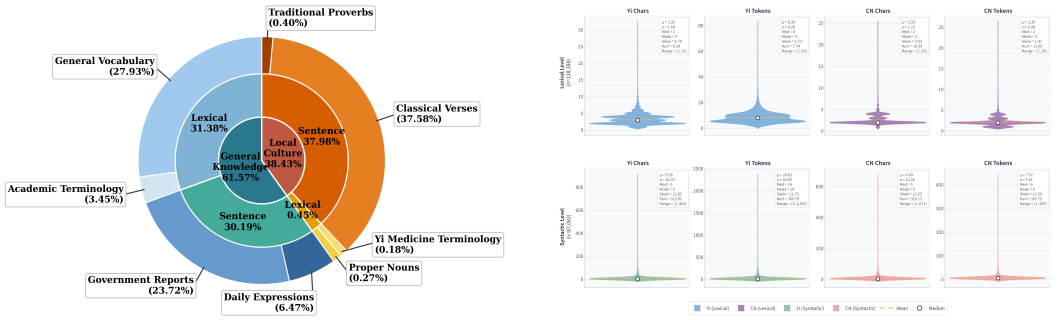
Furthermore, Figure 2b shows distributional statistics of the corpus. First, the length-related measures exhibit clear right-skewed and heavy-tailed distributions, consistent with the long-tailed nature of natural language corpora (Yu et al., 2018), reflecting the heterogeneity of Yi-script texts from short news items to long classical materials. Second, the token-to-character ratio of 2.63 ± 0.30 describes the Qwen3 tokenizer's token expansion on Yi-script. Additionally, the average Yi-script proportion of 0.90 describes the script composition of the corpus. These properties demonstrate the value of the Yi-script monolingual corpus for LLM development by providing reliable target-language exposure for CPT-based adaptation, and laying the foundation for subsequent instruction tuning and cross-lingual task learning.

3.3 Cross-Lingual Alignment Corpus for SFT

The cleaned cross-lingual alignment data support SFT by adapting the Yi-script representations learned during CPT to instruction-following and cross-lingual tasks (Gururangan et al., 2020). Collected from expert-provided resources and reliable web sources, including *SWUN*, *Yi Homeland*, *Yi Studies Network*, and *People's Daily (Yi Edition)*, the data consist of 128,388 Yi–Chinese vocabulary pairs, 87,083 Yi–Chinese sentence pairs, and 3,051 Yi–English sentence pairs. These resources cover vocabulary- and sentence-level alignments across both general-domain and Yi local cultural knowledge, supporting lexical mapping, translation, and culturally grounded semantic learning (Figure 3b).

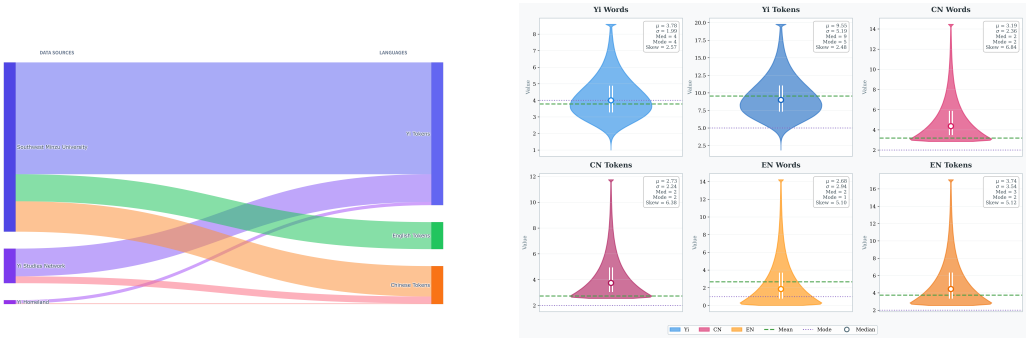
Figure 3b shows right-skewed distributions in the Yi–Chinese parallel data at both vocabulary and sentence levels, with Yi sequences about 3–4$\times$ longer than Chinese under the Qwen3 tokenizer. Due to the scarcity of Yi–English parallel data, we further use Chinese as a pivot language for alignment augmentation. Specifically, we translate Chinese annotations into English using OPUS-MT (Tiedemann et al., 2024), adding 53,663 Yi–Chinese–English samples (47,407 vocabulary-level and 6,256 sentence-level samples). This yields 200,674 additional English tokens and expands the alignment corpus to 175,795 vocabulary pairs and 96,390 sentence pairs, forming richer Yi–Chinese–English alignment resources for broader cross-lingual knowledge transfer (Figure 4a). Figure 4b also shows clear token-scale differences across Yi-script, Chinese, and English. These tokenization asymmetries characterize the alignment data used to connect Yi-script with higher-resource language representations during SFT.

To make the alignment resources usable for instruction tuning, we propose a Task-Oriented Instruction Expansion Algorithm (TIEA; Appendix D) to convert raw parallel data into Alpaca-format (Taori et al., 2023) instruction samples. Instead of relying on parallel mappings alone, TIEA expands the alignment corpus into diverse task forms, including translation, question answering (QA), multiple-choice, error correction, etc. This transformation increases data scale and task-format diversity,



(a) **CPT content-category distribution.** (b) **Yi–Chinese parallel corpus statistics.**

Figure 3: Corpus content and cross-lingual alignment analysis. *(a)* Token shares in the CPT data by content category. *(b)* Statistical distributions of Yi–Chinese parallel data at vocabulary and sentence levels, showing cross-lingual token length asymmetries.



(a) **Yi–Chinese–English token distribution.** (b) **Yi–Chinese–English token statistics.**

Figure 4: Yi–Chinese–English trilingual corpus analysis. *(a)* Token distribution of the 53,663 trilingual alignment samples and token proportions across Yi-script (68.7%), Chinese (18.3%), and English (13.0%) under the Qwen3 tokenizer. *(b)* Word-length and token distributions across Yi-script, Chinese, and English, showing cross-lingual differences in length distribution and tokenization efficiency.

providing task-oriented learning signals across vocabulary, syntax, semantics, and pragmatics for cross-lingual task scenarios.

Before task expansion, we split the *post-augmentation* alignment inventory at the source-unit level: 175,795 vocabulary-level and 96,390 sentence-level pairs obtained after pivot-based English expansion of 218,522 raw parallel records (§3.3). Specifically, 154,104 vocabulary-level and 73,450 sentence-level pairs are allocated to SFT construction, with the remaining units assigned to disjoint held-out benchmark pools. All derivatives follow their source unit's allocation; Appendix J reconciles these pools. Training-side TIEA expands the SFT construction pool into 1,110,064 Yi instruction samples and approximately 166M tokens. On the benchmark side, TIEA expands only NuosuEdu's separate 8,110 vocabulary units into 73,992 items; NuosuLex, NuosuSen, NuosuEpic, and NuosuGov retain original gold references from their respective held-out source pools.

In summary, **Nuosu-Datasets** consists of two complementary components: a 78.6M-token monolingual CPT corpus and a 166M-token cross-lingual SFT corpus. The CPT component supports target-language representation learning, while the SFT component supports instruction following, cross-lingual alignment, and task generalization. Together, these two components provide a foundational training resource for developing and adapting LLMs to Yi-script.

4 NUOSU-BENCHMARK

4.1 DESIGN PRINCIPLES

Evaluating LLMs on Yi-script requires a benchmark that reflects both its low-resource status and its multilingual usage environment. For low-resource languages, cross-lingual alignment is foundational to digital development and knowledge transmission, as it connects the target language with high-resource languages and makes its linguistic and cultural content more accessible. Therefore, Nuosu-Benchmark adopts a cross-lingual alignment evaluation paradigm across Yi–Chinese/English contexts.

Guided by this goal, Nuosu-Benchmark follows a multi-level diagnostic design. It evaluates five complementary capability levels: word-level alignment, sentence-level understanding, cultural discourse interpretation, government-register comprehension, and comprehensive task reasoning. This structure avoids over-reliance on a single translation task and enables skill-wise evaluation from surface lexical mapping to deeper cultural-semantic and task-oriented understanding.

4.2 SUB-BENCHMARKS

Following the cross-lingual alignment evaluation paradigm, Nuosu-Benchmark is organized into five sub-benchmarks so that evaluation is *factorized* rather than collapsed into a single mixed score. Table 1 aligns each subset with a diagnostic *level*, its primary *task* families, and the *capabilities* it is meant to stress; the discussion below follows that layout. The motivation for five partitions is pragmatic and linguistic: cross-lingual competence in a low-resource setting is not one homogeneous skill, accurate word mapping can coexist with weak sentence integration, fluent modern prose with poor epic or administrative register control, and strong local translation with brittle multi-format reasoning. Separating lexical, compositional, culturally grounded, institutionally marked, and integrative layers yields complementary diagnostic profiles across separate held-out source pools.

All sub-benchmarks use mutually disjoint held-out source pools. NuosuLex–NuosuGov retain original gold answers from 13,581 vocabulary and 22,940 sentence units. Benchmark-side TIEA is applied exclusively to NuosuEdu's separate 8,110 vocabulary units, yielding 73,992 items (Appendix J). **NuosuLex** instantiates the table's *lexical/semantics* row: translation and disambiguation isolate the smallest alignment units, so performance here reflects terminology consistency and sense disambiguation without requiring long-range discourse. **NuosuSen** corresponds to *syntax/semantics*, combining translation with oral-expression prompts to test compositional grammar and pragmatics in modern usage, a layer orthogonal to bare lexicon because failures often arise from agreement, word order, and situational appropriateness rather than unknown lemmas. **NuosuEpic** occupies the *pragmatic* epic register (Yi canonical texts such as *Mamu Teyi* and *Lewo Teyi*): poetic translation and exegesis probe culturally anchored interpretation that modern sentences and word lists do not substitute, since allusion, formulaic diction, and ethnoliterary context drive meaning. **NuosuGov** is also *pragmatic* but targets a non-literary axis, official translation and comprehension on government-report style language where abstract policy wording, fixed administrative collocations, and formal tone define a separate competence from both everyday syntax and epic discourse. **NuosuEdu** is the *comprehensive* tier in Table 1: it expands its separate held-out vocabulary source pool into multiple task families (MCQ, judgment, QA, reorder) to approximate examination-style integration of lexicon, grammar, and reasoning; Appendix D details its TIEA construction. Its format diversity complements the original-reference evaluation on the first four subsets.

4.3 EVALUATION METRICS AND PROTOCOL

We construct a multi-dimensional metric suite for Nuosu-Benchmark. Yi-script's tokenization characteristics motivate complementary lexical and character-based measurements. We combine 12 metrics across word-level, character-level, semantic, and edit-distance dimensions to assess reference agreement through lexical, character, alignment, and edit-based views. This metric design supports reference-based Yi-script cross-lingual evaluation. Detailed metric definitions are provided in Appendix F.

All 14 models, including BIMO-8B, use the same standardized 10-shot in-context learning protocol, with two labeled examples randomly sampled from each sub-benchmark. Following prior work on few-shot prompting (Brown et al., 2020; Wei et al., 2022; Hendrycks et al., 2020), this setup

Table 1: Nuosu-Benchmark composition.

Sub-bench.	Samples	Level	Tasks	Capabilities
NuosuLex	13,581	Lexical/Semantics	Translation, disambiguation	Lexical mapping & alignment
NuosuSen	3,584	Syntax/Semantics	Translation, oral expression	Grammar, pragmatics
NuosuEpic	17,644	Pragmatic	Poetic translation, exegesis	Cross-lingual, cultural context
NuosuGov	1,712	Pragmatic	Official translation, comprehension	Abstract semantics, terminology
NuosuEdu	73,992	Comprehensive	MCQ, judgment, QA, reorder	Linguistic & logical reasoning

supplies a common demonstration structure for comparison under low-resource conditions. The prompt template is provided in Appendix H.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Training BIMO-8B We evaluate the utility of Nuosu-Datasets by training BIMO-8B, a domain-adapted 8B model initialized from Qwen3-8B (Yang et al., 2025). Training is implemented with LLaMAFactory (Zheng et al., 2024c) in two stages. We first conduct continual pre-training on monolingual CPT corpus for 5 epochs with an 8,192-token context window, a learning rate of 2×10^{-4}, and a batch size of 64. We then perform full-parameter supervised fine-tuning on 1,110,064 Yi instruction samples plus 222,012 general-purpose examples (1,332,076 in total) for 1 epoch with a 1,280-token context window, corresponding to the 99.9th percentile of Yi SFT sample lengths, a learning rate of 6×10^{-5}, and a micro-batch size of 32. Both stages use cosine learning-rate decay with 10% linear warmup and are conducted on 4× NVIDIA A800-80G GPUs. CPT follows the natural corpus distribution, and replay is introduced only during SFT (Appendix K).

Evaluated models We evaluate 14 representative LLMs covering three groups: (1) *same-scale open-source LLMs*, including Qwen3-8B, Ministral-8B, LLaMA 3.1-8B (Dubey et al., 2024), and Gemma3n-E4B (Gemma Team et al., 2025); (2) *advanced general-purpose LLMs*, including Gemini 2.5-Pro (Comanici et al., 2025), Claude 4 Opus, LLaMA 4-Maverick, GPT-5, DouBao-Seed-1.6, DeepSeek-v3.1 (Liu et al., 2024a), GLM-4.5 (Zeng et al., 2025), Kimi-K2-0905 (Kimi Team et al., 2025), and Qwen3-235B (Yang et al., 2025); and (3) *our domain-adapted model*, BIMO-8B. All models are evaluated under the same Nuosu-Benchmark protocol. General capability results are reported in Appendix I.

5.2 MAIN RESULTS

Figure 5 shows model performance on Yi-script tasks. BIMO-8B achieves the highest average score of 50.11, outperforming the second-ranked Claude 4 Opus (39.87) by 10.24 points. Advanced general-purpose models, including GPT-5, Gemini 2.5-Pro, and DeepSeek-v3.1, cluster within a narrow 36–40-point range, while same-scale open-source models such as Qwen3-8B and Ministral-8B generally fall below the overall mean of 31.95. Relative to Qwen3-8B (29.58), BIMO-8B gains 20.53 points, demonstrating the utility of language-specific adaptation with an 8B backbone. Excluding NuosuEdu, BIMO-8B ranks first across the four non-TIEA subsets with a mean score of 49.99, followed by Gemini 2.5-Pro (41.09) and Claude 4 Opus (38.50) (Table 8), demonstrating that its leading performance extends to evaluation with original gold references.

The error bars further reveal substantial variation in model performance across sub-benchmarks. Models exhibit task-specific deviations (e.g., BIMO-8B ±11.5, DeepSeek-v3.1 ±11.7), indicating uneven strengths across linguistic levels, whereas models (e.g., Qwen3-8B ±1.8, Gemma3n-E4B

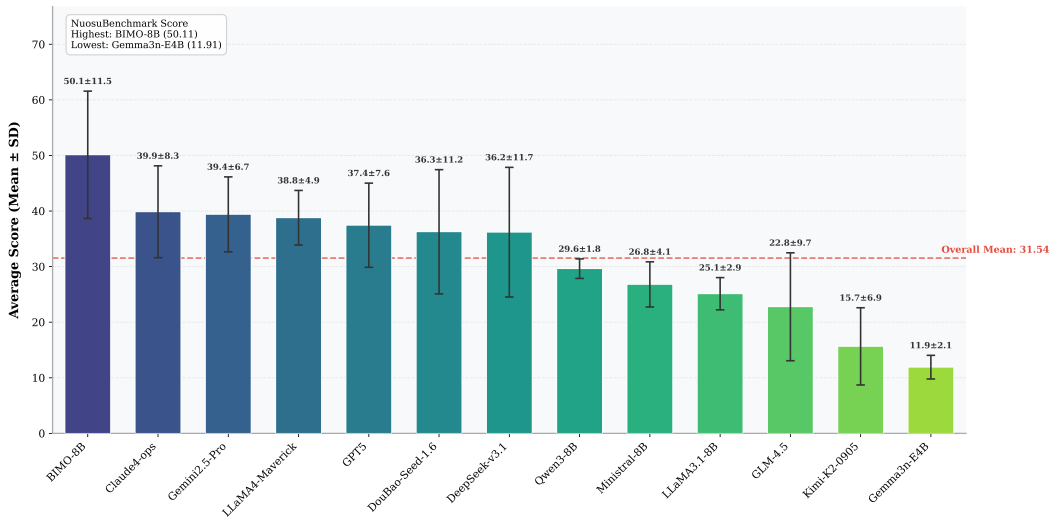


Figure 5: LLM performance on Nuosu-Benchmark. Average scores and population standard deviations across five sub-benchmarks for 14 models. The dashed line indicates the overall mean score across all evaluated models. Error bars denote cross-sub-benchmark variation.

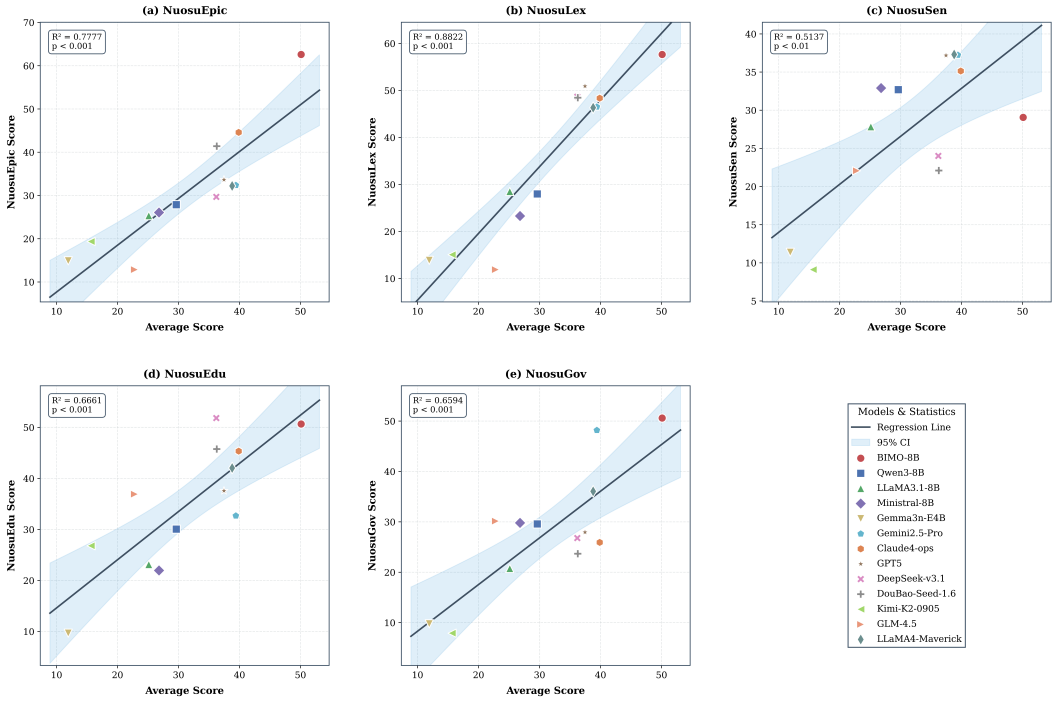


Figure 6: Relationship between sub-benchmark scores and overall performance. Each subplot shows a linear regression between model scores on one sub-benchmark and the overall Nuosu-Benchmark score across 14 evaluated models. Each panel reports R^2 as a descriptive association; the overall score includes the displayed subset.

$\pm$2.1) have more closely clustered subset scores. LLaMA 4-Maverick shows a relatively balanced profile among general-purpose LLMs (38.8$\pm$4.9).

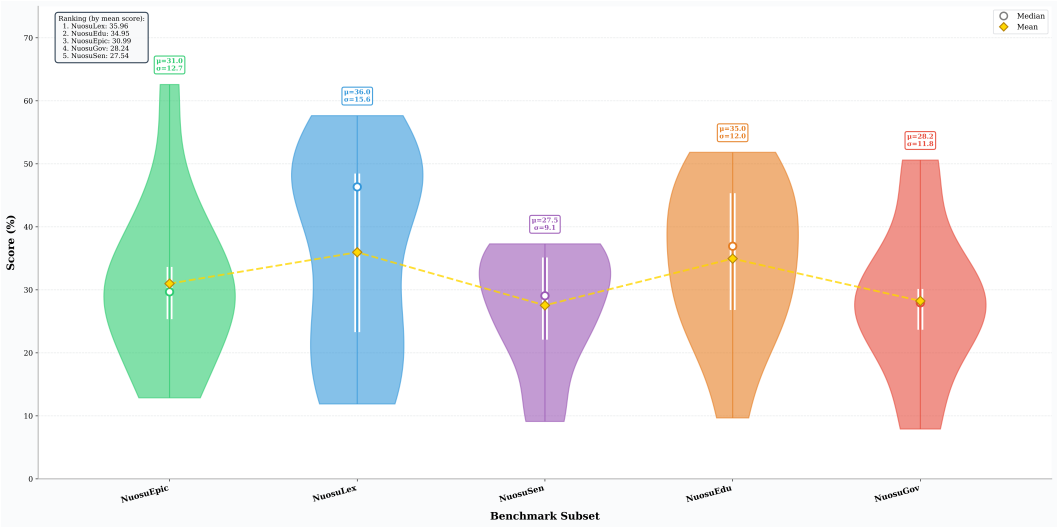


Figure 7: Score distributions across sub-benchmarks. Violin plots show the distribution of model scores on each sub-benchmark. Internal box plots mark quartiles; white dots indicate medians and yellow diamonds indicate means. The dashed line traces the mean trend across sub-benchmarks.

5.3 Sub-Benchmark Analysis

Figure 6 summarizes linear relationships between each sub-benchmark and overall Nuosu-Benchmark scores. Descriptive association is highest for NuosuLex ($R^2 = 0.8849$) and lowest for NuosuSen ($R^2 = 0.5226$); the overall score includes the displayed subset. Composite scores separate models into distinct bands: BIMO-8B near ∼50 points; closed-source models (Claude 4 Opus, GPT-5, Gemini 2.5-Pro) at 36–40; open-source models such as Qwen3-8B at 25–35; and several general-purpose models below 20. Residual patterns add model-specific detail: DeepSeek-v3.1 scores above its regression prediction on NuosuEdu (51.83 vs. 39.63), GLM-4.5 scores above the fitted lines on NuosuGov and NuosuEdu, and BIMO-8B scores 29.06 on NuosuSen.

Figure 7 reports mean scores and dispersion by sub-benchmark. Means fall into three tiers: NuosuLex (μ=36.70) and NuosuEdu (μ=35.58); NuosuEpic (μ=31.13); and NuosuSen (μ=27.91) with NuosuGov (μ=28.45), spanning 8.79 points from the highest to the lowest mean. Population standard deviations are largest on NuosuLex (σ=15.28; the observed score range exceeds 40 points) and smallest on NuosuSen (σ=8.85). BIMO-8B leads on NuosuEpic by 18.00 points and scores 29.06 on NuosuSen; DeepSeek-v3.1 peaks on NuosuEdu (51.83) relative to its 36.19 overall mean.

The lexical-level score spread and concentrated sentence-level NuosuSen distribution make sentence-level alignment a concrete target for further Yi-script LLM development; we expand interpretive discussion in Appendix B.

6 Conclusion

We present Nuosu-Datasets and Nuosu-Benchmark, the first large-scale training-and-evaluation resource suite for standardized Yi-script LLM research. Nuosu-Datasets provides approximately 244.6M tokens of monolingual and cross-lingual alignment instruction data for continual pre-training and supervised fine-tuning, while Nuosu-Benchmark provides a 110,513-sample evaluation suite with five sub-benchmarks and 12 complementary metrics. Evaluations of 14 LLMs show distinct Yi-script task profiles, with BIMO-8B trained on Nuosu-Datasets achieving the best overall performance. These findings highlight the need for language-specific training resources and multi-level evaluation in low-resource language modeling. These resources support reproducible research on Standard Yi-script and related languages.

REFERENCES

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. COIG-CQIA: Quality is all you need for chinese instruction fine-tuning. https://huggingface.co/datasets/m-a-p/COIG-CQIA, 2024.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Xiao Bi, Zhuo Sun, and Zhigang Chen. A novel unsupervised contrastive learning framework for ancient Yi script character dataset construction. *npj Heritage Science*, 13(1):39, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. When is multilinguality a curse? Language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4074–4096, 2024.

Shanxiong Chen, Yi Yang, Xiaofang Liu, et al. Dual discriminator GAN: Restoring ancient Yi characters. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4): 1–23, 2022.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

Cheng Cui, Ting Sun, Mingkun Lin, et al. PaddleOCR 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pp. 138–145, 2002.

Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. SVTR: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*, 2022.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The LLaMA 3 herd of models. *arXiv e-prints, arXiv:2407.21783*, 2024.

Suzanne Duncan, Gianna Leoni, Lee Steven, Keoni Mahelona, and Peter-Lucas Jones. Fit for our purpose, not yours: Benchmark for a low-resource, indigenous language. *Advances in Neural Information Processing Systems*, 37:31543–31561, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376, 2006.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8342–8360, 2020.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*, 2023.

Jianming Huang. The history and characteristics of Yi script. *Literature and History Knowledge*, (2): 64–67, 2009.

iamtarun. Python code instructions 18k alpaca. Hugging Face dataset, `https://huggingface.co/datasets/iamtarun/python_code_instructions_18k_alpaca`, 2023.

Shaoxiong Ji, Zihao Li, Jussi Paavola, et al. EMMA-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*, 2024.

Kimi Team, Yuntao Bai, Yejie Bao, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Xiangqing Kong. Prospects of Yi script standardization. *Journal of Minzu University of China (Philosophy and Social Sciences Edition)*, 31(4):126–131, 2004.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Aixin Liu, Bei Feng, Bin Xue, et al. DeepSeek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Xiaofang Liu, Xu Han, Shanxiong Chen, et al. Ancient Yi script handwriting sample repository. *Scientific Data*, 11(1):1183, 2024b.

Yinquan Lu, Wenhao Zhu, Lei Li, Lei Yu, Lingui Qian, Rui Wang, and Fei Huang. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*, 2024.

M-A-P. CodeFeedback-Filtered-Instruction. Hugging Face dataset, `https://huggingface.co/datasets/m-a-p/CodeFeedback-Filtered-Instruction`, 2024.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

Open R1. OpenR1-Math-220k. Hugging Face dataset, `https://huggingface.co/datasets/open-r1/OpenR1-Math-220k`, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Maja Popović. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, 2015.

Yuzhi Qian. A standardized encoding scheme for the Yi script. *Chinese Information*, (3):40–41, 1990.

Wei Shang, Ling Ling Chen, and Wei Xiang. Yi printed character recognition based on deep learning. *Procedia Computer Science*, 242:584–591, 2024.

shibing624. Alpaca-zh. Hugging Face dataset, `https://huggingface.co/datasets/shibing624/alpaca-zh`, 2023.

Hao Sun, Xiaoqing Ding, Jun Sun, et al. A method for detecting and recognizing Yi character based on deep learning. *Computers, Materials & Continua*, 78(2), 2024.

Hao Sun, Xiaoqing Ding, Zheng Li, et al. Linguistic-visual based multimodal Yi character recognition. *Scientific Reports*, 15(1):11874, 2025.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*, 2023.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, et al. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, 58(2):713–755, 2024.

TiLamb Team. TiLamb: A Tibetan large language model based on incremental pre-training. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pp. 254–267, 2024.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2014.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Colin White, Samuel Dooley, Manley Roberts, et al. LiveBench: A challenging, contamination-limited LLM benchmark. *arXiv preprint arXiv:2406.19314*, 2024.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Wei Xiao, Shanxiong Chen, Yan Ma, et al. Integrating path signature and pen-tip trajectory features for online handwriting Yi text recognition. *Heritage Science*, 12(1), 2024.

Gang Xu, Zhaoyi Su, Zihan Zhang, et al. CMHG: A dataset and benchmark for headline generation of minority languages in China. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12350–12357, 2025.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shuntaro Yu, Chunshan Xu, and Haitao Liu. Zipf's law in 50 languages: Its structural pattern, linguistic interpretation, and cognitive motivation. *arXiv preprint arXiv:1807.01855*, 2018.

Aohan Zeng, Xiao Lv, Qinkai Zheng, et al. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Chaoqun Zhang, Ming Tao, Zijian Liao, et al. MiLiC-Eval: Benchmarking multilingual LLMs for China's minority languages. *arXiv preprint arXiv:2503.01150*, 2025.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yue Yan, and Yang Feng. BayLing: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*, 2023.

Shaolei Zhang, Kehao Zhang, Qingkai Fang, et al. BayLing 2: A multilingual large language model with efficient language alignment. *arXiv preprint arXiv:2411.16300*, 2024.

Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. OpenCodeInterpreter: Integrating code generation with execution and refinement. arXiv preprint arXiv:2402.14658, 2024a.

Wenzhen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, and Ming Zhou. Breaking language barriers: Cross-lingual continual pre-training at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7725–7738, 2024b.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024c.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Celia Cerdà-i Cucchietti, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024.

A Research Opportunities

(1) Building on the resource suite's $\sim$244.6M tokens, future corpus development can expand coverage of spoken and dialectal variation; independently authored tasks and native-speaker assessment can further probe cultural appropriateness. (2) BIMO-8B's NuosuSen profile motivates studies of high-quality sentence-level parallel corpora, domain-balanced adaptation, and alternative training strategies for sentence-level alignment. (3) Future studies can examine Yi-script segmentation and compare the contributions of CPT, instruction expansion, pivot translation, and SFT replay through component-wise experiments.

B Qualitative Discussion of Sub-Benchmark Patterns

The regression pattern for NuosuLex versus NuosuSen provides complementary descriptions of lexical and sentence-level performance in this benchmark. The tiered composite scores further illustrate variation across the current LLM ecosystem. Model-specific residuals invite task-focused analysis: DeepSeek-v3.1's lift on NuosuEdu, GLM-4.5's above-line scores on NuosuGov and NuosuEdu, and BIMO-8B's NuosuSen profile identify distinct settings for further study.

The violin distributions show the widest cross-model spread on lexical tasks and a narrower spread at sentence level. BIMO-8B's epic lead alongside its NuosuSen score and DeepSeek-v3.1's NuosuEdu peak motivate further analysis of sentence-level alignment and examination-style task formats. These perspectives motivate dedicated sentence-level alignment strategies for low-resource languages.

C Post-Cleaning Corpus Statistics

Table 2: **Corpus composition and distribution after cleaning.** *Top*: Monolingual CPT corpus (78.6M tokens total, consistent with §3.1), with news comprising 79.54%, followed by literature (11.84%), classical texts (7.37%), and government reports (1.24%); these are genre shares. *Bottom*: Parallel alignment corpus (218,522 samples total: 128,388 Yi–Chinese vocabulary pairs, 87,083 Yi–Chinese sentence pairs, and 3,051 Yi–English sentence pairs), with vocabulary-level pairs accounting for 58.75% and sentence-level pairs (including Yi–English) for 41.25%. Sample percentages are computed over all parallel samples. Token counts computed with the Qwen3 tokenizer.

(a) Monolingual corpus

Split	Genre	Tokens	%	Source
CPT	News	62.5M	79.54	Liangshan Daily; People's Daily (both Yi); Yi literary texts
	Literature	9.3M	11.84	
	Classical texts	5.8M	7.37	
	Government	1.0M	1.24	
Total		**78.6M**	**100**	

(b) Bilingual parallel corpus

Level	Category	Samples	%	Source
Vocabulary	Proper nouns	794	0.36	SWUN
	Technical terms	12,419	5.68	SWUN, YSN
	General vocabulary	115,175	52.71	SWUN, YH
	Subtotal	*128,388*	*58.75*	
Syntactic	Literary classics	73,558	33.66	SWUN
	Official documents	6,611	3.03	People's Daily
	Daily expressions	6,914	3.16	YH, YSN
	Yi–English (sentence)	3,051	1.40	SWUN, YH, YSN
	Subtotal	*90,134*	*41.25*	
Total		**218,522**	**100**	

C.1 Subject Coverage Within News

The eight categories in Table 3 partition the news component and sum to 100% *within news*. News itself accounts for 79.54% of all CPT tokens. This finer description makes the subject coverage of the largest genre directly inspectable. Training follows the collected genre distribution, retaining the corpus's original mixture of published content.

Table 3: Subject composition within the news component of CPT.

News subject	Share within news (%)
Top News	27.63
Science and Technology	16.42
Holiday and Lifestyle	13.87
Politics	11.59
Education	8.74
Economy	8.34
Ecology	7.06
Investigative Reporting	6.35
Total	**100.00**

C.2 CPT Acquisition, Cleaning, and Deduplication

The CPT corpus draws on authoritative published materials, including the Yi-language editions of *Liangshan Daily* and *People's Daily*, together with Yi poetry, prose, and articles. Source-specific crawlers extract structured text from these complementary resources. For *Liangshan Daily*, the crawler traverses issue dates and newspaper pages, extracts article titles and body text through HTML parsing, and retains extracted records of at least ten characters, preserving article context. For selected WeChat collections, article URLs are deduplicated before paragraph extraction. The newspaper preprocessing pipeline retains records containing characters in the Yi syllables block (U+A000–U+A48F), removes Han characters, and standardizes text formatting. Spaces and line breaks are regularized, while non-breaking spaces, byte-order marks, and placeholder boxes are removed to produce consistent plain text. Records retaining at least five characters after cleaning are serialized as UTF-8 JSONL with a `text` field. Article-level hash deduplication consolidates repeated articles in the CPT collection. The final corpus contains 27,430 samples and 78.6M tokens, providing substantial Yi-script exposure while preserving the collected genre distribution during training.

D Data Augmentation Algorithm (TIEA)

The **Task-Oriented Instruction Expansion Algorithm** (TIEA) converts trilingual parallel JSON records into Alpaca-format $\{\texttt{instruction}, \texttt{input}, \texttt{output}\}$ triples (Taori et al., 2023). In the main text, TIEA scales the alignment corpus for supervised fine-tuning (SFT). The *same* pipeline also builds the comprehensive sub-benchmark **NuosuEdu**, inspired by modern Yi-script examinations: it starts from its separate pool of 8,110 held-out vocabulary-level alignment samples and runs TIEA to materialize seven evaluation families, translation, multiple-choice, cloze completion, true/false judgment, error correction, word-order restoration, and daily-conversation-style interaction. The released code factorizes these into six parallel generator routines (daily-conversation items are emitted by the `translation` module using informal template pools, as in §3.3). The resulting NuosuEdu split contains 73,992 samples and tests whether models transfer Yi-script lexical, syntactic, and semantic knowledge across heterogeneous task formats, with reference-based measurements of instruction following.

The reference implementation is in the NuosuBench code release.[2] Each JSON record provides `yi` (standardized Yi-script), `chinese`, optional machine-translated `english`, and optional explanatory strings `content` / `english_content` (used for culturally grounded or definitional extensions). The six routines, `translation`, `choice`, `cloze`, `correct`, `judge`, and `reorder`, each consume the corpus $\mathcal{D}$ and emit lists of triples. A process pool runs them in parallel; outputs are concatenated, globally shuffled, and written to disk. Instruction wording is randomized from hand-crafted template pools (mixed colloquial and formal Chinese/English), which yields diverse surface forms for the same semantic task.

`translation`. For every $(\texttt{yi}, \texttt{chinese})$, the generator emits Chinese→Yi-script and Yi-script→Chinese samples; Yi-script→English and English→Yi-script are also emitted if `english` is present. If `content` is set, two additional samples append Chinese cultural/semantic explanation after the translation in both directions. If `english_content` is also set, two further samples pair English↔Yi-script translations with bilingual explanations. Templates are drawn uniformly at random from fixed lists; outputs may echo the bare target string or wrap it in short explanatory Chinese/English.

`choice`. Items with $|\texttt{yi}| \geq 30$ characters are skipped (to keep options compact). The number of options K is sampled uniformly from $\{4, \ldots, 8\}$. Distractors are $K-1$ distinct wrong records sampled uniformly from $\mathcal{D} \setminus \{\text{current}\}$. Four multiple-choice tasks are emitted when `english` exists: (i) given `yi`, choose the matching Chinese; (ii) given `chinese`, choose the matching Yi-script; (iii) given English, choose the matching Yi-script; (iv) given `yi`, choose the matching English. Options are letter-prefixed (`A.`, `B.`, ...), shuffled, and the gold label is the correct letter.

`cloze`. Items with $|\texttt{yi}| \leq 15$ are skipped. Let $L = |\texttt{yi}|$. The algorithm masks m random distinct character positions, where $m \sim \text{Unif}\{\lceil 0.1L \rceil, \ldots, \lfloor 0.5L \rfloor\}$ (clipped to $[1, L-1]$), replacing selected

[2] `https://anonymous.4open.science/r/NuosuBench-3D83`

glyphs with '_'. The model must output the full original `yi`; the instruction optionally cites the Chinese gloss for disambiguation.

`correct`. With probability 0.8 and only if $|\texttt{yi}| > 15$, a corrupted string is formed by applying one operation sampled from {`replace`, `delete`, `insert`, `swap`} at the character level: *replace* substitutes a random position with a glyph drawn uniformly from the union of all Yi-script appearing in $\mathcal{D}$; *delete* removes one character; *insert* inserts one random corpus glyph; *swap* exchanges two adjacent characters. Otherwise the input is the clean `yi` and the target output is the fixed phrase indicating no error (Chinese). In all cases the reference output is the gold `yi` when corruption was applied.

`judge`. For each record the generator emits several binary judgment instances: correct Yi-script↔Chinese pairs (positive); Yi-script paired with a wrong Chinese taken from another record (negative); correct Chinese↔Yi-script (positive); Chinese paired with a wrong Yi-script (negative). If `english` exists, additional items judge whether `yi` matches `english` versus a wrong English gloss sampled from another English-bearing record. Gold answers are short natural-language strings (e.g., "correct", "incorrect, the translation should be ...").

`reorder`. Items with $|\texttt{yi}| \leq 15$ are skipped. The characters of `yi` are uniformly shuffled; the model must recover the original string, optionally prompted with the Chinese meaning.

Algorithm 1: TIEA / trilingual parallel-to-Alpaca expansion

Input: JSON parallel corpus $\mathcal{D}$ with fields `yi`, `chinese`, `english`, `content`, and `english_content` (optional fields). **Output:** Shuffled Alpaca dataset $\mathcal{A} = \{(\texttt{instruction}, \texttt{input}, \texttt{output})\}$

$\mathcal{A} \leftarrow \emptyset$
parallel for $f \in \{\texttt{translation}, \texttt{choice}, \texttt{cloze}, \texttt{correct}, \texttt{judge}, \texttt{reorder}\}$
$\mathcal{A} \leftarrow \mathcal{A} \cup f(\mathcal{D})$ *// each f implements the rules in the paragraphs above*
SHUFFLE($\mathcal{A}$); **return** $\mathcal{A}$

Module summary:
- `translation`: template-randomized Yi-script↔Chinese↔English translation; informal pools for daily-conversation-style prompts; optional `content` / `english_content` for glosses
- `choice`: $K \in [4, 8]$-way MCQ with cross-example distractors (skip long Yi-script)
- `cloze`: random 10–50% Yi-script character masks (skip short Yi-script)
- `correct`: 80% character-level noise (replace/delete/insert/swap) else "no error"
- `judge`: correct/incorrect translation judgments for Zh/Yi/En pairs
- `reorder`: unshuffle permuted Yi-script (skip short Yi-script)

Figure 8: TIEA control flow for SFT data and NuosuEdu construction.

E OCR Model Details

To extract text from scanned Yi-script PDF documents, we fine-tune PaddleOCRv5 (Cui et al., 2025) using a synthetic data generation approach. Clean text from the web-crawled corpus (§3.1) is segmented into lines of 25 characters and rendered using the NuosuSIL-Regular.ttf font at 28pt on white backgrounds with 32px image height. Gaussian blur is applied with 50% probability for augmentation, producing 2,219,891 training and 235,740 test images.

The model uses SVTR HGNet (Du et al., 2022) (PPHGNetV2_B4 backbone) with a multi-head design integrating CTC (Graves et al., 2006) and NRTR decoders. Multi-scale training (320×32/48/64) handles variable aspect ratios. Training uses Adam (Kingma & Ba, 2014) (β_1=0.9, β_2=0.999), initial LR 1×10^{-4} with cosine annealing, L2 regularization (3×10^{-5}), for 5 epochs on 4× A800-80G GPUs (batch size 128 per device). The resulting Nuosu-OCR achieves 97% accuracy on the held-out synthetic text-line test set.

F Evaluation Metrics

Yi token lengths are 3–4× those of Chinese under the Qwen3 tokenizer, motivating explicit segmentation conventions and complementary character-level metrics. Furthermore, Nuosu-Benchmark spans heterogeneous tasks (literary translation, terminology alignment, educational QA) where no single metric is uniformly appropriate. We therefore adopt 12 complementary metrics across four dimensions (Table 4):

- **Word-level n-gram** (BLEU-1/2/3/4 (Papineni et al., 2002), NIST (Doddington, 2002), ROUGE (Lin, 2004), GLEU (Wu et al., 2016)): Measure lexical overlap and precision; NIST additionally weights rare vocabulary via information-theoretic scoring.
- **Character-level n-gram** (chrF, chrF++ (Popović, 2015)): Measure character-sequence agreement; chrF++ also includes word n-grams.
- **Statistical-semantic** (METEOR (Banerjee & Lavie, 2005), CIDEr (Vedantam et al., 2014)): Measure token alignment and TF-IDF-weighted n-gram similarity.
- **Edit distance** (TER): Quantifies minimum edit operations to transform model output into the reference, reflecting practical post-editing workload.

Table 4: Conceptual evaluation metric profiles across five dimensions. Morphology-related matching depends on the language resources used by the metric.

Metric	Spelling	Morphology	Word Order	Word Import.	Tok.-Free
BLEU-1/2/3/4	Weak	Weak	Medium	No	No
NIST	Weak	Weak	Medium	Yes	No
ROUGE	Weak	Weak	Medium	No	No
GLEU	Weak	Weak	Medium	No	No
CIDEr	Weak	Weak	Medium	Yes	No
METEOR	Weak	Strong	Strong	No	No
chrF	Strong	Strong	Weak	No	Yes
chrF++	Strong	Strong	Weak	No	No
TER	Medium	Medium	Strong	No	No

The five robustness dimensions address specific challenges in Yi evaluation. *Spelling robustness*: Yi phonograms are susceptible to visually or phonetically similar substitutions; character-level metrics (chrF/chrF++) tolerate such deviations without excessive penalization. *Morphological robustness*: METEOR's stem and synonym matching depends on its language-specific resources. *Word-order robustness*: Yi–Chinese word-order differences mean that semantically equivalent translations may differ structurally; TER and METEOR capture such reordering. *Word importance recognition*: Information-weighted metrics (NIST, CIDEr) reward accurate reproduction of low-frequency, high-information content words, advantageous for Yi cultural and technical vocabulary. *Tokenization dependency*: chrF matches at character level, and chrF++ additionally uses word n-grams.

For standardized Yi-script, we adopt character-level segmentation; for Chinese and English, word-level segmentation (Jieba for Chinese). The reported metrics retain their individual scales, including the metric-specific ranges of NIST and CIDEr and the higher-is-better TER similarity $1/(1 + \mathrm{TER})$. The overall score is the arithmetic mean of the five reported subset scores; the original-reference aggregate averages NuosuLex, NuosuSen, NuosuEpic, and NuosuGov. Derived means, population standard deviations, and descriptive regressions use the displayed subset scores.

Formal definitions.

BLEU (Papineni et al., 2002): BLEU $= BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$, where p_n is the modified n-gram precision, $BP = \min(1, e^{1-r/c})$ is the brevity penalty, c and r are candidate and reference lengths, and $w_n = 1/N$.

CIDEr (Vedantam et al., 2014): $\text{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \cdot \frac{1}{m} \sum_{j=1}^{m} \frac{\mathbf{g}^n(c_i) \cdot \mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\| \cdot \|\mathbf{g}^n(s_{ij})\|}$, where $\mathbf{g}^n$ is the TF-IDF weighted n-gram vector.

GLEU (Wu et al., 2016): GLEU $= \min(P_{\text{ng}}, R_{\text{ng}})$, where P_{ng} and R_{ng} are clipped n-gram overlap precision and recall over the combined 1–4-gram counts.

METEOR (Banerjee & Lavie, 2005): METEOR $= F_{\text{mean}} \cdot (1 - \text{Penalty})$, where $F_{\text{mean}} = \frac{10PR}{R+9P}$ with precision $P = m/|c|$ and recall $R = m/|r|$, and Penalty $= 0.5 \cdot (ch/m)^3$.

NIST (Doddington, 2002): $\text{NIST} = \sum_{n=1}^{N} \frac{\sum_{\text{n-gram} \in c} \text{Info(n-gram)} \cdot \delta(\text{n-gram} \in R)}{\sum_{\text{n-gram} \in c} 1} \cdot BP_{\text{NIST}}$, where $\text{Info}(w_1 \cdots w_n) = \log_2 \frac{\text{Count}(w_1 \cdots w_{n-1})}{\text{Count}(w_1 \cdots w_n)}$.

ROUGE (Lin, 2004): ROUGE-N $= \frac{\sum_{S \in \text{Ref}} \sum_{\text{n-gram} \in S} \text{Count}_{\text{match}}(\text{n-gram})}{\sum_{S \in \text{Ref}} \sum_{\text{n-gram} \in S} \text{Count(n-gram)}}$.

chrF / chrF++ (Popović, 2015): $\text{chrF}_\beta = (1 + \beta^2) \cdot \frac{\text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}}$, with character n-gram precision and recall averaged over orders 1 to N (default N=6, β=2). chrF++ extends this with word-level n-grams (default N_w=2).

TER : The minimum number of edit operations (insertions, deletions, substitutions, shifts) to transform the candidate into the reference, normalized by reference length.

G Per-Metric Results

Table 5 reports the mean scores of representative models across all 12 evaluation metrics on Nuosu-Benchmark. BIMO-8B achieves the best performance on the majority of metrics, ranking first in BLEU-1/2/3/4, CIDEr, GLEU, Meteor, ROUGE, chrF-F, and chrF-F++, demonstrating a clear overall advantage in lexical overlap and n-gram precision. DeepSeek-v3.1 and DouBao-Seed-1.6 share the highest NIST score (167.41), providing a complementary information-weighted view of reference agreement. Among the closed-source frontier models, Gemini 2.5-Pro and Claude 4 Opus occupy the second tier, with Gemini 2.5-Pro attaining second place on BLEU-2/3/4, GLEU, and chrF metrics, while Claude 4 Opus ranks second on BLEU-1, CIDEr, Meteor, and ROUGE. Notably, the TER-based metric 1/(1+TER) presents a contrasting pattern: GLM-4.5 achieves the highest score (50.44), followed closely by DouBao-Seed-1.6 (50.30) and DeepSeek (49.49), whereas BIMO-8B scores relatively lower (45.58), suggesting that edit-distance-based evaluation captures different aspects of translation quality than n-gram-based metrics. Gemma3n-E4B records 5.43 on CIDEr and 18.71 on 1/(1+TER), further illustrating the range of metric-specific model profiles.

H Prompt Template

The 10-shot prompt template used for all 14 model evaluations, including BIMO-8B, is shown below. It contains a system instruction establishing the model as a Yi-script expert, followed by 10 examples (2 per sub-benchmark) covering translation, multiple-choice, judgment, cloze completion, and word reordering tasks.

Table 5: Mean scores per metric for selected models on Nuosu-Benchmark. Best results are in **bold**; second-best are underlined. Higher scores are better (↑).

	BLEU↑									chrF↑		
Model	**1**	**2**	**3**	**4**	**CIDEr↑**	**GLEU↑**	**Meteor↑**	**Nist↑**	**ROUGE↑**	**F**	**F++**	**1/(1+TER)↑**
BIMO-8B	**33.51**	**25.66**	**21.21**	**18.21**	**169.00**	**22.80**	**39.80**	147.34	**35.47**	**26.08**	**22.58**	45.58
Claude 4 Opus	25.20	16.70	13.33	11.84	141.05	16.80	24.91	138.20	26.57	13.02	10.96	47.38
GPT-5	20.56	13.83	11.68	10.80	119.58	14.63	19.93	152.44	21.66	14.45	12.35	49.19
Gemini 2.5-Pro	23.50	17.63	14.26	12.94	131.05	17.86	24.30	133.51	24.76	17.86	15.78	49.16
DeepSeek-v3.1	18.97	12.30	9.67	8.59	103.94	13.15	18.87	**167.41**	21.35	13.02	10.87	49.49
DouBao-Seed-1.6	19.89	12.10	8.69	7.79	101.39	13.46	18.92	**167.41**	21.39	14.97	13.00	50.30
Kimi-K2-0905	11.29	4.92	3.01	2.27	47.40	6.34	9.95	61.94	12.85	6.97	5.51	49.24
GLM-4.5	14.88	7.71	5.27	4.78	66.32	10.44	13.66	91.77	17.03	9.96	8.74	**50.44**
LLaMA4-Maverick	21.26	15.16	12.79	11.70	113.54	15.86	22.01	162.42	23.29	15.39	13.29	47.44
Qwen3-235B	22.42	14.62	11.96	10.66	125.27	15.31	21.95	137.23	24.40	15.09	12.87	46.61
Qwen3-8B	19.07	13.52	11.24	10.04	98.02	13.23	20.34	77.12	20.68	13.30	11.27	47.14
LLaMA3.1-8B	18.80	12.78	10.27	9.07	61.53	12.32	22.94	66.66	21.57	14.09	11.84	39.56
Ministral-8B	18.62	14.05	12.20	10.97	83.08	13.31	21.80	60.12	20.85	12.77	10.85	42.96
Gemma3n-E4B	10.72	6.90	5.55	4.58	5.43	6.19	18.48	32.28	14.79	10.29	8.99	18.71

System prompt (translated): You are a senior standardized Yi-script language expert, proficient in bidirectional translation between Yi and Chinese, with deep understanding of Yi linguistic culture, character morphology, and phonological systems.

Core capabilities: Familiarity with 819 basic Yi syllable characters and their combinatorial rules; mastery of Yi–Chinese differences in syntax, word order, and classifier systems; understanding of Yi cultural context, idioms, and rhetorical traditions.

Translation principles: (1) Balance faithfulness, expressiveness, and elegance; (2) Context-first disambiguation; (3) Morphological-phonological-semantic inference for unfamiliar expressions.

Output requirements: Direct output without explanations; maintain register equivalence; use established translations for proper nouns; always provide complete output (never "I don't know").

10 few-shot examples: [2 from each of NuosuLex, NuosuSen, NuosuEpic, NuosuGov, NuosuEdu, covering Yi→Chinese translation, Chinese→Yi translation, English→Yi translation, multiple-choice, judgment, cloze, and word reordering tasks.]

Figure 9: Prompt Template

I General Capability Evaluation

To characterize BIMO-8B's general capability alongside its Yi-script performance, we evaluate on standard benchmarks alongside base and frontier models. Table 6 reports MMLU 74.4, GPQA 38.2, LiveBench 48.1, MATH-500 51.2, and AIME'25 31.3 for BIMO-8B, with a mean of 48.64. Joint reporting supports assessment of target-language adaptation and general-task performance. Each general-benchmark row mean uses its reported numerical entries.

J Source Allocation and Task Construction

Source allocation takes place on the post-augmentation inventory of 272,185 units and precedes TIEA expansion. The SFT construction pool contains 227,554 units, while the benchmark pools together contain 44,631 units. All derivatives follow the allocation of their source unit. The five benchmark source pools are mutually disjoint, and their assignment is separate from the SFT source pool.

NuosuLex–NuosuGov retain original gold reference answers. NuosuEdu uses TIEA to construct task instances from its 8,110-unit pool. This allocation provides a clear interpretation for the paired reporting of the full-suite score and the original-reference four-subset score.

Table 6: Performance comparison on Nuosu-specific and general ability benchmarks (Best and second-best results on Nuosu-Benchmark are highlighted in bold and underline respectively)

Model	Size	Nuosu-specific Ability						General Ability					
		NuosuLex	NuosuSen	NuosuEpic	NuosuGov	NuosuEdu	Avg.	MMLU	GPQA	LiveBench	MATH-500	AIME'25	Avg.
General-purpose Models													
GPT5	—	<u>50.91</u>	37.18	33.64	27.94	37.57	37.45	—	85.70	—	—	94.60	90.15
Claude4-ops	—	48.36	35.14	<u>44.59</u>	25.92	45.34	<u>39.87</u>	91.90	79.60	79.50	94.40	75.50	84.18
Gemini2.5-Pro	—	46.53	<u>37.24</u>	32.38	<u>48.19</u>	32.68	39.40	92.20	84.00	78.90	98.80	86.70	88.12
LLaMA4-Maverick	400B	46.34	**37.30**	32.21	36.06	42.05	38.79	85.20	69.80	55.20	63.30	15.90	57.88
DeepSeek-v3.1	685B	48.67	24.01	29.68	26.78	**51.83**	36.19	89.10	80.70	62.80	90.20	88.40	82.24
Qwen3-235B	235B	46.32	32.76	32.96	31.37	43.77	37.44	92.70	71.10	73.50	98.00	81.50	83.36
GLM-4.5	355B	11.89	22.07	12.86	30.15	36.92	22.78	90.00	79.10	—	98.20	—	89.10
Kimi-K2-0905	—	15.07	9.12	19.36	7.91	26.80	15.65	89.50	75.10	—	97.40	49.50	77.88
DouBao-Seed-1.6	—	48.46	22.09	41.39	23.67	45.75	36.27	91.00	81.50	—	—	86.00	86.17
8B-class Open-source Models													
Qwen3-8B	8B	27.98	32.69	27.86	29.32	30.06	29.58	87.50	62.00	67.10	97.40	67.30	76.26
LLaMA3.1-8B	8B	28.52	27.84	25.35	20.75	23.13	25.12	69.40	30.40	—	51.90	—	50.57
Ministral-8B	8B	23.29	32.90	26.06	29.79	21.96	26.80	65.00	—	—	54.50	—	59.75
Gemma3n-E4B	8B	13.85	11.38	14.89	9.76	9.67	11.91	64.90	23.70	—	—	11.60	33.40
BIMO-8B (Ours)	8B	**57.65**	29.06	**62.59**	**50.66**	<u>50.59</u>	**50.11**	74.4	38.2	48.1	51.2	31.3	48.64

Table 7: Reconciliation of the post-augmentation source inventory. NuosuEdu's 8,110 vocabulary units constitute its complete source pool.

Allocation	**Vocabulary**	**Sentence**	**Evaluation items**
SFT source pool	154,104	73,450	—
NuosuLex	13,581	0	13,581
NuosuSen	0	3,584	3,584
NuosuEpic	0	17,644	17,644
NuosuGov	0	1,712	1,712
NuosuEdu	8,110	0	73,992
Benchmark source pools	21,691	22,940	110,513
Complete inventory	**175,795**	**96,390**	

J.1 Evaluation Excluding NuosuEdu

Table 8 reports performance on NuosuLex, NuosuSen, NuosuEpic, and NuosuGov, which retain original gold references from their disjoint held-out source pools. All 14 models use the same 10-shot evaluation protocol. BIMO-8B leads this four-subset evaluation with a mean of 49.99, followed by Gemini 2.5-Pro (41.09) and Claude 4 Opus (38.50), extending its leading performance to the four non-TIEA subsets.

Table 8: Evaluation excluding NuosuEdu under the shared 10-shot protocol. The four subset scores are taken from the existing evaluations reported in rebuttal Table 7; the mean is their unweighted arithmetic average, rounded to two decimal places. Models are ordered by the four-subset mean. Bold and underlined values indicate the best and second-best results in each column.

Model	NuosuLex	NuosuSen	NuosuEpic	NuosuGov	Mean
BIMO-8B (Ours)	**57.65**	29.06	**62.59**	**50.66**	**49.99**
Gemini 2.5-Pro	46.53	<u>37.24</u>	32.38	<u>48.19</u>	<u>41.09</u>
Claude 4 Opus	48.36	35.14	<u>44.59</u>	25.92	38.50
LLaMA 4 Maverick	46.34	**37.30**	32.21	36.06	37.98
GPT-5	<u>50.91</u>	37.18	33.64	27.94	37.42
Qwen3-235B	46.32	32.76	32.96	31.37	35.85
DouBao-Seed-1.6	48.46	22.09	41.39	23.67	33.90
DeepSeek-V3.1	48.67	24.01	29.68	26.78	32.29
Qwen3-8B	27.98	32.69	27.86	29.32	29.46
Ministral-8B	23.29	32.90	26.06	29.79	28.01
LLaMA 3.1-8B	28.52	27.84	25.35	20.75	25.62
GLM-4.5	11.89	22.07	12.86	30.15	19.24
Kimi-K2-0905	15.07	9.12	19.36	7.91	12.87
Gemma-3n-E4B	13.85	11.38	14.89	9.76	12.47

K SFT Replay and Training Recipe

K.1 General-Purpose Replay Sources

The replay candidate pool combines five public resources (Table 9). Downloaded records are converted to Alpaca-format instruction, input, and output fields and deduplicated, producing 362,819 candidate examples. The resources contribute general Chinese instruction following, mathematical reasoning, and programming tasks to the SFT mixture.

Table 9: Public resources supplying the general-purpose replay pool.

Resource	Contribution
COIG-CQIA (Bai et al., 2024)	Chinese instructions across cultural, educational, social, and professional subjects
OpenR1-Math-220k (Open R1, 2025)	Mathematical problems and reasoning responses
Python Code Instructions 18k Alpaca (iamtarun, 2023)	Python programming instructions and responses
Alpaca-zh (shibing624, 2023)	Chinese instruction following and everyday question answering
CodeFeedback-Filtered-Instruction (M-A-P, 2024; Zheng et al., 2024a)	Programming instructions and code responses

K.2 Domain-Stratified Sampling

Sampling uses the following 15 domain strata:

`chinese_traditional`, `coig_pc`, `douban`, `exam`, `finance`, `human_value`, `logiqa`, `ruozhiba`, `segmentfault`, `wiki`, `wikihow`, `xhs`, `zhihu`, `math`, and `code`.

Approximately equal allocations select about 14,800 examples per stratum. With random seed 42, the final replay set contains 222,012 examples. The sampling target is expressed relative to the Yi SFT inventory:

$$N_{\mathrm{Yi}} = 1{,}110{,}064, \qquad N_{\mathrm{general}} = 222{,}012 \simeq 0.20 N_{\mathrm{Yi}}.$$

Thus, $N_{\mathrm{SFT}} = 1{,}332{,}076$, and the general-purpose share of the combined mixture is

$$\frac{222{,}012}{1{,}332{,}076} \times 100\% \simeq 16.7\%.$$

The reported 1,110,064 examples and 166M tokens describe the Yi SFT resource. The replay examples are an additional training component, providing coverage of the listed general-purpose domains.

K.3 Training Settings

BIMO-8B is initialized from Qwen3-8B and trained through LLaMAFactory on four NVIDIA A800-80G GPUs. Table 10 records the settings of each stage. CPT uses the collected Yi corpus distribution; SFT combines the Yi instruction corpus with the domain-stratified general-purpose replay set. The SFT context length of 1,280 tokens corresponds to the 99.9th percentile of Yi SFT sample lengths.

Table 10: BIMO-8B training settings for each adaptation stage.

Setting	**CPT**	**SFT**
Data	78.6M Yi tokens	1,332,076 mixed instructions
Epochs	5	1
Context length	8,192 tokens	1,280 tokens
Learning rate	2×10^{-4}	6×10^{-5}
Batch setting	Batch size 64	Micro-batch size 32
Schedule	Cosine decay	Cosine decay
Linear warmup	10%	10%
Hardware	4× A800-80G	4× A800-80G