Multitask Mongolian large language model for Chinese-Mongolian translation

Anonymous ACL submission

Abstract

Mongolian, a low-resource language, poses substantial challenges for training large language models (LLMs), especially for translation tasks that demand both syntactic precision and semantic coherence. To address the scarcity of high-quality parallel data and the inefficiency of existing fine-tuning methods, we build a 350,000-sentence Chinese–Mongolian parallel corpus and present a multi-task instruction-tuning framework using parameter-efficient LoRA (Low-Rank Adaptation). Our approach improves Mongolian translation and further generalizes to text summarization, question answering, and text classification. Experiments show that our method significantly outperforms baselines on BLEU-4 and ROUGE. For example, Baichuan-7B fine-tuned with 100% of the data achieves a BLEU-4 score of 0.2288, surpassing BLOOMZ-7B, LLaMA-8B, and ChatGLM-6B. Multi-task evaluations confirm strong instruction-following and robustness, while ablation studies indicate that LoRA fine-tuning and synthetic data augmentation jointly drive the gains. Overall, this work provides an efficient and scalable paradigm for Mongolian LLMs, offering a practical route for low-resource language modeling and a blueprint for the intelligent processing and digital preservation of minority languages.

1 Introduction

Large language models (LLMs) built on the Transformer architecture have advanced a broad set of NLP tasks, including machine translation, question answering, and summarization (Devlin et al., 2019; Raffel et al., 2020). Recent open models with strong Chinese coverage (e.g., LLaMA-family adaptations, ChatGLM, Qwen) are widely used as base models for downstream fine-tuning (Touvron et al., 2023; Du et al., 2021; Bai et al., 2023). However, most pre-training corpora and evaluation suites still center on high-resource languages, leaving low-resource and morphologically rich languages under-served.

Mongolian presents three practical obstacles. (*i*) **Data scarcity**: public, quality-controlled Chinese–Mongolian parallel data are limited and noisy. (*ii*) **Morphology and syntax**: agglutinative morphology (case markers, verbal suffixes) and flexible word order complicate modeling. (*iii*) **Script issues**: the traditional Mongolian script (vertical, context-sensitive shaping) is underrepresented in tokenizers, causing unstable segmentation and degraded semantics. As a result, general-purpose LLMs tend to produce grammatically inconsistent and semantically brittle translations in Mongolian settings.

We integrate corpus construction with parameter-efficient fine-tuning and instruction design. First, we curate a **350,000** pair Chinese–Mongolian parallel corpus with normalization, sentence segmentation, and one-to-one alignment. For instruction tuning, each bilingual pair is rendered into bidirectional prompt–response instances (Chinese→Mongolian and Mongolian→Chinese), yielding about **700k** instructionized samples. To study data scaling, we use a **10%** subset (≈35k pairs; ≈70k instances) and the **100%** set (350k pairs; ≈700k instances). Second, we adopt **LoRA** for parameter-efficient fine-tuning (PEFT), inserting low-rank adapters into attention submodules while freezing backbone weights (Hu et al., 2022). Third, we employ a **unified instruction format** that, beyond translation, includes Mongolian summarization, question answering, text classification, and named entity recognition to encourage cross-task generalization (Wei et al., 2021; Taori et al., 2023).

We study Chinese-centric base models representative of current practice—**Qwen2.5-7B-Instruct**, **Baichuan2-7B**, **ChatGLM3-6B**, and a **Chinese-adapted LLaMA-7/8B**—using identical instruction templates and LoRA configurations across

10% vs. 100% data scales. Translation is evaluated with **BLEU-4** and **ROUGE-1/2/L** (values in $[0, 1]$). For non-translation tasks we report ROUGE (summarization), macro-F1 (classification), and a calibrated automatic rating for QA, with small-sample human checks discussed in the appendix.

Contributions.

- **Corpus and protocol.** A quality-controlled **350k** Chinese–Mongolian parallel corpus and a **bidirectional instructionization** protocol (≈700k instances) suitable for PEFT and instruction tuning.
- **PEFT for Mongolian.** A **LoRA**-based adaptation pipeline that improves Mongolian translation under tight compute and data, with ablations on data scale and task mixture.
- **Unified instruction design.** A **single-format** interface covering translation, summarization, QA, classification, and NER to enhance instruction-following and transfer in Mongolian.
- **Empirical findings.** On multiple Chinese-centric LLMs, instruction+LoRA yields consistent gains over unfine-tuned baselines; scaling from 10% to 100% further improves n-gram and sentence-level metrics. Error analysis highlights remaining issues in case marking, long-range syntax, and domain terminology.

2 Related Work

2.1 Large Language Models and Instruction Tuning

Transformer-based LLMs have advanced translation, QA, and summarization via large-scale pretraining and instruction-following adaptation (Devlin et al., 2019; Raffel et al., 2020; Brown et al., 2020). Open models such as LLaMA, ChatGLM, Baichuan, and Qwen show strong Chinese performance (Touvron et al., 2023; Du et al., 2021; Bai et al., 2023), but low-resource and typologically distant languages remain under-served. Multilingual LMs (mBERT, XLM-R, BLOOMZ) leverage shared subword vocabularies, yet suffer token fragmentation and weakened syntactic modeling on morphologically rich languages like Mongolian.

Instruction tuning unifies tasks under an "instruction–input–response'' interface and improves zero-shot generalization (Wei et al., 2021; Taori et al., 2023; Zhang et al., 2024). However, instruction templates and high-quality resources tailored to minority languages are scarce, and evaluations often overlook script-specific constraints (e.g., traditional Mongolian).

2.2 Parameter-Efficient Fine-Tuning (PEFT)

PEFT adapts LLMs by updating a small subset of parameters—adapters, prefix/prompt tuning, and LoRA (Hu et al., 2022; Liu et al., 2021). LoRA inserts low-rank matrices into attention modules while freezing the backbone, reducing compute and mitigating forgetting; it is effective in English/Chinese settings but remains under-explored for minority, morphologically complex languages. We build on this line by combining LoRA-based adaptation with bilingual instruction tuning for Mongolian.

2.3 Mongolian MT and Resources

Mongolian is agglutinative with rich case/verb morphology and flexible word order, written in both Cyrillic and the traditional vertical script. Early MT used RBMT and PBSMT (Zens et al., 2002; Zhao et al., 2010; Wang et al., 2024), followed by neural Seq2Seq/attention models (Wu et al., 2016; Qing-Dao-Er-Ji et al., 2020). Progress is constrained by fragmented, noisy parallel data, limited domain coverage, and poor tokenizer support for traditional script (context-dependent shaping, non-LTR layout). These factors hinder effective adaptation of multilingual LLMs and motivate cleaner corpora and task-aware fine-tuning protocols for Mongolian.

3 Methodology

3.1 Overview

We propose a compact pipeline for Mongolian adaptation that couples (*i*) corpus construction and *instructionization* of Chinese–Mongolian sentence pairs, (*ii*) parameter-efficient fine-tuning (PEFT) with LoRA on Chinese-centric LLMs, and (*iii*) multi-task training beyond translation to improve instruction following and robustness.

3.2 Data and Instructionization

We curate a quality-controlled Chinese–Mongolian parallel corpus of **350,000** sentence

pairs (Section 2). Each pair is rendered into *two* instruction-style instances: Chinese→Mongolian and Mongolian→Chinese, producing ≈**700k** instances in total. To study data scaling, we use (**10%**) ≈35k pairs (≈70k instances) and (**100%**) 350k pairs (≈700k instances).

Preprocessing. We apply Unicode normalization, punctuation unification, sentence boundary detection for clausal segmentation, and 1–1 sentence alignment. For traditional Mongolian script, we use byte/character-level tokenization downstream to avoid over-segmentation from subword models; script shaping and directionality are handled at rendering time (training uses horizontal presentation).

Unified templates (examples). We follow an instruction–input–response format. Translation templates:
MN→ZH: *"Below is a segment in traditional Mongolian. Translate it into Chinese: {MN}. The translation is: {ZH}."*
ZH→MN: *"Below is a Chinese segment. Translate it into traditional Mongolian: {ZH}. The translation is: {MN}."*

Non-translation templates (summarization, QA, classification, NER) adopt the same interface; full prompts are provided in the appendix.

3.3 Base Models and PEFT with LoRA

We adapt representative Chinese-centric LLMs using a shared configuration: **Qwen2.5-7B-Instruct**, **Baichuan2-7B**, **ChatGLM3-6B**, and a **Chinese-adapted LLaMA-7/8B**. We freeze all backbone weights and train only LoRA modules on attention projections. Table 1 summarizes the model configurations and LoRA targets used in this study.

LoRA formulation. Given a linear map $W \in \mathbb{R}^{d\times k}$, LoRA learns a low-rank update

$$W' = W + \Delta W, \qquad \Delta W = \alpha BA, \quad (1)$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d,k)$. We target $\{\texttt{q\_proj}, \texttt{v\_proj}\}$ in each attention block (unless otherwise noted), which yielded stable gains under tight compute.

3.4 Training Objectives and Setup

We train with standard next-token prediction loss on the response portion while conditioning on instruction and input, as defined in Eq. (2). The key hyperparameters for both data scales are listed in Table 2.

Table 1: Models and LoRA settings (shared unless specified).

Model	Params	LoRA targets
Qwen2.5-7B-Instruct	7.6B	`q_proj`, `v_proj`
Baichuan2-7B	7.0B	`q_proj`, `v_proj`
ChatGLM3-6B	6.0B	`q_proj`, `v_proj`
LLaMA-7/8B (CN)	7–8B	`q_proj`, `v_proj`

Table 2: Key training hyperparameters (single-column for EACL).

Setting	10% data	100% data
Batch size (per GPU)	2	4
Max source / target	512 / 256	512 / 256
LoRA rank r / scale α	8 / 16	16 / 32
Trainable ratio	$\approx 0.26\%$	$\approx 0.53\%$
Epochs	5	2
LR / schedule	2×10^{-4}, warmup	same
Optimization	AdamW, grad. accum.	AdamW, grad. accum.
Early stopping	optional (val loss)	optional (val loss)

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\log p_\theta\Big(y_t^{(i)} \,\Big|\, y_{<t}^{(i)},\, x^{(i)}\Big), \quad (2)$$

where $x^{(i)}$ is the instruction+input, $y^{(i)}$ the target response, and θ includes backbone (frozen) and LoRA parameters.

Implementation notes. We use PyTorch with Accelerate for multi-GPU, PEFT for LoRA, and deterministic seeds. Evaluation uses sacreBLEU (detokenized, case-sensitive) and ROUGE-1/2/L ($[0,1]$ range); chrF/COMET and small-sample human evaluation are included in the appendix.

3.5 Multi-Task Instruction Mixture

To promote instruction following beyond translation, we mix Mongolian tasks—summarization, extractive QA, topic classification, and NER—into training using the same interface. Sampling follows a temperature-based schedule to avoid overwhelming translation with auxiliary tasks. At inference, tasks are inferred from the natural-language instruction. Examples of instructionized templates for each task are shown in Table 3.

3.6 Data Augmentation

We introduce three lightweight strategies to expand coverage and robustness under low-resource constraints.

Table 3: Instructionization at a glance (examples; full prompts in Appendix).

Task	Template snippet
Translate	*"Translate Mongolian into Chinese: {MN}. The translation is: {ZH}."*
Summarize	*"Summarize the following Mongolian text: {text}."*
QA	*"Read the passage and answer: {ctx} Q: {q}."*
Classify	*"Which category does this Mongolian text belong to? {text}."*
NER	*"List persons/locations/orgs in: {text}."*

Back-translation. Given Chinese monolingual sentences x_{zh}, we generate Mongolian $\hat{x}_{\text{mn}}$ and reconstruct $\hat{x}_{\text{zh}}$. Pairs are kept when similarity exceeds a threshold:

$$x_{\text{zh}} \xrightarrow{\text{M(zh}\to\text{mn)}} x_{\text{mn}} \xrightarrow{\text{M(mn}\to\text{zh)}} x_{\text{zh}} \quad \text{if} \quad \text{BLEU}(x_{\text{zh}}, x^*_{\text{zh}}) > \tau \tag{3}$$

Pseudo-instructions. From unlabeled Mongolian text, we synthesize summarization/QA/classification/NER samples using seed prompts and quality filters (length/entropy/consistency).

Adversarial variants. We create hard cases via word-order permutations (syntax-preserving), synonym substitutions, and register shifts (formal↔colloquial), mixed at low ratios to stabilize training.

3.7 Ablations and Controls (Design)

We design compact ablations to fit page limits: (i) LoRA target modules $\{\mathtt{q}, \mathtt{k}, \mathtt{v}, \mathtt{o}\}$, (ii) rank r and scale α, (iii) with/without auxiliary tasks, (iv) per-augmentation on/off. Full grids and exact prompts are deferred to the appendix for reproducibility.

4 Experiments and Results

4.1 Tasks, Baselines, and Settings

Mongolian-language evaluation tasks. We design the following Mongolian-language evaluation tasks to assess the effectiveness of instruction-tuned LLMs across realistic scenarios:

- **Translation**: Bidirectional translation between Chinese and traditional Mongolian (vertical script). For instance, the Mongolian sentence "[illegible]" is translated into Chinese as "你接了此案，就得负责到底。" We report **BLEU** and **ROUGE-L** to measure accuracy and fluency.

- **Summarization**: Given a long Mongolian document, the model generates a concise Mongolian summary. Quality is evaluated with **ROUGE-1/-2/-L**, reflecting lexical overlap and structural consistency with references.

- **Question Answering (QA)**: Given a Mongolian context and a question, the model produces a Mongolian answer. We use **GPT-4o** as an automatic judge to rate *correctness*, *completeness*, *fluency*, and *relevance* on a **1–5** scale (5 = fully correct, fluent, and contextually appropriate).

- **Text Classification**: Given a Mongolian sentence or paragraph, the model assigns a topic label (e.g., politics, education, entertainment). We report **accuracy** and **macro-F1** to reflect overall precision and class balance.

- **Named Entity Recognition (NER)**: Identify and label entities (persons, locations, organizations) in Mongolian text. We compute **precision**, **recall**, and **F1** at the entity level.

These tasks jointly cover generation and comprehension abilities and align with instruction-following Mongolian NLP applications.

Baselines. To establish reliable and representative baselines for Mongolian tasks, we evaluate five widely used Chinese-centric LLMs with strong multilingual capabilities and proven instruction-following performance. They serve as foundations for our fine-tuning and enable comparisons across architectures and training paradigms. **Qwen 2.5-7B** is a 7-billion-parameter model developed by Alibaba Cloud and DAMO Academy, with optimized RoPE, stronger throughput, and multilingual instruction-tuning support(Bai et al., 2023). **LLaMA 3-7B (Chinese-adapted)** is the third-generation LLaMA model with added Chinese data and long-context optimization(Touvron et al., 2023). **ChatGLM 3-6B** is a bilingual dialogue model from Tsinghua KEG and Zhipu AI, supporting LoRA/PEFT and robust multi-turn QA(Du et al., 2021). **Baichuan 2-7B** features improved bilingual pretraining, longer context handling, and strong multilingual generation(Yang et al., 2023).

InternLM-7B offers strong instruction-tuned performance and open accessibility, supporting diverse Chinese instructions(Team, 2023). All baseline models are compatible with instruction tuning and PEFT techniques such as LoRA.

Experimental settings. We fine-tune the Qwen2.5-7B-Instruct model using LoRA. LoRA inserts lightweight trainable matrices into specific attention submodules while freezing the backbone. The base model contains $\sim$ 7.6B parameters; the proportion of trainable parameters is about 0.26% (10% data) and 0.53% (100% data). Two training scales are adopted: 10% (35k pairs) and 100% (700k pairs). Training runs on an AutoDL H20 GPU (96 GB) with PyTorch and `Accelerate`; monitoring is via TensorBoard and Weights & Biases (WandB). We use instruction-style prompt templates for both training and evaluation. While bidirectional Chinese–Mongolian translation is central, we also train four additional instruction tasks to enhance generalization: summarization, QA, text classification, and NER (metrics as defined above). Warm-up scheduling and gradient accumulation are used; early stopping is optional based on validation loss; a unified instruction format enables task inference from natural language.Hyperparameters are summarized in Table 4.

Table 4: Training Hyperparameters

Parameter	10% Dataset	100% Dataset
Batch size	2	4
LoRA rank (r)	8	16
LoRA scaling (α)	16	32
Epochs	5	2
Max source length	512	512
Max target length	256	256
Learning rate	2×10^{-4}	2×10^{-4}
Gradient accumulation	8	4

4.2 Training Dynamics

Overall training efficiency and final loss are listed in Table 5.

Table 5: Training efficiency and final loss under different data scales

Data	Final Loss	Time (h)	Samples/s	Steps
10%	0.3606	14.8	3.28	10k
100%	0.4567	105	3.70	109k

In the 10% data experiment, the model converged after five epochs, with the validation loss reaching a stable level (see Fig. 1 and Fig. 2). The loss reduction trend over steps is further illustrated in Fig. 3.

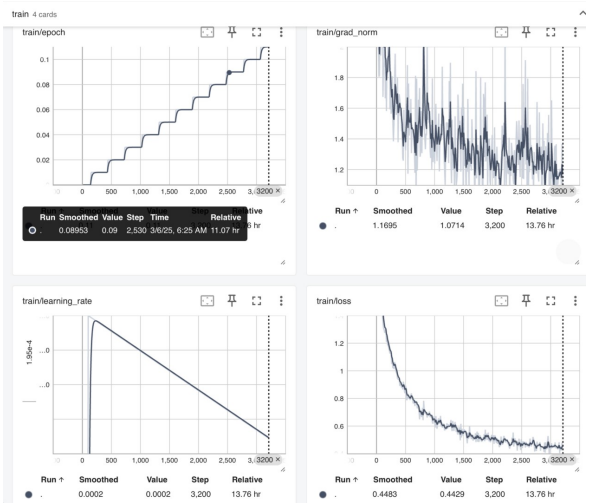


Figure 1: Training curves on the full dataset

After training on 100% of the data, the model's generalization capability improved significantly, making it more suitable for real-world deployment (cf. validation trend in Fig. 2).

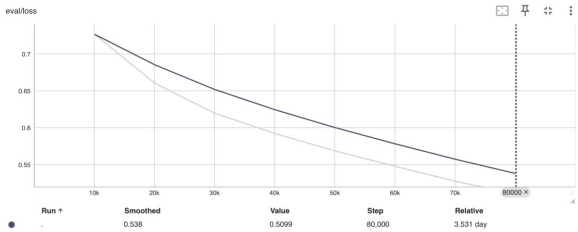


Figure 2: Validation loss curve during training

The validation loss (eval/loss) showed a consistently decreasing trend throughout training. Starting from an initial value of approximately 0.7, the loss gradually declined as training steps progressed and eventually reached 0.5099 at step 80,000. The smoothed loss value was approximately 0.538, indicating stable validation performance without signs of severe overfitting or underfitting. Although the rate of loss reduction slowed during later stages, the overall trend remained positive, reflecting gradual improvements in model generalization.

During the early phase of training (approximately the first 10k steps), the loss rapidly dropped from around 2.0 to 0.5 as the model quickly learned

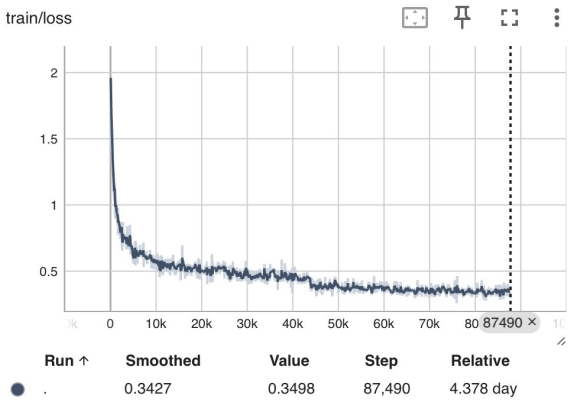


Figure 3: Training loss curve over steps

key features from the data. In the subsequent phase (after 10k steps), the loss plateaued with relatively low fluctuation, indicating convergence and stable training dynamics. Some minor oscillations were observed in later stages, likely due to data noise and inconsistent performance on a small subset of samples.

4.3 Translation Results

We analyze translation quality for the baseline (unfine-tuned), the 10% fine-tuned, and the full-data fine-tuned models using the outputs under `./translation_results_base/`, `./translation_results_0.1/`, and `./translation_results_full/`, respectively. Baseline outputs frequently exhibit literal renderings of Chinese expressions and contextual inconsistencies. By contrast, the fine-tuned models produce more fluent Mongolian-to-Chinese (M2C) translations with markedly improved semantic alignment. Chinese-to-Mongolian (C2M) translations show correct word order and appropriate use of traditional Mongolian grammatical phenomena (e.g., case inflection and subject–object agreement). To comprehensively assess performance, we report **BLEU-4** and **ROUGE-1/2/L** across all base models. The test set contains **2,000** Chinese–Mongolian sentence pairs spanning daily conversations, news, and official documents. Aggregate scores are reported in Table 6 and visualized as single-metric bar charts in Fig. 4a–4d.

From the evaluation results, **Baichuan-7B** and **BLOOMZ-7B** deliver the best overall performance among all baselines. With 100% training data, Baichuan-7B attains the highest ROUGE scores (e.g., **ROUGE-1 = 0.6119**), while BLOOMZ-7B leads slightly on **BLEU-4** (**0.2300**). In contrast, **LLaMA-8B** and **ChatGLM-6B** perform less effectively. All models improve substantially from 10% to 100%, underscoring the importance of data volume. Across metrics, **ROUGE-1** is consistently higher than others; **BLEU-4** is relatively lower, reflecting challenges in longer phrase generation; **ROUGE-2** and **ROUGE-L** fall in between.

4.4 Impact of Training Data Volume

To further investigate how training data volume affects performance, we computed the improvement rates from 10% to 100% training data (Table 7). A complementary visualization of percentage gains is provided in Fig. 5.

The results indicate that **BLOOMZ-7B** and **ChatGLM-6B** are more sensitive to data scaling; **Baichuan-7B** shows robust but moderate relative gains; **LLaMA-8B** exhibits smaller improvements.

4.5 Multi-Task Evaluation

Although translation is primary, real-world Mongolian NLP requires additional capabilities. We construct three Mongolian task sets (summarization, QA, classification) and evaluate Qwen-7B at 10% and 100% (Table 8).

Increasing data consistently boosts performance across tasks; QA shows the largest gain, indicating stronger instruction following and semantic consistency.

4.6 Ablation Study

We conduct four ablation settings on Qwen-7B using 100% data; BLEU-4 is reported in Table 9, and the corresponding bar chart is shown in Fig. 6.

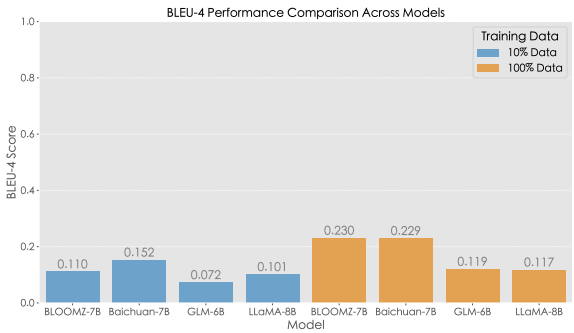
LoRA is the primary driver of accuracy gains; multi-task instruction improves structure and expressiveness; augmentation adds robust but smaller gains (visualized in Fig. 6).

4.7 Error Analysis and Visualizations

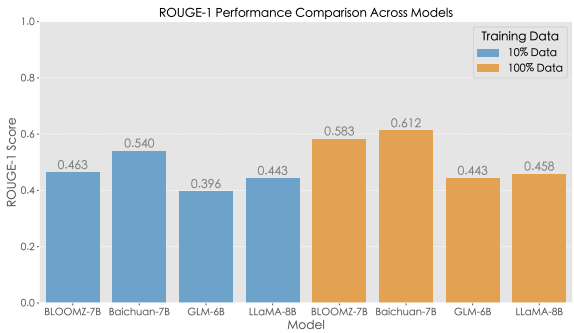
To better understand limitations, we annotate validation outputs. Frequent errors include missing grammatical markers, semantic ambiguity, incomplete sentence structure, and terminology mistranslation. Representative examples and their analyses are summarized in Appendix A.1 (Table 10), while the aggregate error-type distribution is presented in Appendix A.2 (Fig. 7). These results collectively

Table 6: Average translation performance of different models on Mongolian tasks

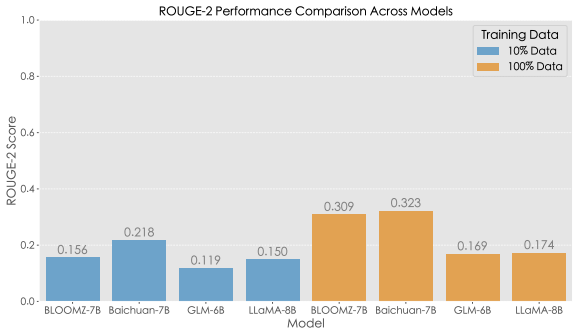
Model	Data Scale	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Baichuan-7B	10%	0.1522	0.5401	0.2184	0.3724
	100%	0.2288	0.6119	0.3228	0.4756
LLaMA-8B	10%	0.1008	0.4426	0.1498	0.2652
	100%	0.1170	0.4581	0.1737	0.2707
ChatGLM-6B	10%	0.0722	0.3960	0.1187	0.2288
	100%	0.1195	0.4430	0.1695	0.2881
BLOOMZ-7B	10%	0.1103	0.4633	0.1563	0.2875
	100%	0.2300	0.5828	0.3095	0.4517



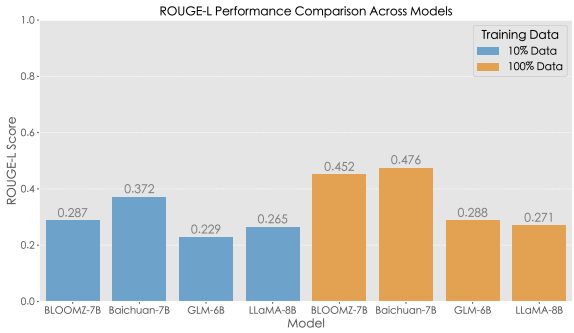
(a) BLEU-4 score comparison



(b) ROUGE-1 score comparison



(c) ROUGE-2 score comparison



(d) ROUGE-L score comparison

Figure 4: Single-metric bar charts combined as a 2×2 grid (filenames 6–9; order preserved).

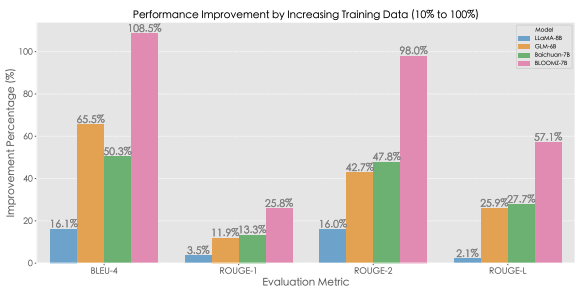


Figure 5: Performance improvement from 10% to 100% training data

Table 7: Percentage performance improvements (10% → 100%)

Model	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
BLOOMZ-7B	108.5%	25.8%	98.0%	57.1%
ChatGLM-6B	65.5%	11.9%	42.7%	25.9%
Baichuan-7B	50.3%	13.3%	47.8%	27.7%
LLaMA-8B	16.1%	3.5%	16.0%	2.1%

Table 8: Multi-task performance of Qwen-7B under different data scales

Task Type	Metric	10%	100%
Summarization	ROUGE-L	0.426	0.481
QA	Human (1–5)	3.21	4.05
Classification	Macro-F1 (%)	68.4	81.2

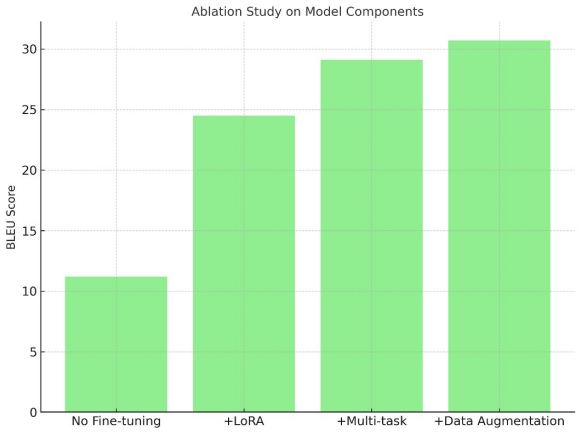


Figure 6: BLEU-4 under ablation configurations

Table 9: BLEU-4 across ablation configurations (incremental gains relative to previous row)

Configuration	BLEU-4	Performance Change
No fine-tuning (baseline)	11.2	—
+ LoRA fine-tuning (translation only)	24.5	+13.3
+ Multitask instruction fine-tuning	29.1	+4.6
+ Data augmentation (back-translation + noise)	30.7	+1.6

highlight typical linguistic challenges encountered by the model in Mongolian-to-Chinese translation, particularly the omission of case markers and ambiguity in complex syntactic constructions.

4.8 Discussion

The results highlight several key insights into adapting Chinese-centric LLMs for Mongolian. First, the strong and consistent gains achieved by LoRA-based fine-tuning confirm that parameter-efficient adaptation is highly effective even under severe data constraints. Second, models pretrained with balanced bilingual data, such as Baichuan-7B and BLOOMZ-7B, exhibit higher transferability and stability across both translation and auxiliary tasks, suggesting that pretraining coverage plays a more critical role than sheer model scale. Third, incorporating instruction-style supervision across multiple Mongolian NLP tasks improves overall linguistic alignment and robustness, enabling the model to generalize beyond translation to tasks like summarization and QA without explicit retraining. Finally, the observed performance trends—especially the diminishing returns of data scaling for certain architectures—point to the need for more linguistically informed adaptation strategies rather than purely increasing corpus size or parameter count. Together, these findings provide empirical evidence that well-designed instruction mixtures and lightweight adaptation can substantially enhance low-resource language modeling efficiency and usability.

5 Conclusion

This study presents an efficient framework for Mongolian translation and multi-task understanding by constructing a 350,000-sentence parallel corpus and applying LoRA-based fine-tuning with multi-task instruction tuning. The approach significantly improves translation quality and generalization on tasks like summarization and QA, offering a practical and reproducible paradigm for low-resource language modeling. Future work will focus on corpus expansion and cross-modal applications to further advance Mongolian NLP.

Limitations

Our framework is limited by the relatively small size and limited domain diversity of the Mongolian corpus compared with high-resource languages. Evaluation mainly relies on automatic metrics (BLEU/ROUGE) and GPT-based judgments, which may not fully capture linguistic adequacy. Traditional Mongolian script is processed in a horizontal layout rather than true vertical rendering, and complex morphology is only indirectly modeled. Finally, our focus is on building a practical and reproducible low-resource adaptation recipe rather than pursuing state-of-the-art results, leaving deeper linguistic modeling and multimodal extensions for future work.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Ren Qing-Dao-Er-Ji, Yila Su, and Nier Wu. 2020. Research on mongolian-chinese machine translation based on the end-to-end neural network. *International Journal of Wavelets, Multiresolution and Information Processing*, 18(01):1941003.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Bailun Wang, Yatu Ji, Nier Wu, Xu Liu, Yanli Wang, Rui Mao, Shuai Yuan, Qing-Dao-Er-Ji Ren, Na Liu, Xufei Zhuang, and 1 others. 2024. Mongolian-chinese neural machine translation based on sustained transfer learning. In *International Conference on Intelligent Computing*, pages 304–315. Springer.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jing Wu, Hongxu Hou, Zhipeng Shen, Jian Du, and Jinting Li. 2016. Adapting attention-based neural network to low-resource mongolian-chinese machine translation. In *International Conference on Computer Processing of Oriental Languages*, pages 470–480. Springer.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *Annual Conference on Artificial Intelligence*, pages 18–32. Springer.

Jinyi Zhang, Ke Su, Haowei Li, Jiannan Mao, Ye Tian, Feng Wen, Chong Guo, and Tadahiro Matsumoto. 2024. Neural machine translation for low-resource languages from a chinese-centric perspective: A survey. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(6):1–60.

Lili Zhao, Jia Men, Congpin Zhang, Qun Liu, Wenbin Jiang, Jinxing Wu, and Qing Chang. 2010. A combination of statistical and rule-based approach for mongolian lexical analysis. In *2010 international conference on asian language processing*, pages 7–10. IEEE.

A Supplementary Visualizations and Error Analysis

A.1 Representative Error Cases

Table 10 lists representative Mongolian-to-Chinese translation errors identified during validation. Each case is categorized by error type and accompanied by a brief analysis describing the underlying issue, such as missing grammatical inflections, semantic ambiguity, or tonal attenuation.

A.2 Aggregate Error-Type Distribution

Figure 7 visualizes the overall distribution of error types observed in the validation set. The figure reveals that grammatical omissions (e.g., missing case particles and verb inflections) constitute the majority of errors, followed by semantic ambiguities and incomplete sentence structures, indicating that morphosyntactic regularities remain challenging for the model to fully capture.

Table 10: Typical errors in Mongolian-to-Chinese translation

Input Mongolian (abbr.)	**Model Output (ZH)**	**Error Type**	**Analysis**
[illegible] …	我不清楚你是谁	Grammatical omission	Missing case marking; incomplete ending
[illegible] …	他很关心教育问题	Semantic ambiguity	“对中小学改革意见” generalized to vague stance
[illegible] …	他们可能有点不同意	Attenuated tone	Missed “strong opposition” polarity

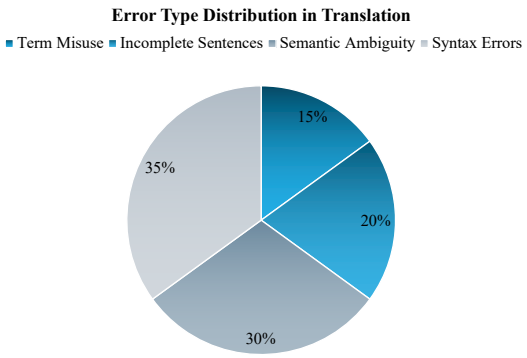


Figure 7: Error-type distribution in Mongolian translation