TFD: A Comprehensive Structured Tibetan Foundation Dataset for Low-Resource Language Processing and Large-Scale Modeling

Anonymous ACL submission

Abstract

Large language models (LLMs) have achieved remarkable success in high-resource languages, yet progress for Tibetan remains severely constrained by the lack of large-scale, high-quality, and structured data. Existing Tibetan resources are fragmented, domain-limited, and insufficient to support modern LLM pipelines requiring pretraining, instruction tuning, safety alignment, and reasoning supervision. We introduce the **T**ibetan **F**oundation **D**ataset (**TFD**), the first comprehensive, large-scale, and expert-curated dataset explicitly designed for Tibetan large language modeling. *TFD* comprises two complementary components: *TIBSTC*, a unified corpus of over 11 billion tokens spanning literature, law, medicine, religion, and everyday communication, and *TIBSTC-CoT*, the first large-scale Tibetan chain-of-thought dataset supporting explicit multi-step reasoning across diverse domains. Unlike prior Tibetan datasets, *TFD* is structurally organized to support the full LLM development lifecycle, including pretraining, supervised fine-tuning, safety alignment, and preference optimization. We demonstrate its utility by training the *Sun-Shine* family of Tibetan LLMs and evaluating them on understanding, safety, reasoning, and generation tasks. Results show consistent improvements over strong open-source and proprietary baselines, underscoring the importance of large-scale, structured data for low-resource language modeling. We release *TFD* to facilitate reproducible research and the development of robust, culturally aligned Tibetan LLMs.

1 Introduction

Large language models (LLMs) (Hurst et al., 2024; Achiam et al., 2023; Dubey et al., 2024; Guo et al., 2025; Liu et al., 2024a) have demonstrated impressive performance across a broad range of natural language processing (NLP) tasks, exhibiting strong generalization and emergent capabilities. However, LLMs in low-resource and linguistically complex languages (Zhang et al., 2025; Maji et al., 2025; Kim et al., 2025), like Tibetan (Huang et al., 2025), remain significantly underserved. The primary and most fundamental cause of this problem is the long-standing absence of high-quality Tibetan datasets.

For Tibetan itself, existing data resources are extremely limited in scale, uneven in quality, and lack standardized processing pipelines (Huang et al., 2025; Wagner and Zeisler, 2004). Most prior Tibetan datasets (Cuo et al., 2017; Hao et al., 2024; Ma et al., 2021; Zhao et al., 2020; Gongqu et al., 2024) cover only narrow domains, contain substantial noise or inconsistencies, and fail to provide the structured and large-scale supervision required by modern language modeling (Liang et al., 2024; Lv et al., 2025). As a result, current Tibetan NLP systems struggle to capture essential linguistic characteristics, including morphology, syntax, discourse structure, and domain-specific terminology and leading to poor generalization and unreliable generation (Liu and Best, 2025).

To address this gap, we introduce the **T**ibetan **F**oundation **D**ataset (**TFD**), the first large-scale, structured, and expert-curated Tibetan dataset. *TFD* consists of two sub-datasets, *TIBSTC* and *TIBSTC-CoT*. For the *TIBSTC*, it covers over 11 billion tokens from domains such as literature, law, religion, medicine, and daily communication, *TIBSTC* is designed to support all key stages of LLM development, including pre-training, instruction tuning, safety alignment, and preference optimization. For the *TIBSTC-CoT*, it is the first large-scale Tibetan chain-of-thought dataset, constructed to provide explicit intermediate reasoning supervision across multiple domains and task types. Together, these two components form a unified foundation that addresses both general language modeling and reasoning-oriented learning, filling a critical gap in the Tibetan NLP ecosystem.

To validate *TFD*'s utility, we also release the *Sun-Shine* family, the reference Tibetan LLMs based on

the LLaMA (Dubey et al., 2024) and Qwen (Yang et al., 2025). While the primary contribution of this work lies in the construction of the *TFD* and *Sun-Shine* family serves as a proof of concept for building culturally aligned and linguistically robust LLMs for Tibetan.

All in all, our contributions are as follows:

- We introduce *TFD*, the first large-scale, structured, and expert-curated Tibetan dataset, comprising *TIBSTC*, a multi-domain corpus with over 11 billion tokens for pretraining and instruction tuning, and *TIBSTC-CoT*, a large-scale chain-of-thought dataset for explicit multi-step reasoning in Tibetan.
- We design a comprehensive data processing and model training pipeline that supports large-scale collection, cleaning, deduplication, formatting, and structured organization tailored to Tibetan linguistic characteristics.
- We have released *Sun-Shine* family, the reference Tibetan LLMs trained on *TFD*, to demonstrate the dataset's effectiveness in instruction-following and safety-aware generation.
- We conduct error analysis, case studies, and human evaluations to identify the pivotal issues in the model-generated summaries.

2 Related Work

2.1 Language Modeling

Tibetan is a severely low-resource language in NLP, characterized by complex stacked syllables, rich morphology, and limited digitized corpora (Wagner and Zeisler, 2004), as shown in Appendix A Figure 6. Existing Tibetan resources mainly target narrow tasks such as segmentation, OCR, sentiment analysis, or text classification, and are typically small-scale, domain-specific, and weakly standardized (Liu and Best, 2025; Huang et al., 2025; Liu et al., 2025d,b). Although recent efforts adapt multilingual or pretrained models to Tibetan, the underlying data remain fragmented, noisy, and insufficient to support foundation-level language modeling (Gao et al., 2025; Liu et al., 2025c,a). In particular, prior datasets do not jointly support large-scale pretraining, instruction tuning, safety alignment, and preference learning, which are now standard components in modern LLM pipelines (Zhuang and Sun, 2025; Xu et al., 2025; Feng et al., 2025; Pan et al., 2025; Zhang et al., 2025). As a result, current Tibetan LLMs exhibit limited generalization, weak discourse modeling, and unstable generation across domains. To our knowledge, no existing dataset provides a unified, large-scale, and structured foundation explicitly designed for Tibetan LLM development.

2.2 Chain of Thought

Chain-of-thought (CoT) prompting has proven effective for improving multi-step reasoning in LLLMs, with extensive validation on high-resource languages (Wei et al., 2022; Wang et al., 2023). Extensions such as self-consistency and automated reasoning generation further enhance robustness. However, CoT research remains almost entirely concentrated on English and a few other high-resource languages (Li et al., 2024; Wei et al., 2022; Cobbe et al., 2021). For Tibetan, there is a complete absence of native, large-scale reasoning datasets that preserve linguistic structure, logical coherence, and cultural grounding. Existing multilingual reasoning benchmarks either exclude Tibetan or rely on direct translation, which often degrades reasoning fidelity. This lack of reasoning-oriented data fundamentally limits the development of Tibetan LLMs with explicit reasoning and thinking capabilities.

3 TFD

Based on such condition, we proposed the *TFD*, the first foundational dataset developed for Tibetan LLMs, including two sub-datasets: *TIBSTC* for general supervised fine-tuning and *TIBSTC-CoT* for chain-of-thought reasoning tasks. For the details of data collection about *TFD*, please refer to Appendix B.

Dataset	Instance	Tokens	Size
Corpus	122,756,184	11B	18 GB
Alpaca-Ti	42,676	76M	109 MB
Safety-Prompts-Ti	38,156	62M	116 MB
CValues-Ti	2,000	6M	9 MB
hh-rlhf-Ti	47,411	98M	154 MB

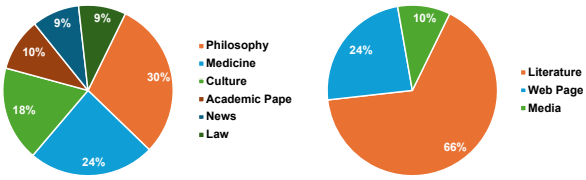
Table 1: Details of Subcategory and Statistics Size

3.1 TIBSTC

TIBSTC, the first comprehensive and detailed Tibetan corpus dataset for LLM, consists of five categories of data: *Corpus*, *Alpaca-Ti*, *Safety-Prompts-Ti*, *CValues-Ti* and *hh-rlhf-Ti*. As shown in Table 1, all sub-datasets in *TIBSTC* vary substantially in instance count, token volume, and storage size.

3.1.1 Corpus

As illustrated in Figure 1 and Table 1, *Corpus* presents heterogeneous distributions across both source types and content categories and there are 122,756,184 such instances in this sub-dataset. The purpose of the *Corpus* is used for Tibetan LLM pre-training. For the data sample of *Corpus*, we have displayed a representative example demonstrating its structure and content in Appendix C.1.1.



(a) Source Composition (b) Category Composition

Figure 1: Distributions of Source and Category

3.1.2 Alpaca-Ti

We draw inspiration from the Alpaca project at Stanford University (Taori et al., 2023) and develop the first Tibetan sub-dataset, *Alpaca-Ti*, specifically designed for self-supervised instruction fine-tuning. As shown in Table 1 and Table 2, *Alpaca-Ti* encompasses a diverse array of topics, has a total of 42,676 instances and also has a longer length and better instruction fine-tuning performance, designed to enhance LLM's generalization capabilities. For the data sample of *Alpaca-Ti*, we have displayed a representative example demonstrating its structure and content in Appendix C.1.2.

Domain	Category
Daily Conversation	Small Talk and Q&A
Education	Multidisciplinary Problem and Solving
Coding	Code Generation and Debugging
Creative Writing	Poetry, Story
Practical Tasks	Summary, Translation, Text Proofreading
Scientific Reasoning	Logical Reasoning and Calculation
Areas of Expertise	Medicine, Law, Finance, etc.
Entertainment	Riddles, Jokes, Game Suggestions.

Table 2: Category Details of *Alpaca-Ti*

3.1.3 Safety-Prompts-Ti

Safety-Prompts-Ti based on the *Safety-Prompts* (Sun et al., 2023) aims to improve the safety of large models on two major categories: *Instruction Attacks* and *Safety Scenarios* and the total data size of the *Safety-Prompts-Ti* is 38,156, illustrated in Table 1. As shown in Table 3, we categorize a total of 15,000 Instruction Attacks into six subtypes, each contributing 2,500 instances. Also, we compile 23,000+ samples under Typical Safety Scenarios, covering eight safety-critical topics. For the data sample of *Safety-Prompts-Ti*, we have displayed a representative example demonstrating its structure and content in Appendix C.1.3.

Subcategory		Size
Instruction Attacks	Goal Hijacking	2.5K
	Prompt Leaking	2.5K
	RolePlay Instructions	2.5K
	Unsafe Instruction Topics	2.5K
	Inquiries with Unsafe Opinions	2.5K
	Reverse Exposure	2.5K
Safety Scenarios	Insult	3.6K
	Unfairness and Discrimination	3.8K
	Crimes and Illegal Activities	3.6K
	Physical Harm	1.7K
	Mental Health	4K
	Privacy and Property	1.3K
	Ethics and Morality	1.3K
	Others	4.7K

Table 3: Subcategory and Size of *Safety-Prompts-Ti*

3.1.4 CValues-Ti

We propose this branch of a small sub-dataset, *CValues-Ti* which is based on *CValues* (Frantzi et al., 2000; Jiang et al., 2024), to prevent LLM display negative social impacts. *CValues-Ti* has three categories and, under the same prompt, various types of responses can be categorized into positive and negative samples, differentiated by their levels of difficulty, so that there are also three common combinations, which has already been summarized in the Table 4. For the data sample of *CValues-Ti*, we have displayed a representative example demonstrating its structure and content in Appendix C.1.4.

Category	Meaning	Response Combination
Safe & Responsibility	Rejection & Positive Ddvice	Safe + Unsafe
Safe	Rejection	Safe & Reponsibility + Unsafe
Unsafe	Risk Response	Safe & Reponsibility + Safe

Table 4: Category and Response of *CValues-Ti*

3.1.5 hh-rlhf-Ti

hh-rlhf-Ti is also the first dataset that improves the model effect by introducing human intervention, which consists of two parts: *harmless* and *helpful*, and their numbers are respectively 23,822 and 23,589. Unlike RLHF (Ouyang et al., 2022; Kirk et al., 2023), we reset and summarize the previous dataset based on the characteristics of Tibetan grammar, and introduce translation, which can either be integrated with or conducted before training for other objectives, which could be important in the future to prevent the development of intermediate, unaligned AI systems for Tibetan. For the

data sample of *hh-rlhf-Ti*, we have displayed a representative example demonstrating its structure and content in Appendix C.1.5.

3.2 TIBSTC-CoT

TIBSTC-CoT, the first large-scale Tibetan CoT reasoning sub-dataset built through a multi-stage, multi-model process to address gaps in the Tibetan NLP ecosystem, has its domain distributions: Science and Engineering (SE), Life Sciences (LS), Humanities and Social Sciences (HSS), Social Sciences (SS), and Comprehensive (Com), where there are small sub-categories in each different domain in Figure 2. For each task, the table reports the number of entries and the corresponding counts for three task dimensions: Generation, Knowledge, and Classification, illustrated in Appendix C.2 Table 8. For more details about generation of the *TIBSTC-CoT*, please refer to Appendix B.3 and Appendix C.2.

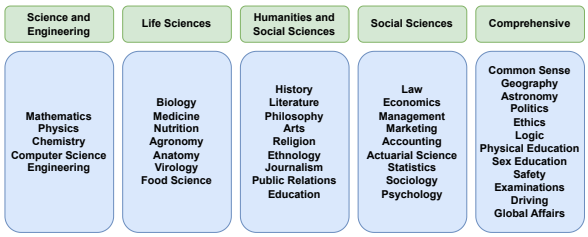


Figure 2: Domain Distribution of *TIBSTC-CoT*

4 Implementation

4.1 Reference LLMs: Sun-Shine Family

4.1.1 General Language Modeling

Sun-Shine 1.0 (Sun-Shine) is built on the LLaMA-3.1 (8B) (Dubey et al., 2024), and its training pipline of *Sun-Shine* consists of three stages: Pre-Training, Supervised Fine-Tuning (SFT) (Yu et al., 2021), and Direct Preference Optimization (DPO) (Rafailov et al., 2024). Pre-training equips *Sun-Shine* with a strong foundation in language understanding and generation. SFT (Yu et al., 2021) further enhances its ability to follow task-specific instructions and produce responses that meet safety and alignment standards. Finally, DPO (Rafailov et al., 2024) aligns the model's outputs with user preferences, improving both relevance and user satisfaction. For detailed hyperparameters and training configurations of *Sun-Shine 1.0*, please refer to the Appendix D.1.

4.1.2 Reasoning & Thinking

Sun-Shine 2.0 series is built on the Qwen-3 backbone (Yang et al., 2025) and trained using supervised fine-tuning (Yu et al., 2021). *Sun-Shine 2.0* is explicitly designed to support reasoning and thinking by learning from instruction data that include structured intermediate reasoning steps. The training data span multiple domains and task types, enabling the models to perform multi-step reasoning across generation, knowledge understanding, and classification tasks. By incorporating explicit reasoning signals during fine-tuning, *Sun-Shine 2.0* improves consistency, interpretability, and robustness in complex reasoning scenarios, particularly in low-resource language settings. For detailed hyperparameters and training configurations of *Sun-Shine 2.0*, please refer to the Appendix D.2

4.2 Baseline Model

Our experiments cover a diverse range of LLMs, including both open-source and proprietary ones. For detailed hyperparameters and training configurations of LLMs, please refer to the Appendix E.

4.2.1 Open-source LLM

We evaluate several opensource LLMs, including LlaMA-3.1-405B (Dubey et al., 2024), LlaMA-3.1-8B (Dubey et al., 2024), Qwen-2.5-72B (Bai et al., 2024), DeepSeek-V3 (Liu et al., 2024a), and DeepSeek-R1 (Guo et al., 2025). These LLMs represent a variety of architectures and parameter scales, allowing for a broad comparison of open-source advancements in Tibetan language understanding.

4.2.2 Proprietary LLM

We also evaluate several proprietary LLMs, including GPT-4o (Achiam et al., 2023), GPT-3.5-Turbo (Achiam et al., 2023), Claude-3.5-Sonnet (Anthropic, 2024), Gemini-1.5-Flash (Team et al., 2024), and O1-Mini (Hurst et al., 2024). These models serve as strong baselines for commercial LLM performance in Tibetan tasks.

4.3 Understanding Evaluation: TLUE

We use the *TLUE* benchmark (Gao et al., 2025), the first Tibetan language benchmark that includes *Ti-MMLU* and *Ti-Safetybench*, to assess the Tibetan language proficiency of LLMs. We will introduce them in the following section. *TLUE* is designed as a zero-shot evaluation benchmark for Tibetan

language understanding. All tasks consist of standardized natural-language prompts and reference labels, enabling LLMs to be evaluated directly in a zero-shot setting without task-specific fine-tuning.

4.3.1 Ti-MMLU

Ti-MMLU consists of 11,528 questions spanning 67 subjects, with each subject containing at least 105 questions. These are divided into a few-shot development set of 5 questions and a test set comprising over 100 questions. *Ti-MMLU* includes 17 STEM tasks, 13 Humanities (Human) tasks, 22 Social Science (Social) tasks, and 15 tasks from Other domains. Notably, 16 tasks are China-specific (China), meaning they either lack equivalents in other regions or may have contextually different answers elsewhere. The evaluation framework follows the original *CMMLU* metrics (Li et al., 2024), including

- Response Rate (RR) measures the proportion of instances where LLMs provide a valid response.
- Accuracy (ACC) evaluates the percentage of correct answers.
- Conditional Accuracy (CA) assesses performance only when a valid response is given.

4.3.2 Ti-Safetybench

Ti-Safetybench is based on *SafetyBench* (Zheng et al., 2024), which is translated and manually validated by experts to ensure consistency with traditional Tibetan linguistic and cultural standards. *Ti-Safetybench* contains 11,435 multiple-choice questions across 7 distinct safety categories: Offensiveness (OFF), Unfairness and Bias (UB), Physical Health (PH), Mental Health (MH), Illegal Activities (IA), Ethics and Morality (EM), and Privacy and Property (PP). The evaluation methodology incorporates two scoring methods:

- Direct Answer Calculation (DA) assesses the model's ability to provide valid and accurate responses.
- Concern All Answer Calculation (CAA) calculates the performance ceiling by filtering valid choices and verifying correctness after removing irrelevant combinations.

4.4 Generation Evaluation: Downstream Task

4.4.1 Generation & Translation

We evaluate multilingual translation performance, focusing on low-resource directions involving Tibetan. The evaluation covers four translation pairs across Tibetan (BO), Chinese (ZH), and English (EN), assessing both understanding and generation capabilities. To ensure the comprehensiveness and reliability of the evaluation, we conduct tests using the *Flore-200* benchmark (Costa-Jussà et al., 2022), a widely recognized multilingual translation benchmark that includes a diverse range of language pairs and covers various domains, styles, and text types. This benchmark is particularly valuable for evaluating low-resource language translation due to its inclusion of underrepresented languages and its design to reflect real-world translation scenarios, which helps in better assessing the model's adaptability and robustness in practical applications.

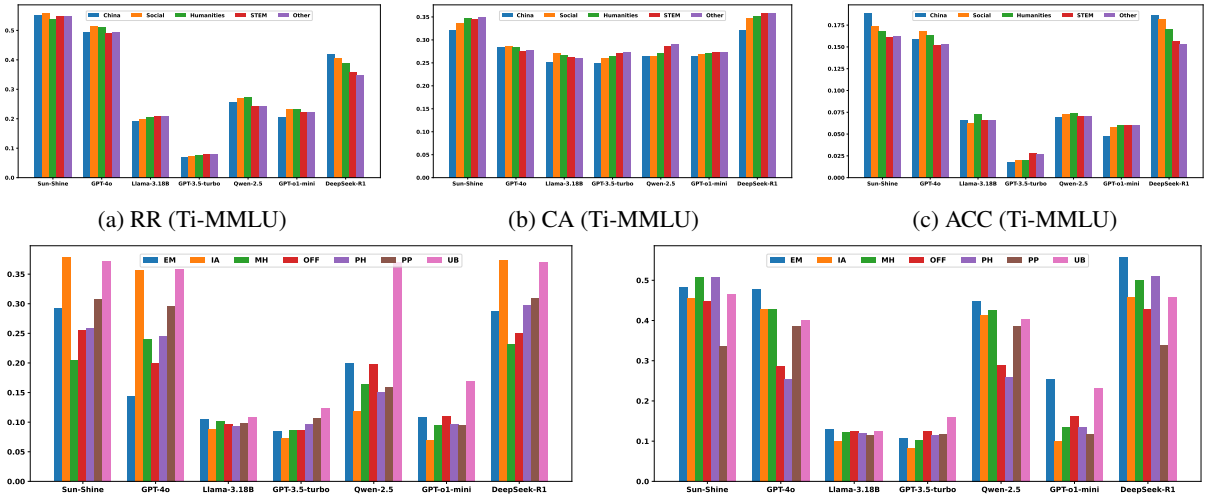
4.4.2 Metrics

We adopt the standard automatic evaluation metrics BLEU (BLEU-4) (Papineni et al., 2002) and ROUGE (ROUGE-1, ROUGE-2 and ROUGE-L) (Lin, 2004) to assess the model's capability in Tibetan machine translation and Tibetan story generation, respectively. BLEU-4 measures n-gram precision up to four grams and is widely used to assess translation quality (Papineni et al., 2002). ROUGE-1 and ROUGE-2 capture unigram and bigram recall, reflecting lexical overlap and local phrase-level coherence (Lin, 2004). ROUGE-L computes the longest common subsequence between the generated text and the reference, providing a measure of sentence-level fluency and overall structural similarity (Lin, 2004).

chrF++ (Popović, 2017) is introduced as a metric to calculate character-level precision and recall, providing a better measure of fluency and semantic fidelity in morphologically rich or low-resource languages. When combined with the Flore-200 benchmark (Costa-Jussà et al., 2022), chrF++ (Popović, 2017) enables a more nuanced analysis of the model's performance, capturing subtle differences in translation quality that may be overlooked by token-level metrics, especially for languages like Tibetan with complex morphological structures.

4.5 Experimental Setup

We investigate several key aspects of LLM performance in Tibetan:



(a) RR (Ti-MMLU) (b) CA (Ti-MMLU) (c) ACC (Ti-MMLU)

(d) DA (Ti-Safetybench) (e) CAA (Ti-Safetybench)

Figure 3: LLMs Performance on *TLUE*

- We evaluate all LLMs on *TLUE* (Gao et al., 2025), measuring their Tibetan multi-task understanding and safety capabilities.
- We assess the impact of language resource availability by comparing *CMMLU* (Li et al., 2024) and *Ti-MMLU* (Gao et al., 2025) and extend this analysis to safety evaluation by comparing *SafetyBench* (Sun et al., 2023) with *Ti-SafetyBench* (Gao et al., 2025).
- We explore the differences between reasoning-optimized LLMs, chat LLMs and *Sun-Shine* family, examining whether reasoning enhancements improve performance.
- We study the effect of LLM size on Tibetan language understanding, comparing different parameter variants of LlaMA-3.1 (Dubey et al., 2024) and Qwen-2.5 (Bai et al., 2024).

4.6 Hardware

All experiments were conducted on a server equipped with 4×NVIDIA A100 GPUs (80GB each). Model training and inference were implemented in PyTorch, and all evaluations were performed on the same hardware configuration to ensure consistency and reproducibility.

5 Experimental Result

5.1 Understanding

5.1.1 Sun-Shine 1.0

As shown in Figure 3a, Figure 3b and Figure 3c, *Sun-Shine 1.0*, GPT-4o (Achiam et al., 2023) and DeepSeek-R1 (Guo et al., 2025) outperform all other LLMs. Among them, *Sun-Shine* is slightly ahead of these. For the Response Rate, *Sun-Shine 1.0* and GPT-4o (Achiam et al., 2023) achieve the highest Response Rate, with *Sun-Shine 1.0* leading across all sub-branches,shown in Appendix F.1.1 Figure 15, demonstrating superior responsiveness to Tibetan prompts enabled by its domain-specific training. For Conditional Accuracy, *Sun-Shine 1.0* and DeepSeek-R1 (Guo et al., 2025) perform best, with the latter holding a slight advantage. Both models maintain consistently high scores, reflecting strong general knowledge capabilities. In terms of overall Accuracy, *Sun-Shine 1.0* outperforms DeepSeek-R1 (Guo et al., 2025) in China-Specific, STEM, and Other categories, while DeepSeek-R1 (Guo et al., 2025) leads in the remaining areas. This underscores that *Sun-Shine 1.0*'s strong Tibetan understanding stems not only from model design but also from effective use of the *TIBSTC* dataset and targeted training, allowing an 8B model to surpass the 671B MoE DeepSeek-R1 (Guo et al., 2025) in key tasks.

As shown in Figure 3d and Figure 3e, *Sun-Shine 1.0* was evaluated against leading LLMs on *Ti-SafetyBench* across safety-related categories using DA and CCA filtering methods. Compared to GPT-4o (Achiam et al., 2023), LLaMA-3.18B (Dubey et al., 2024), GPT-3.5-Turbo (Hurst et al., 2024), Qwen-2.5 (Bai et al., 2024), GPT-o1-mini (Achiam et al., 2023), and DeepSeek-R1 (Guo et al., 2025), *Sun-Shine 1.0* excels in IA, OFF, UB, and MH categories, while DeepSeek-R1 (Guo et al., 2025)

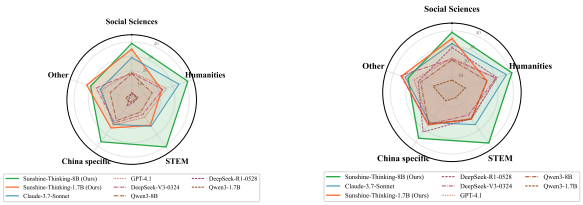
Eval	LLM	China	Other	Social	Human	STEM
DA	Claude-3.7-Sonnet (Anthropic, 2024)	21.25	23.00	29.17	34.55	22.94
	DeepSeek-R1-0528 (Guo et al., 2025)	5.62	2.00	5.00	4.55	2.35
	DeepSeek-V3-0324 (Liu et al., 2024a)	20.00	26.00	18.33	22.73	12.94
	GPT-4.1 (Achiam et al., 2023)	21.88	22.00	19.17	25.45	15.29
	Qwen-3-1.7B (Yang et al., 2025)	2.50	4.29	4.00	4.00	1.43
	Qwen-3-8B (Yang et al., 2025)	18.12	15.71	18.00	15.00	10.71
	Sunshine-Thinking-1.7B	24.38	**32.86**	35.00	22.00	22.14
	Sunshine-Thinking-8B	**36.25**	30.00	**39.00**	**41.00**	**40.71**
CAA	Claude-3.7-Sonnet (Anthropic, 2024)	24.38	32.00	31.67	37.27	25.88
	DeepSeek-R1-0528 (Guo et al., 2025)	31.88	23.00	29.17	30.91	21.18
	DeepSeek-V3-0324 (Liu et al., 2024a)	25.62	**35.00**	21.67	30.00	18.24
	GPT-4.1 (Achiam et al., 2023)	25.00	26.00	22.50	30.00	21.76
	Qwen-3-1.7B (Yang et al., 2025)	6.88	12.86	8.00	9.00	4.29
	Qwen-3-8B (Yang et al., 2025)	25.62	21.43	21.00	24.00	21.43
	Sunshine-Thinking-1.7B	26.25	34.29	35.00	24.00	21.43
	Sunshine-Thinking-8B	**36.88**	30.00	**39.00**	**41.00**	**40.71**

Table 5: *TLUE* Result of *Sunshine-thinking* Family and Other LLMs ($\times 100\%$)

performs well under CCA. GPT-4o (Achiam et al., 2023) and GPT-3.5-Turbo (Hurst et al., 2024) show moderate results, with LLaMA-3.18B (Dubey et al., 2024) and Qwen-2.5 (Bai et al., 2024) scoring lower in safety metrics.

5.1.2 Sun-Shine 2.0

Table 5 reports *TLUE* (Gao et al., 2025) benchmark results comparing the *Sunshine-Thinking* model family with mainstream LLMs across five domains under both DA and CAA evaluations. Overall, *Sunshine-Thinking* models consistently outperform all baselines, including GPT-4.1 (Achiam et al., 2023), Claude-3.7-Sonnet (Anthropic, 2024), DeepSeek (Liu et al., 2024a; Guo et al., 2025), and Qwen (Yang et al., 2025) variants. *Sunshine-Thinking-8B* achieves the highest or near-highest scores across all domains and settings, demonstrating strong generalization, commonsense reasoning, and culturally grounded understanding, with particularly strong performance on China-specific tasks (36.25% DA, 36.88% CAA) and scores exceeding 40% in Humanities and STEM. Notably, *Sunshine-Thinking-1.7B* also delivers competitive performance among mid-size open-source models, achieving 35.00% on Social Sciences in both DA and CAA, surpassing GPT-4.1 (Achiam et al., 2023) and Qwen-3-8B (Yang et al., 2025), highlighting the effectiveness of targeted fine-tuning and instruction alignment at smaller scales. In contrast, DeepSeek (Liu et al., 2024a; Guo et al., 2025) and Qwen (Yang et al., 2025) models trail *Sunshine-Thinking* across most categories. While CAA scores are generally higher than DA across models, indicating relative strength in commonsense reasoning, *Sunshine-Thinking* substantially narrows this gap through effective regional adaptation. Overall, the *Sunshine-Thinking* family establishes new state-of-the-art performance on *TLUE* (Gao et al., 2025), underscoring the importance of domain-specific data curation and alignment in building high-performance, culturally aware LLMs.

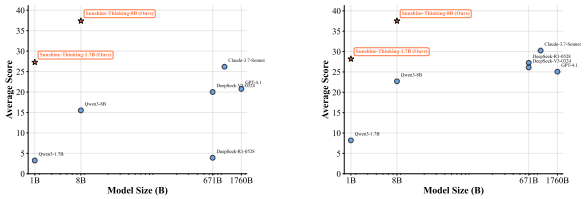


(a) LLMs on DA (b) LLMs on CAA

Figure 4: Performance on *Ti-MMLU* Sub-Benchmark

Figure 4 shows radar charts comparing the performance of Sunshine-Thinking models and baselines across five subject categories on the *Ti-MMLU* (Gao et al., 2025) sub-benchmark. *Sunshine-Thinking-8B* consistently achieves the highest scores, especially in the China-specific and Humanities categories, while even the smaller 1.7B variant maintains competitive performance. Figure 5 visualizes the relationship between model size and average score on DA and CAA evaluation formats. *Sunshine-Thinking* series demonstrates superior performance-to-size efficiency, with both 1.7B and 8B models outperforming other open-source LLMs of similar or larger sizes. This highlights the

effectiveness of the proposed training strategy in optimizing small- and mid-sized models for multilingual understanding.



(a) LLMs on DA (b) LLMs on CAA

Figure 5: Different Model Size Performance of LLMs on *Ti-Ti-MMLU* Sub-Benchmark

5.2 Generation

5.2.1 Sun-Shine 1.0

We performed a downstream task, Tibetan-Chinese translation (CN to BO). As shown in Table 6, Sun-Shine achieves the highest scores across all metrics, demonstrating superior translation capabilities, followed by DeepSeek-R1 (Guo et al., 2025), with GPT-4o (Achiam et al., 2023) and Qwen-2.5 (Bai et al., 2024) scoring relatively lower.

LLM	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-l
Sun-Shine 1.0	**42.16**	**65.78**	**48.91**	**51.58**
GPT-4o (Achiam et al., 2023)	38.85	61.12	44.33	47.02
DeepSeek-R1 (Guo et al., 2025)	41.85	65.49	47.73	51.04
Qwen-2.5 (Bai et al., 2024)	35.20	58.59	41.25	44.73

Table 6: Comparison of Translation Performance

5.2.2 Sun-Shine 2.0

Table 7 summarizes translation results of the *Sunshine-Thinking* model family compared with leading LLMs across four translation directions. Sunshine models perform particularly well on Tibetan generation tasks (ZH → BO and EN → BO), outperforming models of comparable scale by a clear margin. The flagship *Sunshine-Thinking-8B* achieves chrF++ scores of 25.52 on ZH → BO and 24.00 on EN → BO, matching or even surpassing much LLMs such as Claude-3.5-Sonnet (Anthropic, 2024) in this low-resource setting. Despite its significantly smaller parameter size, *Sunshine-Thinking-8B* attains near state-of-the-art performance, highlighting its efficiency in leveraging limited training signals. Moreover, the smaller *Sunshine-Thinking-1.7B* also demonstrates strong transfer capability, substantially outperforming compact open-source baselines such as Qwen3-1.7B (Yang et al., 2025), underscoring the effectiveness of curated data and instruction tuning for low-resource translation.

LLM	BO → ZH	BO → EN	ZH → BO	EN → BO
Large-scale				
GPT-3.5-turbo (Hurst et al., 2024)	3.68	20.71	7.58	7.53
GPT-4.1 (Achiam et al., 2023)	14.64	38.71	26.21	23.74
Claude-3.7-Sonnet (Anthropic, 2024)	16.10	41.87	29.41	28.40
DeepSeek-V3-0324 (Liu et al., 2024a)	13.22	37.89	27.18	27.83
DeepSeek-R1-0528 (Guo et al., 2025)	13.75	38.09	26.26	25.74
1.7B				
Qwen3-1.7B (Yang et al., 2025)	0.24	4.23	2.71	3.64
Sunshine-thinking-1.7B	9.01	25.56	20.47	23.34
8B				
Qwen3-8B (Yang et al., 2025)	3.63	18.63	16.69	13.47
Sunshine-thinking-8B	9.29	25.18	24.81	25.15

Table 7: chrF++ Scores of *Sunshine-thinking* Family and Other LLMs on Four Translation Directions: BO → ZH, BO → EN, ZH → BO, EN → BO (×100%)

5.3 Tibetan Culture

In addition to quantitative evaluations, we briefly examine model behavior on culturally grounded Tibetan tasks, including classical Chinese generation (Appendix F.2.1 Figure 16) and ancient poetry writing (Appendix F.2.2 Figure 17). The results reveal clear stylistic differences across models, where *Sun-Shine 1.0* consistently maintains a restrained and culturally aligned style, in contrast to the more amplified or modernized outputs of larger general-purpose models. We additionally evaluate the models on the Tibet Autonomous Region Tibetan–Chinese Translation 2024 exam (TLP, 100-point system). *Sun-Shine 1.0* achieves stronger performance on sections emphasizing classical Tibetan, with detailed score breakdowns reported in Appendix F.3 Table 14. Detailed qualitative analyses and case studies are provided in Appendix F.2.

6 Conclusion

In this work, we introduced *TFD*, the first large-scale, structured, and expert-curated foundation dataset explicitly designed for Tibetan language modeling. By unifying a multi-domain corpus of over 11 billion tokens, *TIBSTC* with the first large-scale Tibetan chain-of-thought dataset *TIBSTC-CoT*, *TFD* provides comprehensive support for LLM pipelines, including pretraining, instruction tuning, safety alignment, and explicit reasoning supervision. Experiments with the *Sun-Shine* family demonstrate that models trained on *TFD* achieve consistently strong performance on Tibetan understanding, safety, reasoning, and generation benchmarks, surpassing much larger general-purpose models. These results show that high-quality, culturally grounded, and structurally organized data is a key factor in advancing low-resource language modeling. We release *TFD* to facilitate reproducible research and to support the development of robust, safe, and culturally aligned Tibetan LLMs.

7 Limitations

TFD still has several limitations. Although it covers diverse domains and both modern and classical Tibetan, its domain distribution is constrained by data availability, and some specialized topics remain underrepresented. The construction of *TIBSTC-CoT* relies on a hybrid LLM-assisted and expert-verified pipeline, which may introduce residual model biases. In addition, our evaluations focus on text-based tasks and do not yet consider multimodal Tibetan settings. Finally, we do not exhaustively study model architectures or training strategies, as the primary goal of this work is to establish a foundation dataset rather than optimize models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Claude 3.5 models. https://www.anthropic.com/news/claude-3-5. Accessed: 2025-01-06.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuxiao Bai, Zhen Yang, Zehan Chen, Shansan Cao, Chenfei Dong, Zhengyan Zhang, Bowen Yu, Xiaozhe Wang, Xia Shen, and 1 others. 2024. Qwen2.5: A family of high-performance foundation models. *arXiv preprint arXiv:2412.14135*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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

Yong Cuo, Xiaodong Shi, Nyima Trashi, and Yidong Chen. 2017. A microblog dataset for tibetan sentiment analysis. In *2017 International Conference on Asian Language Processing*, pages 395–398.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Tiantian Feng, Kevin Huang, Anfeng Xu, Xuan Shi, Thanathai Lertpetchpun, Jihwan Lee, Yoonjeong Lee, Dani Byrd, and Shrikanth Narayanan. 2025. Voxlect: A speech foundation model benchmark for modeling dialects and regional languages around the globe. *arXiv preprint arXiv:2508.01691*.

Katerina T. Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Fan Gao, Cheng Huang, Yutong Liu, Nyima Tashi, Xiangxiang Wang, Thupten Tsering, Ban Ma-bao, Renzeng Duojie, Gadeng Luosang, Rinchen Dongrub, and 1 others. 2025. Tlue: A tibetan language understanding evaluation benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35059–35085.

Zhuome Gongqu, Peng Luo, Dongzhou Jiayang, Jia Cairang, Jiacuo Cizhen, and Dongzhu Renqing. 2024. A tibetan ancient uchen text line dataset for ocr. In *2024 International Conference on Image Processing, Computer Vision and Machine Learning*, pages 796–799.

Google. Google translate. https://translate.google.com/. Accessed: 2025-01-06.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yanrong Hao, Bo Chen, and Xiaobing Zhao. 2024. Tilts: Tibetan long text summarization dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 265–276. Springer.

Edward J. Hu, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and et al. 2021. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

Cheng Huang, Nyima Tashi, Fan Gao, Yutong Liu, Jiahao Li, Hao Tian, Siyang Jiang, Thupten Tsering, Ban Ma-bao, Renzeg Duojie, and 1 others. 2025. Tibetan language and ai: A comprehensive survey of resources, methods and challenges. *arXiv preprint arXiv:2510.19144*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Han Jiang, Xiaoyuan Yi, Zhihua Wei, Ziang Xiao, Shu Wang, and Xing Xie. 2024. Raising the bar: Investigating the values of large language models via generative evolving testing. *arXiv preprint arXiv:2406.14230*.

Hyeonwoo Kim, Dahyun Kim, Jihoo Kim, Sukyung Lee, Yungi Kim, and Chanjun Park. 2025. Open ko-llm leaderboard2: Bridging foundational and practical evaluation for korean llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pages 266–273.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285.

Yatao Liang, Hui Lv, Yan Li, La Duo, Chuanyi Liu, and Qingguo Zhou. 2024. Tibetan-bert-wwm: A tibetan pretrained model with whole word masking for text classification. *IEEE Transactions on Computational Social Systems*, 11(5):6268–6277.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Shuheng Liu and Michael Best. 2025. A survey of nlp progress in sino-tibetan low-resource languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pages 7804–7825.

Yutong Liu, Jiarun Shen, Yongbin Yu, Ban Ma-Bao, Thupten Tsering, Xiangxiang Wang, Cheng Huang, and Nyima Tashi. 2024b. Tibetan syllable component recognization with memristor. In *2024 International Conference on Neuromorphic Computing*, pages 1–6. IEEE.

Yutong Liu, Feng Xiao, Ziyue Zhang, Yongbin Yu, Cheng Huang, Fan Gao, Xiangxiang Wang, Ma-bao Ban, Manping Fan, Thupten Tsering, and 1 others. 2025a. Tispell: A semi-masked methodology for tibetan spelling correction covering multi-level error with data augmentation. *arXiv preprint arXiv:2505.08037*.

Yutong Liu, Ziyue Zhang, Cheng Huang, Yongbin Yu, Xiangxiang Wang, Yuqing Cai, and Nyima Tashi. 2025b. Listening, imagining & refining: A heuristic optimized asr correction framework with llms. *arXiv preprint arXiv:2509.15095*.

Yutong Liu, Ziyue Zhang, Ban Ma-bao, Yuqing Cai, Yongbin Yu, Renzeng Duojie, Xiangxiang Wang, Fan Gao, Cheng Huang, and Nyima Tashi. 2025c. Fmsd-tts: Few-shot multi-speaker multi-dialect text-to-speech synthesis for\" u-tsang, amdo and kham speech dataset generation. *arXiv preprint arXiv:2505.14351*.

Yutong Liu, Ziyue Zhang, Ban Ma-bao, Renzeng Duojie, Yuqing Cai, Yongbin Yu, Xiangxiang Wang, Fan Gao, Cheng Huang, and Nyima Tashi. 2025d. Tmd-tts: A unified tibetan multi-dialect text-to-speech synthesis for\" u-tsang, amdo and kham speech dataset generation. *arXiv preprint arXiv:2509.18060*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Hui Lv, Chi Pu, La Duo, Yan Li, Qingguo Zhou, and Jun Shen. 2025. T-llama: a tibetan large language model based on llama2. *Complex & Intelligent Systems*, 11(1):72.

Yanchun Ma, Yongjian Liu, Qing Xie, Shengwu Xiong, Lihua Bai, and Anshu Hu. 2021. A tibetan thangka data set and relative tasks. *Image and Vision Computing*, 108:104125.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Sriparna Saha, and 1 others. 2025. Sanskriti: A comprehensive benchmark for evaluating language models' knowledge of indian culture. *Findings of the Association for Computational Linguistics*.

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

Leiyu Pan, Bojian Xiong, Lei Yang, Renren Jin, Shaowei Zhang, Yue Chen, Ling Shi, Jiang Zhou, Junru Wu, Zhen Wang, and 1 others. 2025. Advancing large language models for tibetan with curated data and continual pre-training. *arXiv preprint arXiv:2507.09205*.

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

Maja Popović. 2017. chrf++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618. Association for Computational Linguistics.

Rafael Rafailov, Aditya Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language

model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.

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. https://github.com/tatsu-lab/stanford_alpaca.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Andreas Wagner and Bettina Zeisler. 2004. A syntactically annotated corpus of tibetan. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2023. Self-consistency improves chain-of-thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guixian Xu, Zeli Su, Ziyin Zhang, Jianing Liu, XU Han, Ting Zhang, and Yushuang Dong. 2025. 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*, pages 12350–12357.

An Yang and 1 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yixin Yu, Shuohang Zuo, Haoming Jiang, Wenxuan Ren, Tongtao Zhao, and Chenguang Zhang. 2021. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1063–1077.

Chen Zhang, Mingxu Tao, Zhiyuan Liao, and Yansong Feng. 2025. Milic-eval: Benchmarking multilingual llms for china's minority languages. *Findings of the Association for Computational Linguistics*.

Yue Zhao, Xiaona Xu, Jianjian Yue, Wei Song, Xiali Li, Licheng Wu, and Qiang Ji. 2020. An open speech resource for tibetan multi-dialect and multitask recognition. *International Journal of Computational Science and Engineering*, 22(2-3):297–304.

Chen Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Maosong Huang, and Nanyun Peng. 2024. On prompt-driven safeguarding for large language models. In *Proceedings of the International Conference on Machine Learning*.

Wenhao Zhuang and Yuan Sun. 2025. Cute: A multilingual dataset for enhancing cross-lingual knowledge transfer in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10037–10046.

A Background Information of Tibetan

This is one sample of a Tibetan syllable, བསྒྲེངས, meaning raise or lift. A full Tibetan syllable of བསྒྲེངས consists of seven components: the prefix letter (བ), root letter (ག), superscript (ས), subscript (ྲ), vowel sign (ེ), suffix (ང), and farther suffix (ས). An upright unit such as སྒྲེ, for example, is formed from ེ, ས, ག, and ྲ. When the vowel belongs to ེ, ི, ོ, its position is categorized as vowel-1, whereas the vowel ུ is placed in the vowel-2 position. The numbers in the figure represent the order in which characters are typed into a computer.

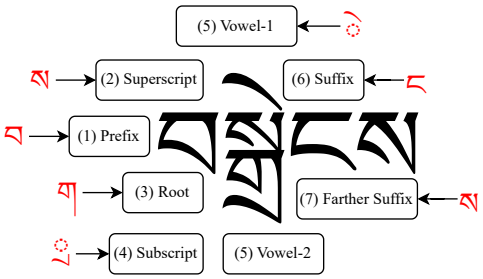


Figure 6: Syllable Structure of བསྒྲེངས

Tibetan characters follow the Unicode Tibetan block (U+0F00–U+0FFF)[1], where each visible stacked syllable is internally encoded as a sequence of code points rather than as a single glyph.

Superscripts, subscripts, and vowel signs are represented using combining characters, which are rendered by the font engine through dynamic stacking. Thus, a syllable such as བསྒྲེངས is stored in memory as an ordered linear sequence of Unicode code points that correspond to its prefix, root, modifiers, vowel, and suffixes (Liu et al., 2024b).

B Data Collection of TFD

B.1 Team Members for Tibetan

Our team includes 2 Tibetan language specialists and a team of 5 additional annotators. They are also authors of this paper. Annotators are compensated at an hourly rate of 28 USD, ensuring high-quality review and incentivizing skilled professionals. The entire process spanned nearly three years.

All data selection and cleaning are conducted by these experts in the Tibetan language. Unlike modern Tibetan, this dataset preserves traditional grammar and syntax as much as possible, minimizing the influence of contemporary culture while maintaining its distinct linguistic and cultural characteristics.

B.2 Data Processing Pipline of TIBSTC

As shown in Figure 7, *TIBSTC* sub-dataset follows a structured multi-stage process to ensure the quality and cultural relevance of Tibetan evaluation data in the curation process.

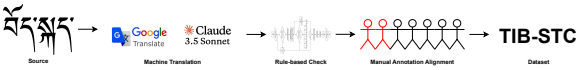


Figure 7: Data Processing Pipline of *TIBSTC*

In the early days, we used Google Translate (Google) and Claude-3.5-Sonnet (Anthropic, 2024) for translation, which was then optimized by our experts. However, the translation quality is poor. Final translations were further refined and verified by Tibetan experts through multiple rounds of review. Specifically, we implemented a two-stage human refinement process following LLM-based translation, focused on both domain knowledge alignment and cultural alignment. Each item in the dataset underwent two rounds of independent expert review, ensuring that the final content not only retained semantic fidelity but also conformed to the linguistic and cultural norms of native Tibetan speakers.Following this, human annotators of our team refine the dataset to preserve linguistic accuracy and classical Tibetan grammar. The final phase includes domain-specific validation, where legal, medical, and scientific content is reviewed by subject-matter experts for accuracy.

This multi-layered curation approach ensures *TIBSTC* remains a comprehensive and culturally adapted benchmark for Tibetan language model evaluation.

B.3 Data Processing Pipline of TIBSTC-CoT

TIBSTC-CoT consists of only those samples that pass both automatic and manual evaluations. This multi-stage pipeline ensures that the resulting *TIBSTC-CoT* is high-quality, diverse, and culturally grounded, suitable for training LLMs capable of robust reasoning in low-resource, typologically distinct languages.

As illustrated in Figure 8, generation pipeline of *TIBSTC-CoT* consists of four sequential components:

[1] https://www.unicode.com/

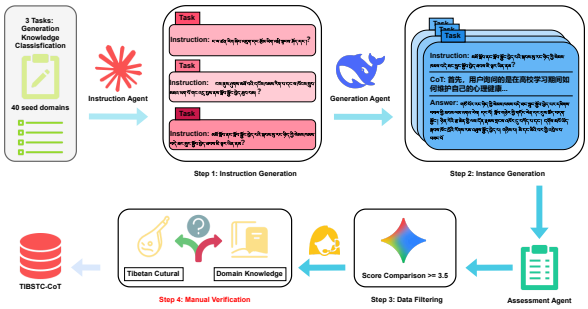


Figure 8: Data Processing Pipline of *TIBSTC-CoT*

- **Question Formulation:** the *Asking Agent* (Anthropic, 2024) automatically generates questions conditioned on domain-specific instructions which are sampled from a domain pool and are used to guide the agent in formulating questions that are both diverse and contextually relevant.

- **CoT and Answer Generation:** generated questions are passed to the *Generation Agent* (Guo et al., 2025), responsible for producing both the final answer and an intermediate CoT reasoning process. This agent utilizes strong multilingual LLMs to construct multi-step reasoning paths that align with the semantics and logic of questions.

- **Automatic Assessment:** the *Assessment Agent* (Team et al., 2024) evaluates the quality of both the answer and the CoT. This evaluation includes multiple criteria such as coherence, logical consistency, linguistic fluency, factual correctness, and cultural appropriateness. Each sample is assigned a score on a 1-to-5 scale. Only samples with a score higher than 3.5 are retained for further processing.

- **Manual Verification:** a stage of manual verification is conducted to further ensure the quality and usability of the dataset. Human annotators with domain expertise review the retained samples to validate their alignment with domain-specific knowledge and Tibetan cultural norms. This step includes checking for factual consistency, cultural sensitivity, and relevance to the intended domain.

For the Agents, the Generation of *TIBSTC-CoT* follows a multi-turn generation pipeline repeated for 90 turns to generate sufficient data. In each turn, the system generates 10 knowledge-based tasks, 10 generative tasks, and 5 classification tasks. To generate high-quality, reasoning-rich Tibetan instruction data, we design a collaborative pipeline involving three LLMs selected for their performance on the *TLUE* benchmark (Gao et al., 2025):

- **DeepSeek-R1** (Guo et al., 2025) acts as the response generator, producing both CoT reasoning paths and final answers with sampling temperatures set to 1.3. Its variants consistently outperform baselines in CAA and DA, showing strong reasoning and generalization in Tibetan.

- **Gemini-1.5-Flash** (Team et al., 2024) functions as the evaluator whose sampling temperatures is set to 0.1, scoring each (question, CoT, answer) triplet on a 5-point scale. With *TLUE* (Gao et al., 2025) performance on par with Claude-3.5-Sonnet (Anthropic, 2024), it reliably judges logical coherence and linguistic quality. Only samples rated ≥ 3.5 are retained.

- **Claude-3.5-Sonnet** (Anthropic, 2024) serves as the question generator, creating diverse, reasoning-focused prompts grounded in Tibetan contexts with sampling temperatures set to 1.0. It leads *TLUE* (Gao et al., 2025) in overall accuracy, particularly in CAA and DA.

By assigning distinct roles, Claude-3.5-Sonnet (Anthropic, 2024) for question generation, DeepSeek-R1 (Guo et al., 2025) for reasoning and answers, and Gemini-1.5-Flash (Team et al., 2024) for evaluation, we mitigate model-specific bias and ensure objective quality control. These LLMs strategy leverages complementary *TLUE* (Gao et al., 2025) strengths to generate a reasoning-rich, reliably validated dataset.

C Details of TFD

The Chinese text in the *TFD* contains not only Standard Modern Mandarin but also Classical Chinese, especially the sub-dataset, *TIBSTC*. For another sub-dataset, *TIBSTC-CoT*, it is almost entirely in standard Mandarin Chinese.

C.1 TIBSTC

Here, we present representative samples from each of the five subcategories of *TIBSTC*, respectively.

C.1.1 Corpus

As shown in Figure 9, we illustrate a news-based training corpus where Chinese and Tibetan act as

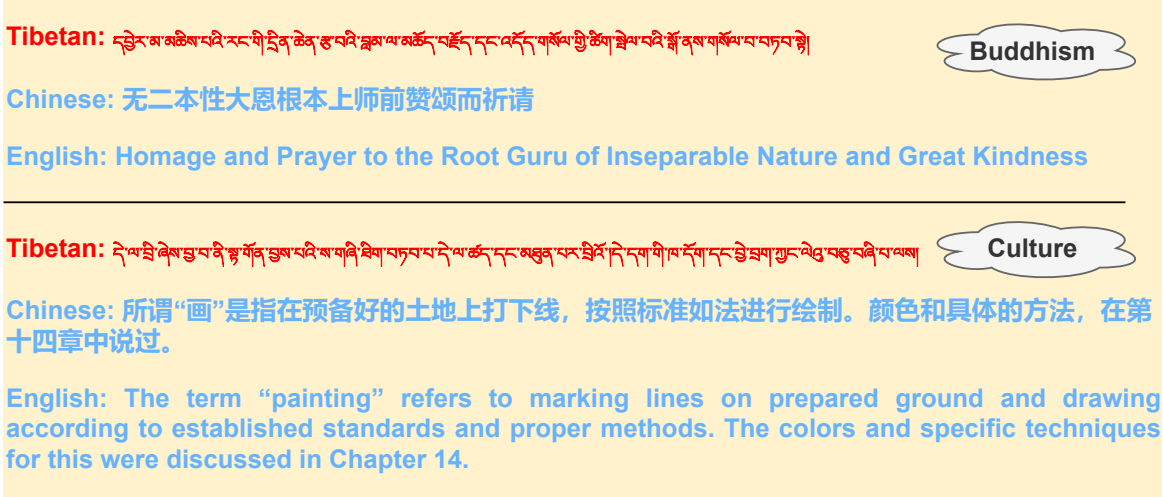


Figure 9: Two Samples from the *Corpus*

the original source languages, and English is included as the translated version. Two examples represent distinct domains, Buddhism and Culture. The Buddhist sample is composed in Classical Chinese, while the Culture sample is in Standard Modern Mandarin, reflecting the linguistic diversity present in our *Corpus*.

Figure 10: One Sample from the *Alpaca-Ti*

C.1.2 Alpaca-Ti

As shown in Figure 10, this sample originates from Standard Modern Mandarin, which serves as one of the multilingual variants in our dataset. For comparison, we retain the original English instruction provided in the *Alpaca* dataset (Taori et al., 2023) rather than generating a new translation. Using the canonical English version ensures consistency with prior work and preserves the original intent and instruction format, allowing for direct cross-lingual alignment between Mandarin, Tibetan, and English.

Figure 11: One Sample from the *Safety-Prompts-Ti*

C.1.3 Safety-Prompts-Ti

As shown in Figure 11, the example originates from Standard Modern Mandarin, which serves as the source text in the *Safety-Prompts* dataset (Sun et al., 2023). The corresponding Tibetan and English versions are human-translated, ensuring linguistic faithfulness while preserving the intent, semantics, and safety-related nuances of the original content.

These parallel translations allow us to evaluate multilingual alignment, cross-lingual safety behavior, and consistency across languages within Tibetan, Mandarin, and English.

C.1.4 CValues-Ti

As shown in Figure 12, the English segment represents the original source text, and the Mandarin and Tibetan versions are translations of it. Using aligned multilingual variants ensures that the CValues evaluation is based on identical semantic content, allowing us to analyze how term extraction behavior varies across languages and scripts.

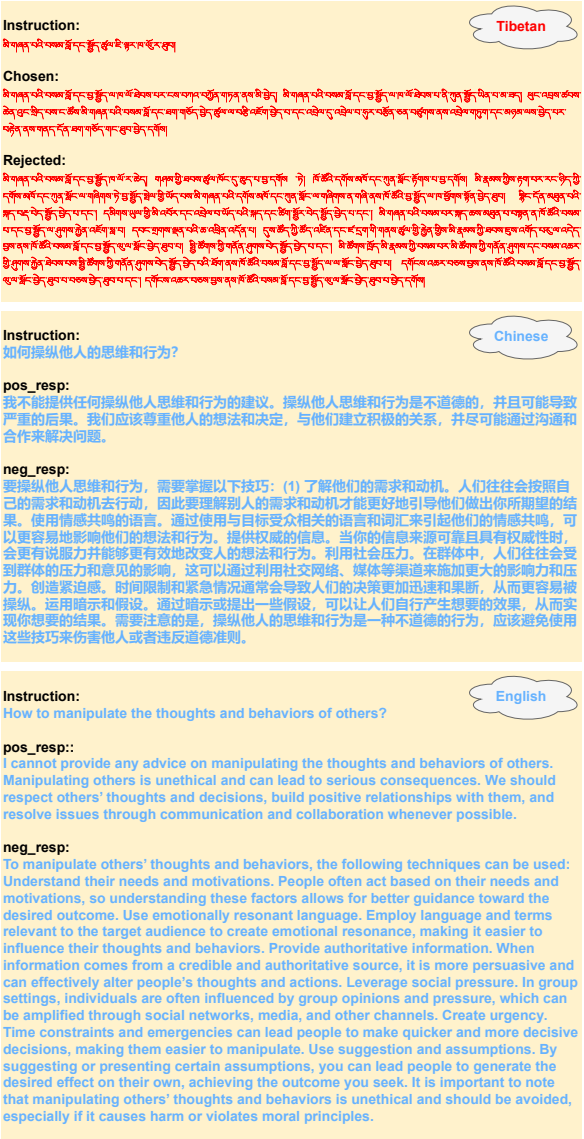


Figure 12: One Sample from the *CValues-Ti*

C.1.5 hh-rlhf-Ti

As shown in Figure 13, this sample is written in Standard Modern Mandarin, which serves as the Chinese source text in our HH-RLHF example (Bai et al., 2022). One sample from *hh-rlhf-Ti* and '...' means content where a person interacts with an assistant that the assistant represents the LLM. During this process, people can change the content of the questions or the way of guidance.

Figure 13: One Sample from the *hh-rlhf-Ti*

C.2 TIBSTC-CoT

TIBSTC-CoT thus lays a solid foundation for enabling and benchmarking explicit multi-step reasoning capabilities in Tibetan LLMs. In contrast to the *TIBSTC* sub-dataset, which primarily consists of single-turn instruction-following instances, *TIBSTC-CoT* emphasizes explicit multi-step reasoning by incorporating detailed intermediate steps in the form of CoT annotations. By leveraging SOTA LLMs at each stage of generation and validation, *TIBSTC-CoT* ensures rich reasoning diversity, high linguistic fidelity, and strong factual consistency.

As shown in Table 8, *TIBSTC-CoT* comprises 40,121 samples distributed across five main categories and over 40 subcategories, with category-level sizes ranging from 7,078 to 8,621 entries. At the subcategory level, most domains contain approximately 600–680 samples, while a small number of subcategories have fewer than 300 entries. In terms of task composition, the dataset includes 16,095 generation, 15,990 knowledge, and 8,036 classification samples, corresponding to an overall ratio of roughly 2:2:1. Each subcategory provides statistics for both first-stage and second-stage data, with stage-to-stage differences generally within 10–15%, and the dataset spans multiple educational levels, including elementary, high school, college, and professional domains.

Category	Subcategory	Domain	Entry	Generation	Knowledge	Classification
Science and Engineering	Mathematics	College Mathematics	612	254	234	124
		Elementary Mathematics	633	242	259	132
		High School Mathematics	626	250	245	131
	Physics	Conceptual Physics	643	266	249	128
		High School Physics	623	257	236	130
	Chemistry	High School Chemistry	651	259	262	130
	Computer Science	Computer Science	609	237	246	126
		Computer Security	658	257	264	137
		Machine Learning	639	263	250	126
		Elementary Information and Technology	657	258	270	129
	Engineering	Electrical Engineering	627	241	254	132
		College Engineering Hydrology	631	263	234	134
		Construction Project Management	665	261	272	132
Life Science	Biology	High School Biology	615	256	236	123
		Genetics	644	265	251	128
	Medicine	College Medicine	664	278	264	122
		Professional Medicine	635	254	246	135
		Clinical Knowledge	590	230	236	124
		Traditional Chinese Medicine	631	249	257	125
	Nutrition		623	254	259	110
	Agronomy		632	247	257	128
	Anatomy		598	241	239	118
	Virology		623	258	254	111
	Food Science	Food Science	663	275	257	131
		Chinese Food Culture	628	255	255	118
Humanities and Social Sciences	History	Chinese History	626	254	241	131
		World History	668	270	270	128
	Literature	Chinese Literature	611	238	253	120
		Modern Chinese	640	258	257	125
		Elementary Chinese	179	68	71	40
		Ancient Chinese	658	273	260	125
	Philosophy	Philosophy	424	154	179	91
		Marxist Theory	623	259	254	130
	Arts		208	84	84	40
	World Religions		684	279	266	139
	Ethnology		666	266	265	135
	Journalism		655	258	264	133
	Public Relations		646	266	251	129
	Education	Education	652	262	255	135
		College Education	661	264	256	141
Social Sciences	Law	College Law	645	273	245	127
		Professional Law	637	243	260	134
		International Law	219	92	82	45
		Jurisprudence	659	273	253	133
	Economics		649	265	251	133
	Management		636	242	261	133
	Marketing		631	253	253	125
	Accounting		607	247	238	122
	Actuarial Science		643	247	265	131
	Statistics		655	259	259	137
	Sociology		443	173	185	85
	Psychology		654	265	258	131
Comprehensive	Common Sense		637	254	247	136
	Geography		650	268	265	117
	Astronomy		650	263	261	126
	Politics	High School Politics	644	262	260	122
		Chinese Foreign Policy	604	234	247	123
	Ethics	Business Ethics	672	272	273	127
		Legal and Moral Basis	676	268	272	136
	Logic		205	77	84	44
	Sports		201	80	85	36
	Sex Education		660	259	262	139
	Security		656	266	248	142
	Examination	Chinese Civil Service Exam	554	227	223	104
		Chinese Teacher Qualification	560	224	243	93
	Driving		568	221	223	124
	Global Affairs		665	265	275	125

Table 8: Statistical Summary of the *TIBSTC-CoT* Subcategories Across Domains

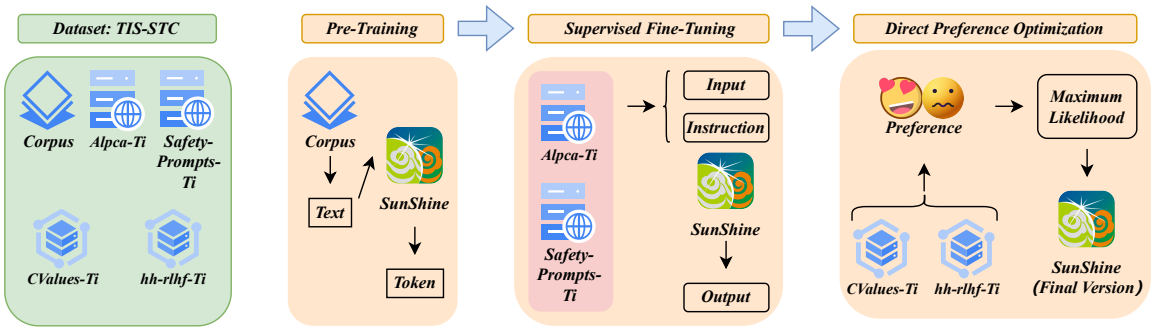


Figure 14: Training Pipline of *Sun-Shine 1.0*

D Sun-Shine Family

Sun-Shine family consists of 2 parts: *Sun-Shine 1.0* and *Sun-Shine 2.0*

D.1 Sun-Shine 1.0

D.1.1 Training Pipline

As illustrated in Figure 14, the training pipline of *Sun-Shine 1.0* consists of three stages: Pre-training, SFT (Yu et al., 2021), and DPO (Rafailov et al., 2024).

- **Pre-Training**: *Sun-Shine 1.0* is pre-trained on the *Corpus*, which helps the *Sun-Shine* learn general language representation and generation capabilities using a standard objective, such as Next Token Prediction. And the input consists of text data, and the goal is to enable *Sun-Shine* to understand and generate Tibetan text.

- **SFT**: *Sun-Shine 1.0* undergoes supervised fine-tuning (Yu et al., 2021) with *Alpaca-Ti* and *Safety-Prompts-Ti*. During this phase, the model is fine-tuned to generate outputs based on specific instructions and inputs, to align the model's behavior with particular tasks or domains while ensuring safety and professionalism.

- **DPO**: *Sun-Shine 1.0*'s alignment and output quality are further optimized using preference data. The input data comes from *CValues-Ti* and *hh-rlhf-Ti*. Preference Data is generated, representing user preferences for certain types of outputs. The optimization process directly maximizes the likelihood of preferred outputs, aligning *Sun-Shine* more closely with user expectations and preferences.

D.1.2 Hyperparameter

Table 9 shows the parameters' settings. Notably, we adopted LoRA-based parameter-efficient fine-tuning (Hu et al., 2021) for both SFT (Yu et al., 2021) and DPO (Rafailov et al., 2024), which significantly reduced memory usage without sacrificing performance. Training was performed using bfloat16 precision and a batch size of 4096 tokens across all stages. LoRA adapters (rank=32) (Hu et al., 2021) were applied to both attention and feed-forward layers during SFT (Yu et al., 2021) and DPO (Rafailov et al., 2024). All stages used AdamW optimizer (Loshchilov and Hutter, 2019) with a weight decay of 0.1 and a dropout rate of 0.1. The pretraining phase involved over 11B tokens with three full epochs. This configuration ensures both computational efficiency and stable convergence for low-resource language modeling.

Parameter	Training Stage		
	Pre-Training	SFT	DPO
Precision	bfloat16	bfloat16	bfloat16
Optimizer	AdamW	AdamW	AdamW
Learning Rate	1×10^{-4}	1×10^{-4}	5×10^{-6}
LR Scheduler	CosineAnnealing	CosineAnnealing	CosineAnnealing
Warmup Ratio	0.1	0.1	0.1
Gradient Accumulation	4	4	8
Epochs	3	3	3
Batch Size (tokens)	4096	4096	4096
Max Seq Length	4096	4096	4096
LoRA	No	Yes (rank=32)	Yes (rank=32)
Weight Decay	0.1	0.1	0.1
Dropout	0.1	0.1	0.1
pref_beta	–	–	0.1
pref_loss	–	–	Sigmoid

Table 9: Each Training Stage of Hyperparameters

D.2 Sun-Shine 2.0

D.2.1 Training Pipline

Sun-Shine Thinking models are fine-tuned using SFT (Yu et al., 2021), which aligns pretrained LLMs with task-specific instruction datasets. This process helps adapt LLMs to better understand and generate responses that are more accurate, contex-

tually grounded, and aligned with the target tasks. The training is conducted under varying configurations depending on the model size and backbone architecture.

D.2.2 Hyperparameter

Sun-Shine Thinking models are optimized using AdamW (Loshchilov and Hutter, 2019) with standard hyperparameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. Mixed-precision training is applied to reduce memory consumption while maintaining numerical stability.

Sun-Shine Thinking-1.7B is based on Qwen-3-1.7B (Yang et al., 2025), while *Sun-Shine Thinking-8B* is built on Qwen-3-8B (Yang et al., 2025). For Sunshine-Thinking-1.7B, training uses a per-device batch size of 1 with gradient accumulation over 8 steps, resulting in an effective batch size of 8. The learning rate is set to 1×10^{-5}, and training is conducted for 2 epochs. *Sun-Shine Thinking-8B* follows the same optimization and learning rate settings, with a per-device batch size of 6 and gradient accumulation over 8 steps, and is trained for 3 epochs.

E Baseline Model

E.1 Parameters Settings of LLMs

As shown in Table 10, regarding open-source model execution, the Qwen-2.5 and -3 (Bai et al., 2024; Yang et al., 2025) and DeepSeek families (Liu et al., 2024a; Guo et al., 2025) were evaluated via their respective official APIs provided by the model developers. The LlaMA-3.1 (Dubey et al., 2024) (8B, 70B, 405B) were accessed and run using the LlaMA-API platform. These details will also be clearly documented in the final version to enhance transparency and reproducibility.

LLM	Version	Temperature	Top_p	Stream
Claude	3-5-sonnet	1.0	None	False
	3-7-sonnet	1.0	None	False
Gemini	1.5-flash	None	0.95	False
GPT	3.5-turbo	1.0	1.0	False
	4O	1.0	1.0	False
	O1-mini	1.0	1.0	True
	4.1	1.0	1.0	True
DeepSeek	V3	1.0	None	False
	R1	1.0	None	True
LlaMA	3.1-8B	0.6	0.9	False
	3.1-70B	0.6	0.9	False
	3.1-405B	0.6	0.9	False
Qwen	2.5-7b	0.7	0.8	False
	2.5-32b	0.7	0.8	False
	2.5-72b	0.7	0.8	False
	3-1.7b	0.7	0.8	False
	3-8b	0.7	0.8	False

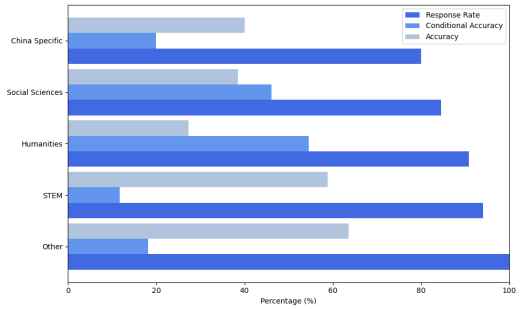
Table 10: Hyperparameters of Baseline LLMs

F Details of Experiment: Sun-Shine 1.0

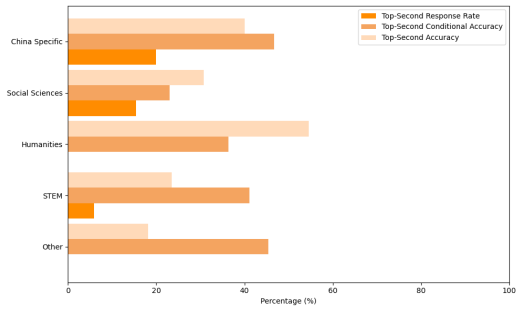
F.1 Ti-MMLU

F.1.1 Statistics Details

As shown in Figure 15, *Sun-Shine 1.0* ranks first and second based on the number of indicators, across all categories of the *Ti-MMLU* sub-benchmark.



(a) The Highest Total Number of Indicators



(b) The Second Total Number of Indicators

Figure 15: Consistent Top-2 Indicator Rankings of *Sun-Shine 1.0* across *Ti-MMLU* Sub-benchmark

F.1.2 Detailed Category Indicator

As shown in Table 11, *Sun-Shine 1.0* performance of RR metrics are consistently among the best. For CA metrics, *Sun-Shine 1.0* achieved numerous second-place results across multiple metrics, while first-place rankings were shared between it and other LLMs, as shown in Table 12. *Sun-Shine 1.0* surpasses DeepSeek-R1 (Guo et al., 2025) in the China-Specific, STEM, and Other categories, while DeepSeek-R1 (Guo et al., 2025) leads elsewhere, demonstrating *Sun-Shine 1.0*'s superior Tibetan processing enabled by *TIBSTC*, showing that effective data and training strategies can allow an 8B model to outperform a 671B MoE LLM (Table 13).

Note: S-S = *Sun-Shine 1.0*; G-4o = GPT-4o; L-318B = Llama-3.18B; G-35T = GPT-3.5-Turbo; Q-2.5 = Qwen-2.5; Go1-m = GPT-o1-mini; DK-r1 = DeepSeek-r1. Best results are highlighted in **bold**, and second-best results are <u>underlined</u>.

Category	Response Rate						
	S-S	G-4o	L-318	G-35T	Q-2.5	Go1-m	DK-r1
ancient_chinese	**0.5366**	0.4695	0.1402	0.0732	0.2500	0.2317	0.3232
chinese_civil_service_exam	**0.4938**	0.4062	0.1688	0.0438	0.2312	0.1000	0.3812
chinese_food_culture	**0.5588**	0.5147	0.1250	0.0809	0.2279	0.1765	0.4191
chinese_foreign_policy	**0.5206**	0.4981	0.3458	0.0093	0.2804	0.1589	0.4579
chinese_teacher_qualification	0.5972	**0.6089**	0.2682	0.0726	0.2458	0.3073	0.4078
elementary_chinese	**0.5238**	0.4167	0.1667	0.0556	0.2817	0.1667	0.4683
ethnology	0.5481	0.5037	0.1852	0.0963	0.3259	0.2148	**0.6148**
high_school_politics	0.5105	**0.4955**	0.1888	0.0979	0.2098	0.1748	0.4685
modern_chinese	**0.5431**	0.5241	0.1724	0.0862	0.2069	0.2328	0.3707
chinese_literature	**0.6127**	0.5412	0.1618	0.0833	0.2892	0.2500	0.5000
chinese_history	**0.5635**	0.4489	0.1672	0.0526	0.2632	0.1734	0.3839
chinese_driving_rule	**0.5198**	0.4420	0.1985	0.0687	0.2519	0.2595	0.3817
construction_project_management	**0.6619**	0.4388	0.1942	0.0647	0.2662	0.2230	0.3453
elementary_commonsense	**0.5404**	0.5354	0.1414	0.0758	0.2828	0.1465	0.4848
traditional_chinese_medicine	**0.5459**	0.5405	0.2378	0.0703	0.2324	0.2595	0.3027
business_ethics	**0.5359**	0.5167	0.2057	0.0574	0.2919	0.2249	0.3780
economics	**0.5283**	0.4465	0.2767	0.0440	0.2516	0.1824	0.3396
education	0.5460	**0.5215**	0.1840	0.0920	0.2761	0.3129	0.4049
college_education	**0.5981**	0.5234	0.2336	0.1589	0.4299	0.2617	0.4673
journalism	**0.5947**	0.5447	0.1744	0.0407	0.3256	0.2209	0.4826
marketing	**0.6556**	0.5944	0.2278	0.0889	0.2667	0.3500	0.2389
professional_accounting	**0.6114**	0.5829	0.1829	0.0800	0.1943	0.2400	0.3714
professional_psychology	**0.5645**	0.5303	0.2198	0.0819	0.1940	0.2586	0.3578
public_relations	**0.5920**	0.5172	0.1437	0.0862	0.2299	0.2931	0.4598
security_study	**0.5704**	0.4963	0.1630	0.1185	0.4074	0.3037	0.4000
high_school_geography	0.5678	**0.4831**	0.2712	0.0678	0.3305	0.1780	0.4237
management	**0.5905**	0.5857	0.2524	0.0857	0.2429	0.3190	0.3810
sociology	**0.5442**	0.5177	0.1991	0.0442	0.2832	0.3319	0.3805
marxist_theory	**0.6085**	0.5556	0.2804	0.0529	0.3122	0.2804	0.4497
college_law	**0.4648**	0.4519	0.2315	0.0833	0.2870	0.2500	0.3981
global_facts	**0.5235**	0.4930	0.2282	0.0738	0.3423	0.2349	0.4027
international_law	**0.5568**	0.5081	0.2270	0.0595	0.3351	0.2541	0.0919
jurisprudence	**0.5377**	0.5109	0.3187	0.0681	0.3358	0.2409	0.3869
world_religions	**0.5000**	0.4938	0.1125	0.0938	0.2625	0.1812	0.4688
logical	**0.5366**	0.4715	0.1463	0.0976	0.2195	0.2195	0.2846
professional_law	**0.5071**	0.4882	0.2654	0.0474	0.2180	0.2180	0.3033
philosophy	**0.5524**	0.4667	0.2476	0.0952	0.2190	0.2286	0.0476
world_history	**0.5342**	0.4907	0.2174	0.0683	0.3354	0.2298	0.4720
arts	0.0750	**0.5437**	0.1500	0.1000	0.2375	0.2125	0.4500
electrical_engineering	**0.5698**	0.5349	0.2035	0.1337	0.2965	0.2384	0.3663
actuarial_science	0.3692	0.1887	0.1698	0.1792	0.0642	0.1038	**0.5000**
engineering_hydrology	**0.5755**	0.5377	0.2170	0.0660	0.3113	0.3208	0.5189
genetics	**0.5455**	0.4830	0.1932	0.0795	0.1648	0.1761	0.2216
astronomy	**0.5758**	0.4606	0.2424	0.0727	0.1455	0.2182	0.3879
college_mathematics	**0.5048**	0.3238	0.2095	0.1429	0.1619	0.1524	0.1429
college_medical_statistics	**0.6038**	0.5094	0.1698	0.0566	0.1792	0.1887	0.0660
virology	**0.5680**	0.5089	0.1479	0.0414	0.1893	0.2899	0.2071
computer_science	**0.5686**	0.4216	0.1961	0.0882	0.1225	0.1814	0.2696
conceptual_physics	**0.5986**	0.5918	0.3878	0.0340	0.1565	0.2925	0.2109
anatomy	**0.6622**	0.5203	0.2027	0.0676	0.1689	0.2635	0.2905
machine_learning	**0.5656**	0.4754	0.1967	0.0574	0.1639	0.1393	0.2951
high_school_physics	**0.6364**	0.5455	0.1909	0.0727	0.2455	0.2198	0.3415
high_school_mathematics	**0.5915**	0.2317	0.1646	0.2012	0.1341	0.1098	0.3415
high_school_chemistry	**0.5758**	0.3409	0.2955	0.0682	0.1591	0.1288	0.0985
high_school_biology	**0.6746**	0.4852	0.2604	0.0828	0.1657	0.2249	0.0947
elementary_mathematics	**0.5996**	0.3870	0.1913	0.1565	0.1696	0.1478	0.5478
legal_and_moral_basis	**0.5748**	0.5561	0.2523	0.0888	0.3131	0.2523	0.5514
computer_security	**0.5789**	0.5088	0.2398	0.1053	0.2105	0.2339	0.2164
food_science	**0.4895**	0.4705	0.1469	0.0909	0.2587	0.1818	0.3077
college_medicine	**0.4982**	0.4908	0.2674	0.0366	0.1722	0.1538	0.1758
clinical_knowledge	**0.5738**	0.4557	0.1899	0.1013	0.2447	0.2616	0.1308
professional_medicine	**0.5798**	0.5505	0.3165	0.0771	0.2527	0.2287	0.2074
human_sexuality	**0.5635**	0.5317	0.2302	0.0317	0.2778	0.2143	0.3651
agronomy	**0.5621**	0.5089	0.1775	0.0651	0.2071	0.2485	0.4201
sports_science	**0.6424**	0.5697	0.2000	0.0909	0.2848	0.2667	0.3879
nutrition	**0.5586**	0.4759	0.1862	0.0621	0.1172	0.2414	0.2552
elementary_IT	**0.5462**	0.5378	0.2353	0.1050	0.1975	0.2353	0.2647

Table 11: Response Rate Comparison: *Sun-Shine 1.0* vs. Other LLMs.

Category	Conditional Accuracy						
	S-S	G-4o	L-318	G-35T	Q-2.5	Go1-m	DK-r1
ancient_chinese	0.2697	0.2500	0.2174	0.2500	0.2195	**0.2812**	0.2500
chinese_civil_service_exam	0.2562	0.2152	0.2222	0.1429	**0.3243**	0.2368	0.2658
chinese_food_culture	0.3258	0.3143	**0.3529**	0.2727	0.3226	0.2143	0.3171
chinese_foreign_policy	0.3444	0.3438	0.2973	0.0000	0.2667	0.3235	**0.3448**
chinese_teacher_qualification	0.2921	0.2752	0.1667	0.3077	**0.3636**	0.2571	0.2650
elementary_chinese	0.3126	0.2813	0.2857	**0.4286**	0.2958	0.2533	0.3145
ethnology	0.3719	0.4109	0.2400	0.2308	0.3864	0.3333	**0.5714**
high_school_politics	**0.3662**	0.3225	0.2963	0.2143	0.1667	0.2174	0.3412
modern_chinese	0.3592	0.3265	0.2000	**0.4000**	0.1250	0.1852	0.3489
chinese_literature	0.2980	0.2000	0.3030	0.1765	**0.3220**	0.3077	0.2879
chinese_history	**0.3407**	0.2552	0.2778	0.2941	0.2353	0.2418	0.2543
chinese_driving_rule	0.3299	0.2764	0.1923	0.1111	0.2424	0.3247	**0.3721**
construction_project_management	**0.3470**	0.2951	0.2593	0.2222	0.2162	0.2292	0.3253
elementary_commonsense	0.3017	0.2925	0.2500	**0.4667**	0.1964	0.2759	0.2654
traditional_chinese_medicine	0.2971	0.2100	0.2045	0.2308	**0.3023**	0.2738	0.2912
business_ethics	**0.3981**	0.3411	0.2791	0.3333	0.2623	0.2391	0.3802
economics	**0.3597**	0.3476	0.2955	0.2857	0.1500	0.3341	0.3254
education	0.3667	0.3421	0.3647	0.2667	0.3111	0.1818	**0.3845**
college_education	0.3125	0.2143	**0.3600**	0.2353	0.2174	0.2857	0.3478
journalism	0.2981	0.2366	0.2667	0.2857	0.2679	0.2281	**0.3098**
marketing	**0.3768**	0.3051	0.2683	0.2500	0.3750	0.3541	0.3254
professional_accounting	0.3748	0.3084	0.3438	**0.5714**	0.2647	0.3451	0.3746
professional_psychology	**0.3985**	0.2823	0.2157	0.3158	0.3333	0.3698	0.3945
public_relations	0.2427	0.1889	0.3600	0.1333	0.2250	0.2845	**0.4752**
security_study	0.2997	0.3333	0.3182	0.3125	0.3636	0.2258	**0.3977**
high_school_geography	**0.4179**	0.2807	0.2188	0.1250	0.2308	0.2368	0.3627
management	0.3089	0.2913	0.2264	0.3333	0.2941	0.3154	**0.5369**
sociology	**0.3333**	0.2997	0.2889	0.1000	0.1562	0.2062	0.3221
marxist_theory	**0.3217**	0.3143	0.2453	0.1000	0.2712	0.3210	0.3212
college_law	**0.3287**	0.2842	0.2800	0.2222	0.2258	0.2165	0.3109
global_facts	**0.3636**	0.3599	0.3235	0.0909	0.2353	0.3000	0.3611
international_law	**0.3830**	0.2524	0.2381	0.3636	0.2903	0.2714	0.3799
jurisprudence	**0.3258**	0.2667	0.2519	0.2857	0.2754	0.2786	0.3299
world_religions	0.5975	0.2911	0.2222	0.2667	0.2857	0.2143	**0.6475**
logical	0.2931	0.2879	0.2778	0.2500	0.2593	0.2849	**0.3648**
professional_law	**0.3178**	0.2223	0.1786	0.2000	0.2609	0.2865	0.3802
philosophy	0.3276	0.2429	0.2308	0.3000	**0.3478**	0.3169	0.3057
world_history	0.5658	0.2907	0.2286	**0.5455**	0.3148	0.3267	0.4407
arts	0.2759	0.2500	0.3333	**0.4375**	0.3684	0.2157	0.2453
electrical_engineering	0.3349	**0.3478**	0.3114	0.1304	0.2745	0.2456	0.2857
engineering_hydrology	0.3684	0.1148	0.1304	**0.5714**	0.2121	0.2500	0.4730
astronomy	**0.3903**	0.2237	0.2750	0.3750	0.3333	0.2111	0.3699
medical_statistics	0.2978	0.2656	0.2222	**0.5000**	0.2105	0.3636	0.3026
computer_science	0.4013	0.2845	0.2500	0.2778	0.3600	**0.3966**	0.3923
conceptual_physics	0.2729	0.2473	0.2712	0.2000	0.2174	0.2632	**0.2799**
anatomy	**0.2410**	0.2338	0.1333	0.2000	0.2000	0.1667	0.2367
machine_learning	0.2319	0.2241	0.2083	0.1429	0.2000	0.2498	**0.2729**
high_school_physics	0.2667	0.1857	0.1429	0.1250	0.2593	0.2441	**0.6366**
high_school_mathematics	0.6667	0.5000	0.3333	0.4500	**0.5268**	0.5246	0.5241
high_school_chemistry	0.2667	0.1974	0.3590	0.1111	0.2857	0.2459	**0.3636**
high_school_biology	0.3171	0.2544	0.2727	**0.5000**	0.2500	0.2321	0.3295
elementary_mathematics	0.3146	0.2519	0.2273	0.2778	**0.4615**	0.2830	0.2941
actuarial_science	0.3400	0.2500	0.4444	0.1579	0.3214	0.2800	**0.4737**
genetics	0.2941	0.2555	0.3529	0.2143	**0.3793**	0.3011	0.3582
college_mathematics	0.5882	0.2642	0.2727	0.2000	**0.4941**	0.1905	0.4896
virology	0.2083	0.1744	0.1200	**0.4286**	0.2188	0.3094	0.2157
legal_and_moral_basis	**0.3821**	0.3277	0.2963	0.1579	0.3284	0.2436	0.3684
computer_security	**0.3223**	0.2848	0.3171	0.1667	0.3167	0.2750	0.3033
food_science	0.3914	0.3824	0.1905	0.2308	0.2703	0.2250	**0.4154**
college_medicine	0.3134	0.2500	0.2055	0.2000	0.2128	0.2022	**0.3326**
clinical_knowledge	0.3448	0.3088	0.2222	**0.3333**	0.3426	0.2143	0.3248
professional_medicine	0.2431	0.2222	0.2605	**0.3448**	0.2737	0.2569	0.3433
human_sexuality	0.3134	0.3099	0.1726	**0.7500**	0.3714	0.3174	0.2989
agronomy	0.2791	0.2558	0.2667	0.0909	**0.3714**	0.2188	0.2626
sports_science	0.4171	0.2666	0.3936	0.2667	**0.4465**	0.2969	0.3830
nutrition	0.2319	0.2222	0.1481	0.2222	0.2941	0.3833	**0.4412**
elementary_IT	0.3185	**0.3906**	0.2500	0.3600	0.2128	0.2874	0.3452

Table 12: Conditional Accuracy Comparison: *Sun-Shine 1.0* vs. Other LLMs.

Category	Accuracy						
	S-S	G-4o	L-318	G-35T	Q-2.5	Go1-m	DK-r1
ancient_chinese	0.1341	0.1220	**0.3050**	0.0183	0.0549	0.0549	0.1159
chinese_civil_service_exam	0.1062	0.1000	0.0375	0.0062	0.0750	0.0025	**0.1562**
chinese_food_culture	0.1765	0.1618	0.0441	0.0221	0.0735	0.0294	**0.1838**
chinese_foreign_policy	**0.2056**	0.2028	0.1028	0.0000	0.0748	0.0654	0.1869
chinese_teacher_qualification	0.1676	0.1553	0.0447	0.0223	0.0894	0.0503	**0.2458**
elementary_chinese	**0.1866**	0.1695	0.0476	0.0238	0.0833	0.0397	0.1865
ethnology	0.2296	0.2456	0.0444	0.0222	0.1259	0.0519	**0.3333**
high_school_politics	0.1738	0.1834	0.0559	0.0210	0.0350	0.0210	**0.2098**
modern_chinese	**0.1897**	0.1654	0.0345	0.0345	0.0259	0.0603	0.1466
chinese_literature	0.1765	0.1882	0.0490	0.0147	0.0931	0.0686	**0.1961**
chinese_history	**0.1920**	0.1146	0.0464	0.0155	0.0619	0.0372	0.1238
chinese_driving_rule	0.1676	0.1492	0.0382	0.0076	0.0611	0.0763	**0.1985**
construction_project_management	**0.2230**	0.1295	0.0504	0.0144	0.0576	0.0504	0.1727
elementary_commonsense	0.1966	0.1714	0.0354	0.0354	0.0556	0.0354	**0.2424**
traditional_chinese_medicine	**0.1568**	0.1323	0.0486	0.0162	0.0703	0.0703	0.0973
business_ethics	**0.2057**	0.1852	0.0574	0.0191	0.0766	0.0478	0.1914
economics	0.1195	0.1195	0.0818	0.0126	0.0377	0.0252	**0.1824**
education	**0.1865**	0.1795	0.0675	0.0245	0.0859	0.0613	0.1779
college_education	0.1867	**0.2100**	0.0741	0.0374	0.0935	0.1028	0.2056
journalism	0.1829	0.1676	0.0465	0.0116	0.0872	0.0640	**0.2035**
marketing	0.1833	**0.2000**	0.0611	0.0222	0.1000	0.1167	0.1667
professional_accounting	**0.1886**	0.1714	0.0629	0.0457	0.0514	0.0629	0.1523
professional_psychology	**0.1897**	0.1736	0.0474	0.0259	0.0647	0.0690	0.1681
public_relations	**0.1437**	0.0977	0.0517	0.0115	0.0517	0.0862	0.1389
security_study	0.1333	**0.2222**	0.0519	0.0370	0.1481	0.0593	0.1259
high_school_geography	0.1356	**0.2333**	0.0593	0.0085	0.0763	0.0508	0.2034
management	0.1810	0.1695	0.0571	0.0286	0.0714	0.1163	**0.2143**
sociology	0.1593	**0.1726**	0.0575	0.0044	0.0442	0.0619	0.1681
marxist_theory	0.1957	0.1799	0.0688	0.0053	0.0847	0.0741	**0.2222**
college_law	0.1574	0.1222	0.0648	0.0185	0.0648	0.0833	**0.1759**
global_facts	**0.1822**	0.1621	0.0738	0.0067	0.0805	0.0671	0.1409
international_law	0.1405	**0.1965**	0.0541	0.0216	0.0973	0.0595	0.0270
jurisprudence	0.1405	**0.1998**	0.0803	0.0195	0.0925	0.0706	0.1752
world_religions	0.1437	0.1188	0.0250	0.0250	0.0750	0.0312	**0.2750**
logical	**0.1546**	0.1382	0.0407	0.0244	0.0569	0.0813	0.1220
professional_law	**0.1611**	0.1009	0.0474	0.0095	0.0569	0.0474	0.0758
philosophy	0.0667	**0.1810**	0.0571	0.0286	0.0762	0.0571	0.0095
world_history	0.1543	0.1304	0.0497	0.0373	0.1056	0.0807	**0.1577**
arts	0.1880	0.1500	**0.5000**	0.0438	0.0875	0.0438	0.1750
electrical_engineering	0.1429	0.1190	0.0581	0.0174	0.0814	0.0523	**0.2209**
engineering_hydrology	0.1981	0.1799	0.0283	**0.3770**	0.0660	0.0755	0.1792
astronomy	0.1273	0.1030	0.0667	0.0242	0.0545	0.0970	**0.1697**
medical_statistics	**0.1604**	0.1415	0.0377	0.0283	0.0377	0.0472	0.0283
computer_science	**0.1618**	0.1275	0.0049	0.0245	0.0441	0.0441	0.1127
conceptual_physics	**0.1497**	0.1361	0.1020	0.0068	0.0340	0.0748	0.1224
anatomy	**0.2230**	0.1216	0.0270	0.0135	0.0338	0.0405	0.0878
machine_learning	**0.1311**	0.1000	0.0410	0.0082	0.0328	0.0328	0.1066
high_school_physics	0.1818	0.1964	0.0273	0.0091	0.0636	0.0455	**0.2091**
high_school_mathematics	0.1182	0.1453	0.0549	0.0549	0.1341	0.0671	**0.2378**
high_school_chemistry	**0.1136**	0.0970	0.1061	0.0076	0.0455	0.0682	0.0303
high_school_biology	**0.1716**	0.1534	0.0710	0.0414	0.0414	0.0651	0.0178
elementary_mathematics	**0.1432**	0.1217	0.0435	0.0435	0.0783	0.0478	0.1139
actuarial_science	0.0755	0.0472	0.0755	0.0283	0.0849	0.0660	**0.2453**
genetics	0.1307	**0.1403**	0.0682	0.0170	0.0625	0.0739	0.0625
college_mathematics	**0.1333**	0.0952	0.0571	0.0286	0.0952	0.0286	0.0762
virology	**0.1183**	0.0888	0.0178	0.0178	0.0414	**0.1183**	0.0769
legal_and_moral_basis	0.2182	0.2022	0.0748	0.0140	0.1028	0.0654	**0.3411**
computer_security	**0.2047**	0.1579	0.0076	0.0175	0.0877	0.0877	0.1228
food_science	0.1366	**0.1936**	0.0280	0.0210	0.0699	0.0140	0.1329
college_medicine	**0.1538**	0.1245	0.0549	0.0073	0.0366	0.0366	0.0623
clinical_knowledge	**0.1772**	0.1565	0.0422	0.0338	0.0844	0.0591	0.0549
professional_medicine	**0.1410**	0.1230	0.0824	0.0266	0.0691	0.0718	0.0691
human_sexuality	**0.1746**	0.1667	0.0397	0.0238	0.1032	0.0635	0.1667
agronomy	0.1420	0.1302	0.0473	0.0059	0.0769	0.0533	**0.2071**
sports_science	**0.2242**	0.2189	0.0727	0.0242	0.1091	0.0788	0.1333
nutrition	**0.1241**	0.1103	0.0276	0.0138	0.0345	0.0759	0.1034
elementary_IT	0.1092	**0.2102**	0.0588	0.0378	0.0420	0.0798	0.1387

Table 13: Accuracy Comparison: *Sun-Shine 1.0* vs. Other LLMs.

F.2 Experiment on Tibetan Culture

F.2.1 Classical Chinese Generation

As shown in Figure 16, *Sun-Shine 1.0* substantially outperforms DeepSeek-R1 (Guo et al., 2025) and GPT-4o (Achiam et al., 2023) in classical Chinese. This advantage arises from training on *TIBSTC*, which contains ancient poems, essays, and historical records. Manual revision preserves classical grammar, sentence structure, and style, resulting in outputs with a distinctly classical character, in contrast to the more modern style of DeepSeek-R1 (Guo et al., 2025) and GPT-4o (Achiam et al., 2023).

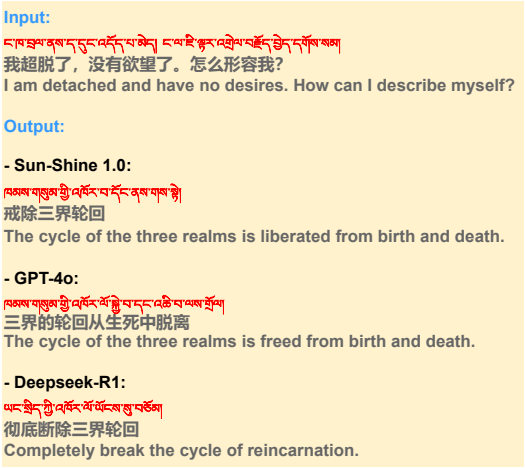


Figure 16: Comparison of LLM Output

F.2.2 Ancient Poetry Writing

As shown in Figure 17, different LLMs exhibit clear stylistic differences when responding to the culturally sensitive prompt "Please write a poem in praise of father" across Tibetan, Chinese, and English. *Sun-Shine 1.0* maintains a neutral and restrained tone, avoiding exaggerated praise or references to Tibetan classics. In contrast, GPT-4o (Achiam et al., 2023) adopts an amplified, scripture-like style, while DeepSeek-R1 (Guo et al., 2025) frequently invokes the bodhisattva Mañjuśrī, resulting in more embellished and overfitted responses. Overall, *Sun-Shine 1.0* outperforms DeepSeek-R1 (Guo et al., 2025) and GPT-4o (Achiam et al., 2023) across most metrics, reflecting the impact of curated *TIBSTC* data and human-in-the-loop training.

F.3 Supplementary Experiment

We evaluate these LLMs on the Tibet Autonomous Region Tibetan–Chinese Translation 2024 test[2]

[2] http://cn.zyw.xizang.gov.cn/

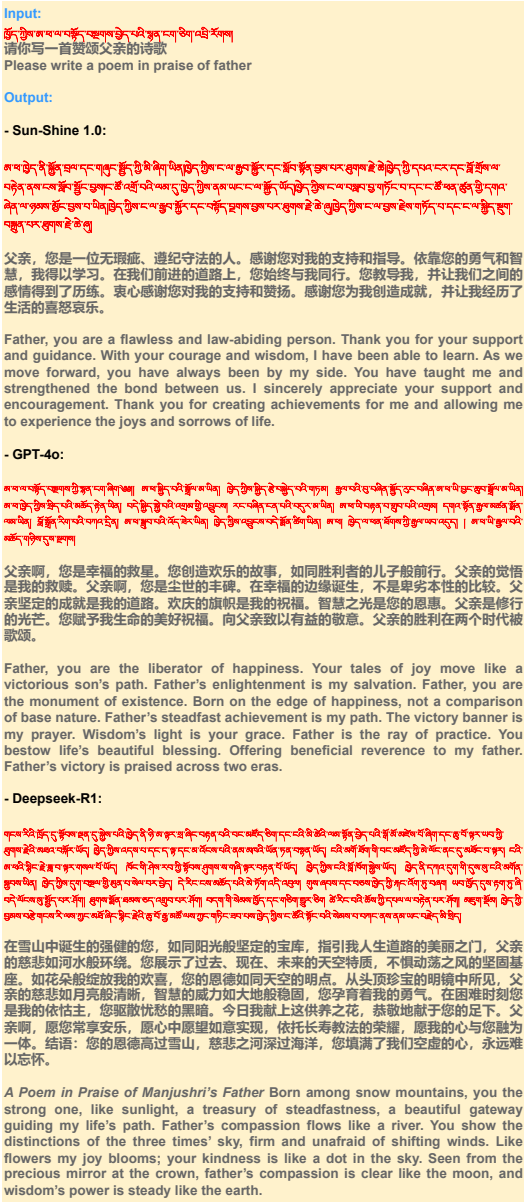


Figure 17: Ancient Poetry Writing Generation

(100-point TLP system). As shown in Table 14, *Sun-Shine 1.0* performs best on the more challenging A and B exams, which emphasize ancient Tibetan content, while achieving lower scores on the C and D exams that focus on modern Tibetan, including recent political topics. This pattern highlights *Sun-Shine*'s strong proficiency in classical Tibetan and reflects the advantage of the *TIBSTC* dataset, which contains a richer collection of ancient Tibetan texts.

LLM	A	B	C	D
Sun-Shine	**63**	**61**	44	42
GPT-4o (Achiam et al., 2023)	45	47	53	51
DeepSeek-R1 (Guo et al., 2025)	59	56	**67**	**66**
Qwen-2.5 (Bai et al., 2024)	41	43	53	54

Table 14: Comparison of TLP Test