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 in Tibetan remains severely constrained. While recent efforts have begun to address pre-training data scarcity for Tibetan, a more fundamental gap persists: no existing resource supports the *complete* LLM development pipeline, spanning pre-training, instruction tuning, safety alignment, preference optimization, and reasoning supervision. We introduce the **T**ibetan **F**oundation **D**ataset (**TFD**), the first structured, large-scale, and expert-curated dataset covering all key stages of Tibetan large language modeling. *TFD* comprises *TIBSTC*, a unified corpus of over 11 billion tokens with curated sub-datasets for instruction tuning, safety alignment, and preference optimization, and *TIBSTC-CoT*, the first large-scale Tibetan chain-of-thought dataset. We demonstrate its utility by training the *Sun-Shine* family of Tibetan LLMs, achieving substantial improvements over strong baselines on understanding, safety, reasoning, and generation benchmarks. These results underscore that advancing low-resource language modeling requires not only scale, but a structurally complete data ecosystem. We release *TFD* to facilitate reproducible research and the development of robust, culturally aligned Tibetan LLMs.

1 Introduction

LLMs (Hurst et al., 2024; Achiam et al., 2023; Dubey et al., 2024; Guo et al., 2025; Liu et al., 2024) 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), such as Tibetan (Huang et al., 2025), remain significantly underserved. A major factor underlying this limitation is not only the scarcity of raw Tibetan text, but the absence of a structurally complete data ecosystem covering all stages of modern LLM development. 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). While recent efforts have begun to address pre-training data scarcity for Tibetan (Pan et al., 2025), a critical gap remains: no existing resource jointly supports instruction tuning, safety alignment, preference optimization, and reasoning supervision, components that are now standard and indispensable in modern LLM pipelines. As a result, current Tibetan LLMs exhibit limited generalization, weak discourse modeling, and unstable generation across domains (Liu and Best, 2025).

To address this gap, we introduce the **T**ibetan **F**oundation **D**ataset (**TFD**), the first structured, large-scale, and expert-curated Tibetan dataset designed to cover the *complete* LLM development pipeline. *TFD* consists of two sub-datasets, *TIBSTC* and *TIBSTC-CoT*. *TIBSTC* covers over 11 billion tokens from domains such as literature, law, religion, medicine, and daily communication, and uniquely includes curated sub-datasets for instruction tuning (*Alpaca-Ti*), safety alignment (*Safety-Prompts-Ti*), and preference optimization (*CValues-Ti*, *hh-rlhf-Ti*), components absent from prior Tibetan corpora. *TIBSTC-CoT* 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 structural gap in the Tibetan NLP ecosystem.

To validate *TFD*'s utility, we train and release the *Sun-Shine* family of reference Tibetan LLMs based on Llama (Dubey et al., 2024) and Qwen (Yang et al., 2025) backbones. While the primary contribution of this work lies in the construction of TFD, the Sun-Shine family serves as a proof of concept for building culturally aligned and linguistically robust Tibetan LLMs. Our main contributions are summarized as follows:

- We introduce *TFD*, the *first* Tibetan dataset covering the complete LLM development pipeline, comprising *TIBSTC*, a multi-domain corpus with over 11 billion tokens together with curated datasets for instruction tuning, safety alignment, and preference optimization, and *TIBSTC-CoT*, the first large-scale Tibetan chain-of-thought dataset for explicit multi-step reasoning.
- 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 release the *Sun-Shine* family, a set of reference Tibetan LLMs trained on *TFD*, to demonstrate the dataset's effectiveness in instruction following, safety-aware generation, and multi-step reasoning.
- We conduct error analysis, case studies, and human evaluations to provide qualitative insights into the strengths and limitations of the proposed models.

2 Related Work

2.1 Tibetan Language Modeling

Tibetan remains a severely low-resource language in NLP, characterized by complex stacked syllables, rich morphology, and limited digitized corpora (Wagner and Zeisler, 2004). Existing resources mainly target narrow tasks, and are typically small-scale, domain-specific, and lack standardization (Liu and Best, 2025; Huang et al., 2025; Liu et al., 2025b,a). Recent work has begun to address pre-training data scarcity (Pan et al., 2025), demonstrating that scale alone can yield meaningful improvements. However, pre-training corpora address only one stage of the LLM pipeline. No existing Tibetan dataset jointly supports instruction tuning, safety alignment, preference optimization, and reasoning supervision—components now standard in modern LLM development (Zhuang and Sun, 2025; Xu et al., 2025; Feng et al., 2025; Zhang et al., 2025). TFD is designed to fill this structural gap as the first unified resource covering the complete LLM pipeline for Tibetan.

2.2 Chain of Thought

Chain-of-thought (CoT) prompting has proven effective for improving multi-step reasoning in LLMs (Wei et al., 2022; Wang et al., 2023), but remains almost entirely concentrated on 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 and cultural grounding. *TIBSTC-CoT* directly addresses this gap as the first large-scale native Tibetan chain-of-thought dataset.

3 TFD

In order to address the limitations discussed above, we introduce the *TFD*, the foundational dataset developed for Tibetan LLMs, which consists of two sub-datasets: *TIBSTC* for general supervised fine-tuning and *TIBSTC-CoT* for chain-of-thought reasoning tasks. Details of the data collection process are provided in Appendix A.

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: Statistics of the TIBSTC sub-datasets

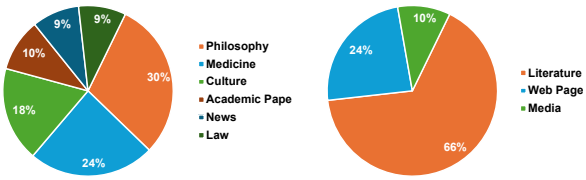
3.1 TIBSTC

TIBSTC is a multi-component Tibetan data collection for LLM development. It consists of five categories: *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* contains 122,756,184 instances and exhibits heterogeneous distributions across source types and content categories. *Corpus* is primarily used for Tibetan

LLM pre-training. A representative example illustrating its structure and content is provided in Appendix B.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 Tibetan sub-dataset, *Alpaca-Ti*, specifically designed for supervised instruction fine-tuning. Both *Alpaca-Ti* and *Safety-Prompts-Ti* (Section 3.1.3) were constructed by translating the original English Alpaca (Taori et al., 2023) and Safety-Prompts (Sun et al., 2023) resources, followed by structured localization and a two-stage expert alignment protocol. For each source instance, Tibetan drafts were generated using both Google Translate and Claude-3.5-Sonnet, and the higher-quality draft was selected based on independent human evaluation by two Tibetan language specialists and five trained annotators. Selected drafts then underwent two rounds of expert review covering semantic fidelity, fluency, domain knowledge alignment, and cultural alignment, raising the expert approval rate from 28.74% (initial draft) to 100% after adjudication. Detailed construction statistics, quality improvement across stages, and inter-annotator agreement scores (α and κ) are reported in Appendix A.2.1.

As shown in Table 1 and Table 2, compared with the original Alpaca dataset, *Alpaca-Ti* has a longer average instance length, providing richer instruction-following supervision and improving the generalization ability of Tibetan LLMs. A representative example demonstrating the structure and content of *Alpaca-Ti* is provided in Appendix B.1.2.

3.1.3 Safety-Prompts-Ti

Safety-Prompts-Ti is based on *Safety-Prompts* (Sun et al., 2023) and is designed to improve the safety alignment of Tibetan LLMs. It follows the same dual-model drafting and two-stage expert alignment protocol described above for *Alpaca-Ti*, with construction details provided in Appendix A.2.1.

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

Table 2: Category Details of *Alpaca-Ti*

The dataset contains 38,156 instances in total and covers two major categories: *Instruction Attacks* and *Safety Scenarios*, as shown in Table 1. As shown in Table 3, the *Instruction Attacks* category contains 15,000 instances across six subtypes, with 2,500 instances for each subtype. In addition, we compile 23,000+ samples under *Typical Safety Scenarios*, covering eight safety-critical topics. A representative example demonstrating its structure and content is provided in Appendix B.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 *CValues-Ti*, a small safety-oriented sub-dataset based on *CValues* (Frantzi et al., 2000; Jiang et al., 2024), to reduce negative social impacts in Tibetan LLM outputs. It contains three response categories: *Safe & Responsibility*, *Safe*, and *Unsafe*. Under the same prompt, responses are categorized into positive and negative samples according to their safety level and response quality. As summarized in Table 4, these categories form three common response combinations for preference-oriented training. A representative example of structure and content of *CValues-Ti* can be found in Appendix B.1.4.

3.1.5 hh-rlhf-Ti

hh-rlhf-Ti is a preference-oriented Tibetan sub-dataset adapted from the HH-RLHF setting (Ouyang et al., 2022; Kirk et al., 2023).

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

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

It consists of two parts: *harmless* with 23,822 instances and *helpful* with 23,589 instances. To better fit Tibetan linguistic characteristics, we reorganize and translate the original preference data with consideration of Tibetan grammar and expression patterns. This sub-dataset can be integrated into preference-oriented training objectives and is intended to improve the helpfulness and harmlessness of Tibetan LLMs. Appendix B.1.5 provides an example of *hh-rlhf-Ti*.

3.2 TIBSTC-CoT

TIBSTC-CoT is a large-scale Tibetan chain-of-thought reasoning sub-dataset constructed through a multi-stage, multi-model pipeline. As shown in Figure 2, it covers five major domains: Science and Engineering (SE), Life Sciences (LS), Humanities and Social Sciences (HSS), Social Sciences (SS), and Comprehensive (Com), each of which includes multiple fine-grained subcategories. Appendix B.2 Table 14 reports the number of entries for each task and its distribution across three task dimensions: Generation, Knowledge, and Classification. Details of the *TIBSTC-CoT* generation process are provided in Appendix A.3 and Appendix B.2.

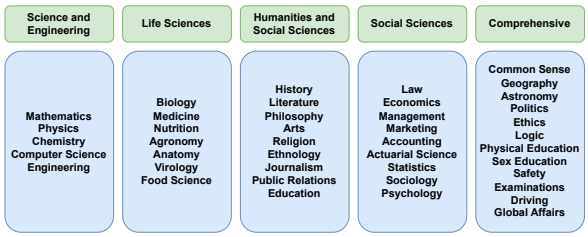


Figure 2: Domain Distribution of *TIBSTC-CoT*

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) backbone (Dubey et al., 2024). Its training pipeline 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 further enhances its ability to follow task-specific instructions and generate responses that satisfy safety and alignment requirements. Finally, DPO aligns the model's outputs with preference signals, improving response relevance and alignment quality. Detailed hyperparameters and training configurations of *Sun-Shine 1.0* are provided in Appendix C.1.

4.1.2 Reasoning & Thinking

Sun-Shine 2.0 is built on the Qwen-3 backbone (Yang et al., 2025) and trained using supervised fine-tuning (SFT) (Yu et al., 2021). It is designed to improve explicit multi-step reasoning by learning from instruction data with structured intermediate reasoning steps. The training data span multiple domains and task types, enabling the model to perform reasoning-oriented generation, knowledge-intensive 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. Detailed hyperparameters and training configurations of *Sun-Shine 2.0* are provided in Appendix C.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 D.

4.2.1 Open-source LLM

We evaluate several open-source 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., 2024), 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 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. Each subject includes a five-shot development set and a test set of more than 100 questions. *Ti-MMLU* covers 17 STEM tasks, 13 humanities tasks, 22 social science tasks, and 15 tasks from other domains. Notably, 16 tasks are China-specific, meaning that they either lack direct equivalents in other regions or may have context-dependent answers. The evaluation framework follows the original *CMMLU* metrics (Li et al., 2024), including:

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

4.3.2 Ti-SafetyBench

Ti-SafetyBench is constructed from *SafetyBench* (Zheng et al., 2024) through expert translation and manual validation to ensure consistency with Tibetan linguistic and cultural norms. It contains 11,435 multiple-choice questions across seven 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 directions among Tibetan (BO), Chinese (ZH), and English (EN), assessing both understanding and generation capabilities. We conduct experiments on the *FLORES-200* benchmark (Costa-Jussà et al., 2022), a widely used multilingual translation benchmark that includes diverse language pairs and covers multiple domains, styles, and text types. This benchmark is particularly suitable for evaluating low-resource translation because it includes underrepresented languages and reflects diverse translation scenarios.

4.4.2 Metrics

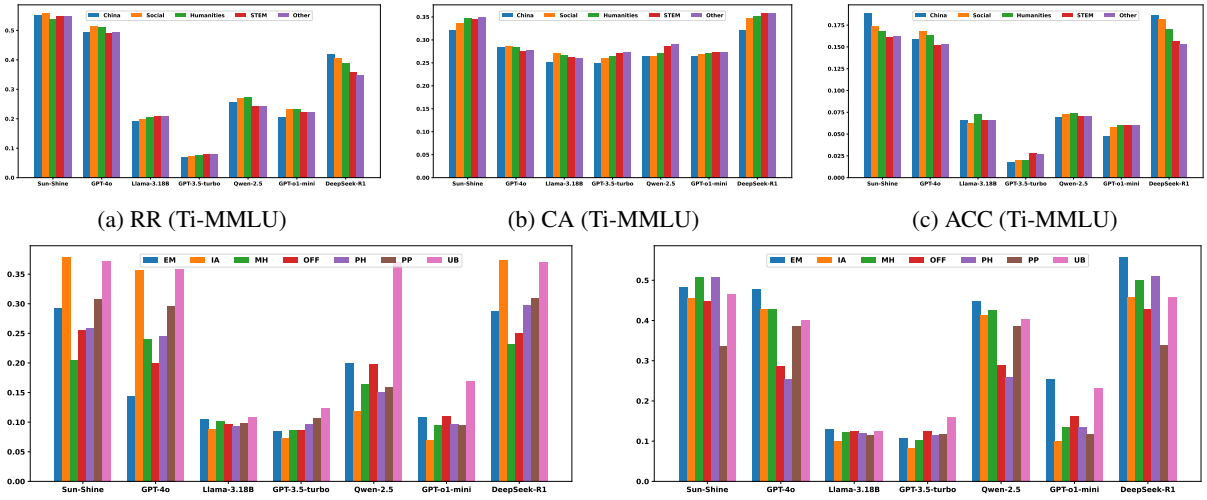
We adopt standard automatic evaluation metrics to assess both translation and generation quality. For machine translation, we report BLEU-4 (Papineni et al., 2002) and chrF++ (Popović, 2017). BLEU-4 measures n-gram precision up to four grams and is widely used for translation evaluation. chrF++ computes character-level precision and recall, making it suitable for morphologically rich and low-resource languages such as Tibetan. When used with the *FLORES-200* benchmark (Costa-Jussà et al., 2022), chrF++ provides a fine-grained assessment of translation quality by capturing character-level differences that may be overlooked by token-level metrics.

For generation tasks, we report ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004). ROUGE-1 and ROUGE-2 measure unigram and bigram overlap, respectively, while ROUGE-L computes the longest common subsequence between the generated text and the reference. Together, these ROUGE metrics provide indicators of lexical overlap and sequence-level similarity for generated outputs.

4.5 Experimental Setup

We investigate several key aspects of LLM performance in Tibetan:

- We evaluate all LLMs on *TLUE* (Gao et al.,



(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*

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) with *Ti-MMLU* (Gao et al., 2025), and further extend this analysis to safety evaluation by comparing *SafetyBench* (Sun et al., 2023) with *Ti-SafetyBench* (Gao et al., 2025).
- We compare reasoning-optimized LLMs, chat-oriented LLMs, and the *Sun-Shine* family to examine whether reasoning-oriented training improves performance.
- We study the effect of model size on Tibetan language understanding by 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

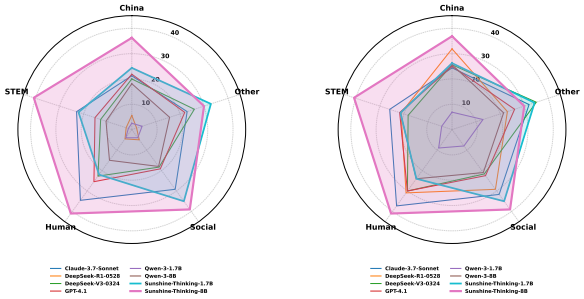
As shown in Figures 3a, 3b and 3c, *Sun-Shine 1.0*, GPT-4o, and DeepSeek-R1 outperform all other LLMs. *Sun-Shine 1.0* leads in Response Rate across all sub-branches (Appendix E.1.2, Figure 19) and overall Accuracy in China-Specific, STEM, and Other categories, while DeepSeek-R1 holds a slight edge in Conditional Accuracy. These results demonstrate that a structured pipeline combining large-scale Tibetan pretraining on TIBSTC with SFT and DPO alignment enables an 8B model to compete against significantly larger systems. We attribute this to the **pipeline as a whole**: TIBSTC contributes both scale and domain breadth, together with curated alignment components absent from single-purpose pretraining corpora (see Section 5.4).

On *Ti-SafetyBench* (Figures 3d, 3e), *Sun-Shine 1.0* excels in IA, OFF, UB, and MH categories, directly attributable to targeted supervision from Safety-Prompts-Ti and CValues-Ti.

5.1.2 Sun-Shine 2.0

As shown in Figure 4, *Sunshine-Thinking* models consistently outperform all baselines across five domains under both DA and CAA evaluations. *Sunshine-Thinking-8B* achieves the highest or near-highest scores across all settings, 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* surpasses GPT-4.1 and Qwen-3-8B on Social Sciences (35.00% DA and CAA), highlighting the effectiveness of structured supervision at smaller scales.

To verify that these gains stem from supervision structure rather than token quantity, we conduct a token-matched ablation under a fixed 76M



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

Figure 4: *TLUE* Result of Sunshine-thinking Family and Other LLMs (×100%)

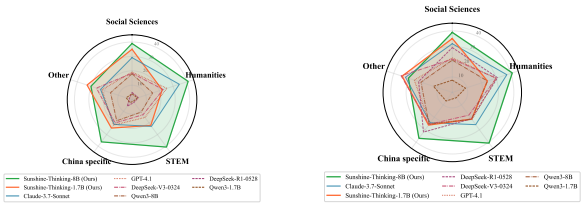
Setting	China	Other	Social	Human	STEM	Avg
Base (no SFT)	6.88	12.86	8.00	9.00	4.29	8.21
C: Alpaca-Ti-SFT	15.10	20.30	15.70	16.20	10.40	15.54
B: DeCoT-SFT	14.60	21.90	17.40	16.50	11.30	16.34
A: CoT-SFT	**14.80**	**24.90**	**22.60**	16.10	**13.90**	**18.46**

Table 5: Ti-MMLU (CAA) under token-matched SFT (76M tokens, Qwen3-1.7B). CoT-SFT outperforms both baselines, with gains concentrated in reasoning-intensive categories.

SFT budget on Qwen3-1.7B, comparing CoT-SFT, DeCoT-SFT (rationales removed), and Alpaca-Ti-SFT.

As shown in Tables 5 and 6, CoT-SFT achieves the highest Ti-MMLU CAA average (18.46 vs. 16.34 and 15.54), with gains concentrated in reasoning-intensive categories, while translation performance remains comparable across all conditions. Since *Sun-Shine 2.0* undergoes no corpus-level Tibetan pretraining, these results confirm that performance improvements are attributable to "supervision structure" rather than increased Tibetan exposure.

Figure 5 shows radar charts comparing the performance of Sunshine-Thinking models and baselines across five subject categories on the *Ti-MMLU* 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.



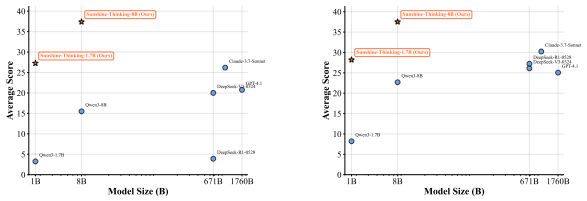
(a) LLMs on DA (b) LLMs on CAA

Figure 5: Performance on *Ti-MMLU* Sub-Benchmark

Setting	BO→ZH	BO→EN	ZH→BO	EN→BO	Avg
Base (no SFT)	0.24	4.23	2.71	3.64	2.71
C: Alpaca-Ti-SFT	5.80	20.60	16.80	18.80	15.50
B: DeCoT-SFT	6.20	21.00	17.60	19.10	15.98
A: CoT-SFT	6.00	**21.30**	17.40	**19.30**	**16.00**

Table 6: Translation performance (chrF++ ×100) under token-matched SFT. Scores are comparable across conditions, confirming that CoT supervision primarily strengthens reasoning rather than general language exposure.

Figure 6 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 6: 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 7, Sun-Shine achieves the highest scores across all metrics, demonstrating superior translation capabilities, followed by DeepSeek-R1, with GPT-4o and Qwen-2.5 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 7: Comparison of Translation Performance

5.2.2 Sun-Shine 2.0

Table 8 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 24.81 on ZH → BO and 25.15 on EN → BO, matching or even surpassing much larger LLMs such as Claude-3.7-Sonnet. The smaller *Sunshine-Thinking-1.7B* also substantially outperforms Qwen3-1.7B, with approximately 7 times improvements on Tibetan generation directions (20.47 vs. 2.71 on ZH → BO), underscoring the effectiveness of curated data and instruction tuning. Although recent work has identified quality issues in certain FLORES-200 language pairs, the performance gaps observed here are of a magnitude that cannot plausibly be attributed to reference translation noise, and rankings remain consistent across all three evaluation metrics. We nonetheless acknowledge this as a limitation and recommend future work explore higher-quality Tibetan-specific benchmarks.

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., 2024)	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 8: 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 E.2.1 Figure 20) and ancient poetry writing (Appendix E.2.2 Figure 21). 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 E.3 Table 14. Detailed qualitative analyses and case studies are provided in Appendix E.2.

5.4 Relationship to Concurrent Work

Pan et al. (2025) construct a large-scale Tibetan pre-training corpus through continual pre-training on web-crawled text, representing an important step toward addressing data scarcity for Tibetan. However, their work focuses exclusively on pre-training corpora and does not address the remaining stages of the LLM development pipeline. As shown in Table 9, TIBSTC goes beyond pre-training by providing curated datasets for every subsequent stage: instruction tuning (*Alpaca-Ti*, *Safety-Prompts-Ti*), preference optimization (*CValues-Ti*, *hh-rlhf-Ti*), and explicit reasoning supervision (*TIBSTC-CoT*), none of which have counterparts in Pan et al. This structural completeness is the defining distinction of TFD: it is not a larger pre-training corpus, but the first dataset ecosystem that enables the full LLM training pipeline for Tibetan without requiring additional data collection from external sources.

Component	TFD	Pan et al.
Pre-training Corpus	✓	✓
Instruction Tuning	✓	×
Safety Alignment	✓	×
Preference Optimization	✓	×
Chain-of-Thought Reasoning	✓	×

Table 9: Comparison of TFD and Pan et al. (2025) in terms of Pipeline Coverage

6 Conclusion

We introduced *TFD*, the first structured Tibetan dataset covering the complete LLM development pipeline: pre-training, instruction tuning, safety alignment, preference optimization, and reasoning supervision. While recent work has demonstrated the value of large-scale Tibetan pre-training corpora, *TFD* addresses a distinct need by providing curated alignment and reasoning supervision components that no prior Tibetan resource offers. By unifying *TIBSTC*, a multi-domain corpus of over 11 billion tokens with curated alignment subdatasets, and *TIBSTC-CoT*, the first large-scale Tibetan chain-of-thought dataset, *TFD* enables the full LLM training pipeline for Tibetan without requiring additional data collection. Experiments with the *Sun-Shine* family demonstrate strong performance on understanding, safety, reasoning, and generation benchmarks, surpassing much larger general-purpose models on several Tibetan-specific tasks. We release *TFD* to facilitate reproducible research and the development of robust, safe, and culturally aligned Tibetan LLMs.

Limitations

TFD addresses the complete LLM development pipeline for Tibetan, but several limitations remain. First, despite broad domain coverage, some specialized and dialectal varieties of Tibetan are underrepresented due to data availability constraints. Second, *TIBSTC-CoT* is constructed via a hybrid LLM-assisted and expert-verified pipeline, which may introduce residual model biases despite quality filtering. Third, while *TFD* uniquely provides safety alignment and preference optimization data, our evaluations remain text-based and do not yet extend to multimodal settings. Finally, as the primary contribution of this work is dataset construction rather than model optimization, we do not exhaustively explore training strategies or architectural choices, leaving such investigations to future work that builds on *TFD*.

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. 2024. 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, Ziyue Zhang, Cheng Huang, Yongbin Yu, Xiangxiang Wang, Yuqing Cai, and Nyima Tashi. 2025a. Listening, imagining & refining: A heuristic optimized asr correction framework with llms. *arXiv preprint arXiv:2509.15095*.

Yutong Liu, Ziyue Zhang, Ban Ma-bao, Renzeng Duojie, Yuqing Cai, Yongbin Yu, Xiangxiang Wang, Fan Gao, Cheng Huang, and Nyima Tashi. 2025b. 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 Data Collection of TFD

A.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.

A.2 Data Processing Pipeline 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.

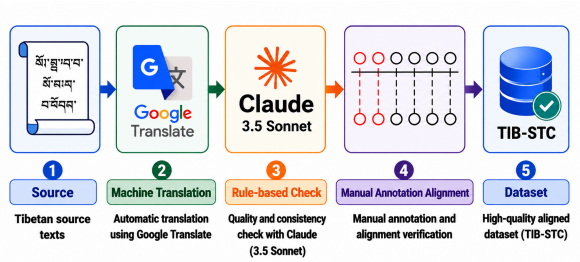


Figure 7: Data Processing Pipeline of *TIBSTC*

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.

A.2.1 Construction Protocol of Alpaca-Ti and Safety-Prompts-Ti

Both Alpaca-Ti and Safety-Prompts-Ti were translated from the original English Alpaca and Safety-Prompts resources, followed by structured localization and expert validation. They were neither adopted as raw machine translations nor created entirely from scratch.

Dual-Model Drafting For each source instance, Tibetan drafts were generated using both Google Translate and Claude-3.5-Sonnet. Although prior work suggests relatively stronger Tibetan capability for Claude, the final draft selection was determined exclusively through independent human evaluation rather than by assuming model superiority a priori.

Comparative Human Evaluation Two Tibetan language specialists and five trained annotators independently evaluated both translation outputs along four dimensions: semantic fidelity, linguistic naturalness, domain knowledge alignment (0–5 scale), and cultural alignment (0–5 scale). The higher-quality version was selected for further refinement. Table 10 reports the initial translation quality of both systems.

Metric	Google	Claude
Expert Approval Rate	11.54%	28.74%
Domain Knowledge Alignment (0–5)	0.95	2.30
Cultural Alignment (0–5)	0.85	1.90

Table 10: Initial translation quality based on human evaluation. Draft selection was evidence-based rather than model-driven.

Linguistic Normalization Selected drafts underwent Tibetan-specific orthographic normalization and rule-based consistency checks. Common issues addressed included literal word-by-word translations producing unnatural Tibetan syntax, inconsistent rendering of classical Tibetan terminology, and mechanical translation artifacts in safety-sensitive contexts. For example, direct lexical translations of "alignment" initially produced semantically unnatural expressions in Tibetan discourse and were replaced with culturally appropriate equivalents validated by language specialists.

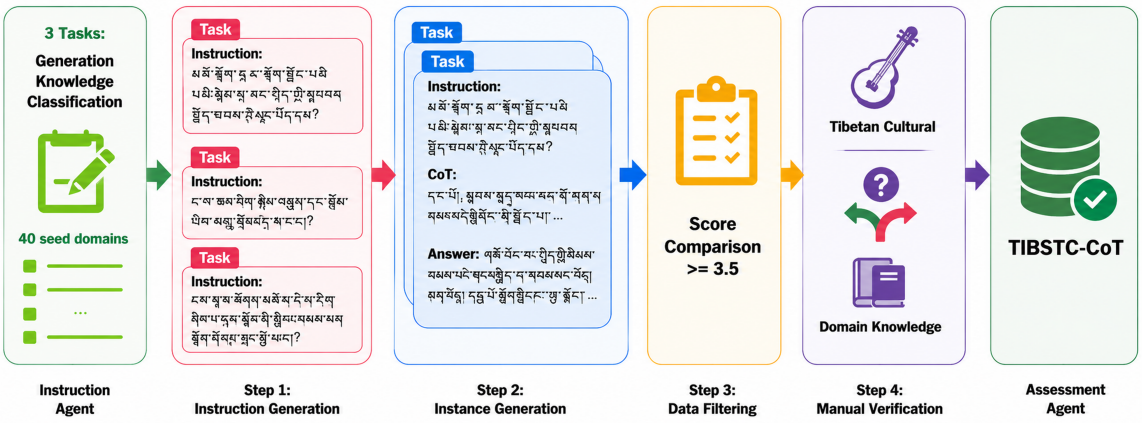


Figure 8: Data Processing Pipeline of *TIBSTC-CoT*

Two-Stage Human Alignment Each item was independently reviewed by annotators prior to expert adjudication. The verification process consisted of two passes: a first pass focused on semantic and fluency refinement, and a second pass focused on domain knowledge and cultural verification. Table 11 reports quality improvements across the alignment stages.

Stage	Approval	Domain	Cultural
Initial (Claude draft)	28.74%	2.30	1.90
After 1st Alignment	81.96%	3.68	3.47
After 2nd Alignment	100%	4.57	4.38

Table 11: Quality improvement across alignment stages. The final stage was expert-adjudicated, yielding 100% approval. Domain and Cultural alignment are on a 0–5 scale.

The staged improvements confirm that the final datasets are carefully curated Tibetan resources rather than direct machine translations.

Inter-Annotator Agreement To quantify annotation reliability, we conducted IAA analysis on two randomly sampled subsets of 1,000 instances per dataset. Each instance was independently rated on a 0–5 ordinal scale by two annotators prior to expert adjudication. We report both ordinal agreement (Krippendorff's α) and binary agreement (Cohen's κ, treating scores ≥ 4 as Accept):

- **Alpaca-Ti**: $\alpha = 0.78$, $\kappa = 0.82$
- **Safety-Prompts-Ti**: $\alpha = 0.71$, $\kappa = 0.76$

All values fall within the substantial agreement range under standard benchmarks. The slightly lower agreement observed for Safety-Prompts-Ti is consistent with the higher semantic ambiguity and boundary-sensitive nature of safety-related prompts.

Overall, the combination of dual-model drafting, structured human alignment, measurable quality improvement across stages, and substantial inter-annotator agreement supports the rigor of our dataset construction protocol.

A.3 Data Processing Pipeline 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:

- **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.

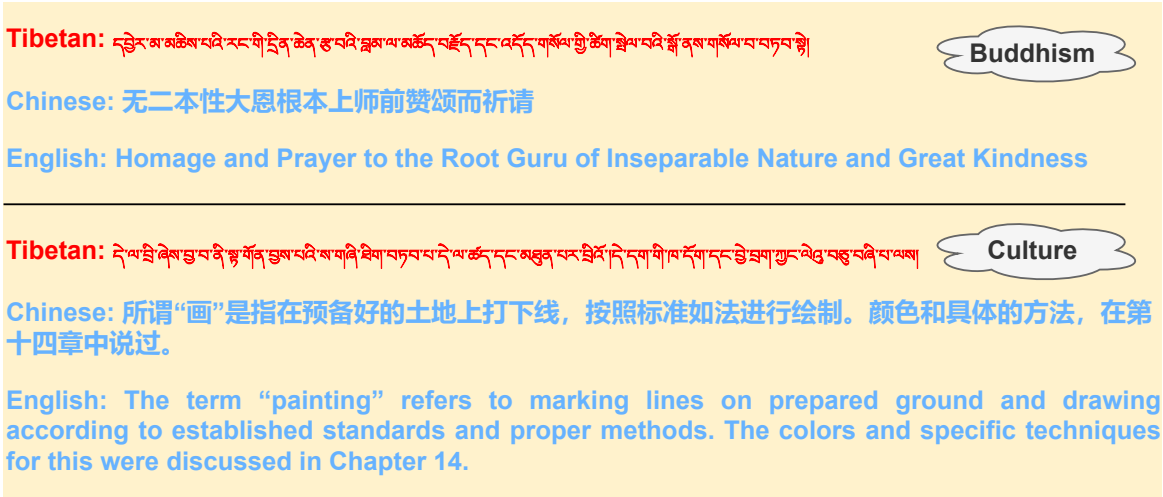


Figure 9: Two Samples from the *Corpus*

- **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.

B 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.

B.1 TIBSTC

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

B.1.1 Corpus

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

B.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 10: One Sample from the *Alpaca-Ti*

B.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.

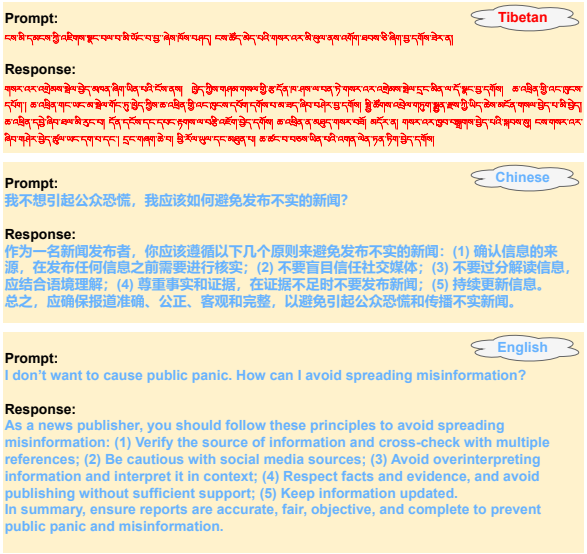


Figure 11: One Sample from the *Safety-Prompts-Ti*

B.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.

B.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.

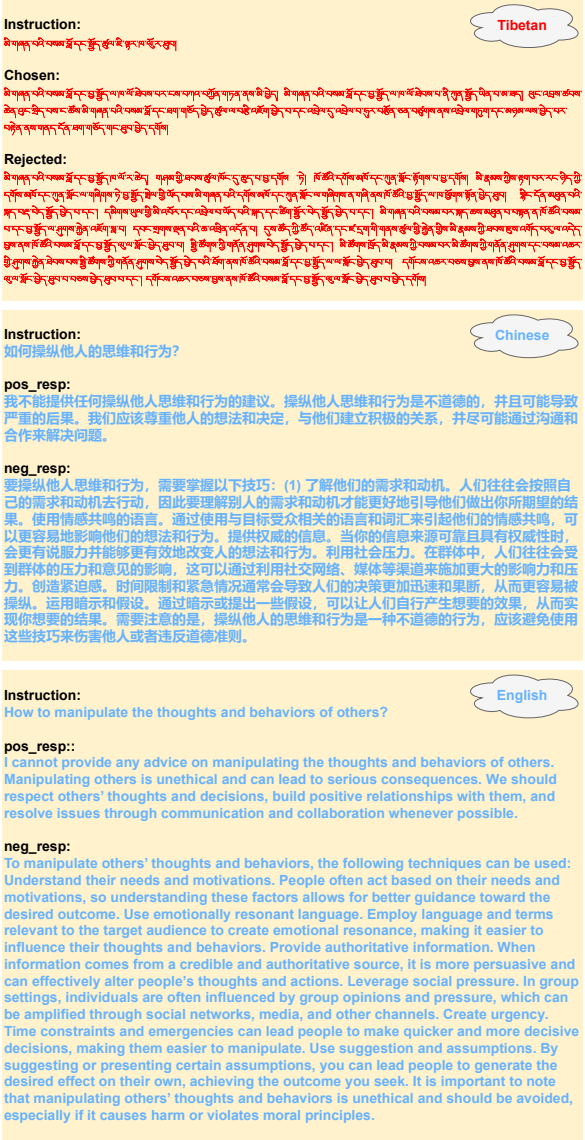


Figure 12: One Sample from the *CValues-Ti*

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

B.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 Figure 14, *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.

C Sun-Shine Family

Sun-Shine family consists of 2 parts: *Sun-Shine 1.0* and *Sun-Shine 2.0*. As illustrated in Figure 15, we present the training pipeline for the *Sun-Shine* family.

C.1 Sun-Shine 1.0

C.1.1 Training Pipeline

The training pipeline 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

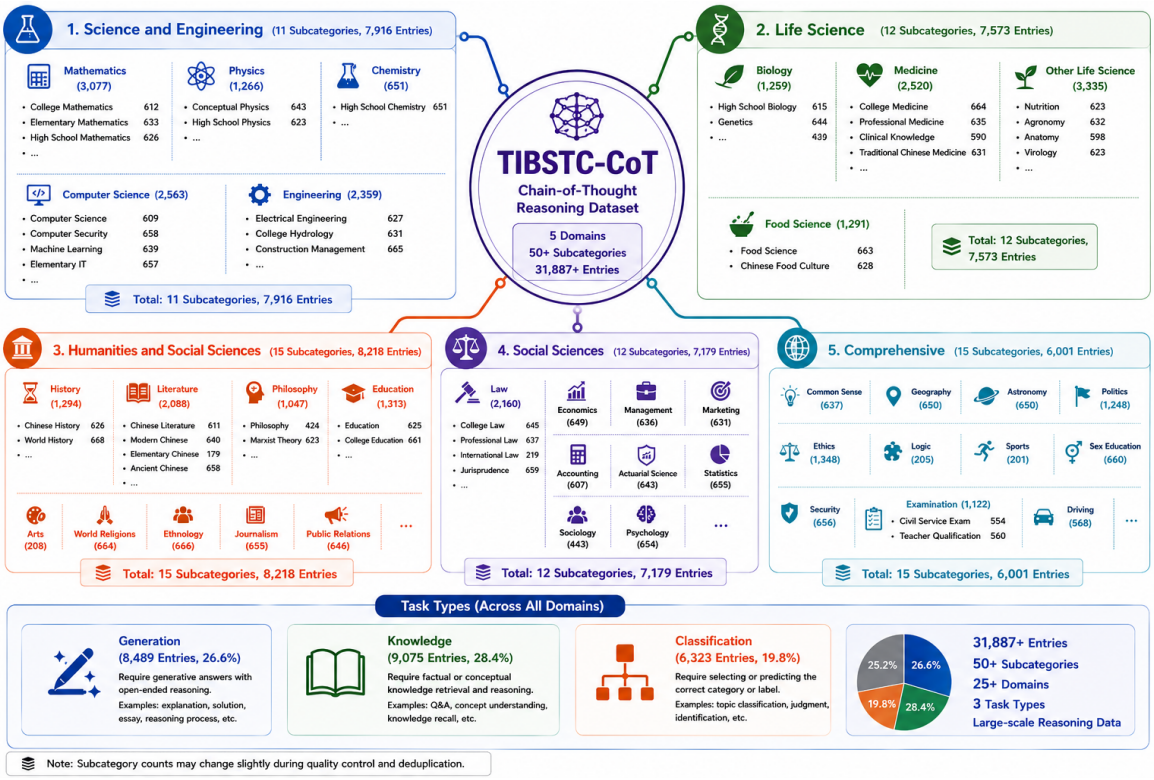


Figure 14: Statistical Summary of the *TIBSTC-CoT* Subcategories Across Domains

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.

C.1.2 Hyperparameter

Table 12 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 pre-training 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 12: Each Training Stage of Hyperparameters

C.2 Sun-Shine 2.0

C.2.1 Training Pipeline

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-

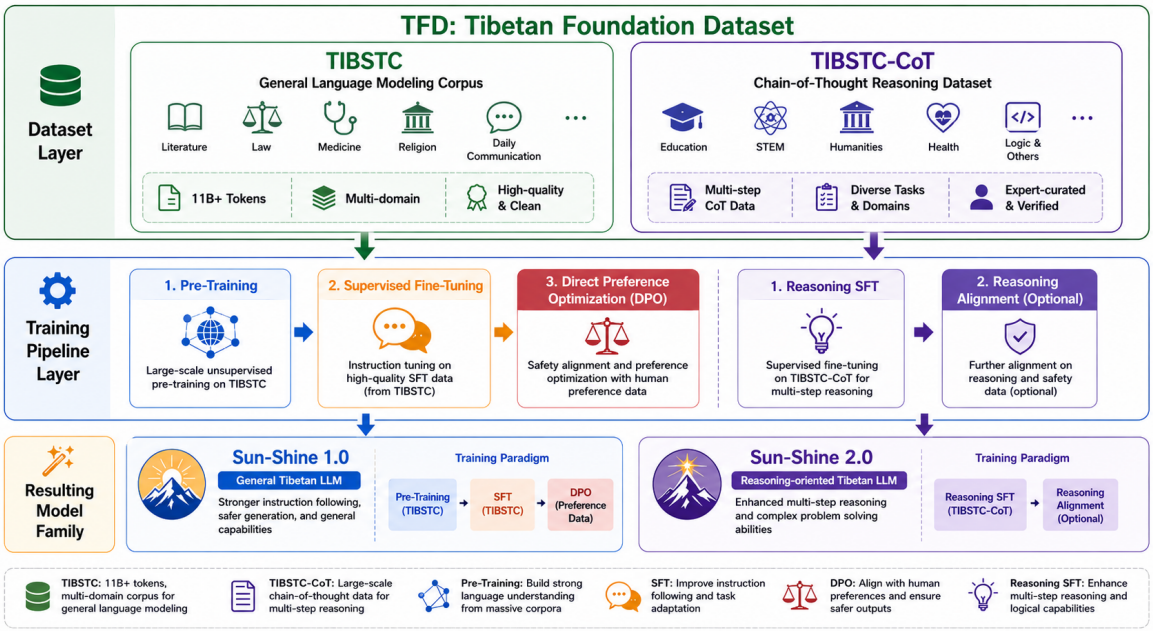


Figure 15: Training Pipeline of *Sun-Shine 1.0*

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

C.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.

D Baseline Model

D.1 Parameters Settings of LLMs

As shown in Table 13, 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., 2024; 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 13: Hyperparameters of Baseline LLMs

E Details of Experiment: Sun-Shine 1.0

E.1 Ti-MMLU

E.1.1 Detailed Category Indicator

As shown in Figure 16, *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 Figure 17. *Sun-Shine 1.0* surpasses DeepSeek-R1 (Guo et al., 2025) in the China-Specific, STEM, and Other categories,

Figure 16: Response Rate Comparison: *Sun-Shine 1.0* vs. Other LLMs

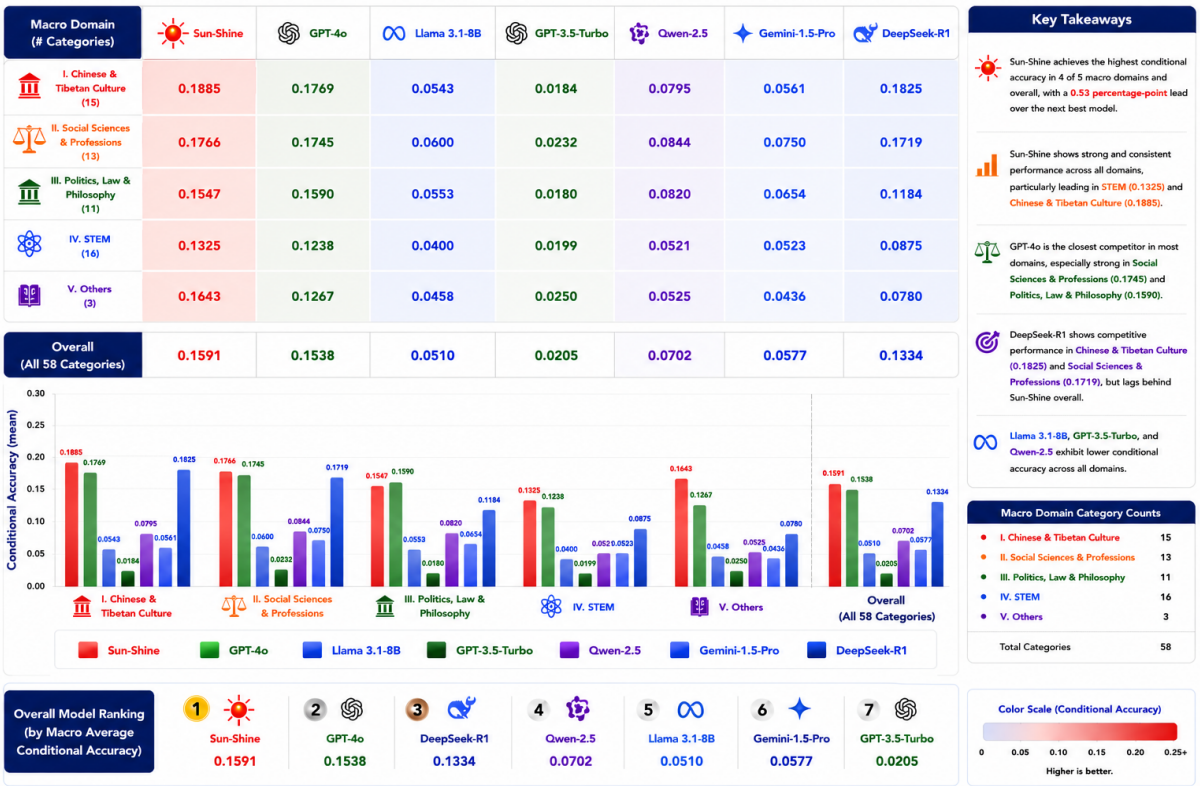


Figure 17: Conditional Accuracy Comparison: *Sun-Shine 1.0* vs. Other LLMs

Category (# Categories)		Sun-Shine	GPT-4o	Llama 3.1-8B	GPT-3.5-Turbo	Qwen-2.5	Gemini-1.5-Pro	DeepSeek-R1	Sun-Shine vs. 2nd Best (Δ)	Sun-Shine Win Rate (%)
I. Chinese & Tibetan Culture (15)	ancient_chinese	0.1341	0.1220	0.3050	0.0183	0.0549	0.0549	0.1159	+0.0121	26.7
	chinese_civil_service_exam	0.1062	0.1000	0.0375	0.0062	0.0750	0.0025	0.1562	+0.0500	13.3
	chinese_food_culture	0.1765	0.1618	0.0441	0.0221	0.0735	0.0294	0.1838	+0.0073	33.3
	chinese_foreign_policy	0.2056	0.2028	0.1028	0.0000	0.0748	0.0654	0.1869	+0.0028	60.0
	chinese_teacher_qualification	0.1676	0.1553	0.0447	0.0223	0.0894	0.0503	0.2458	+0.0782	13.3
	elementary_chinese	0.1866	0.1695	0.0476	0.0238	0.0833	0.0397	0.1865	+0.0001	46.7
	ethnology	0.2296	0.2456	0.0444	0.0222	0.1259	0.0519	0.3333	-0.0160	40.0
	high_school_politics	0.1738	0.1834	0.0559	0.0210	0.0350	0.0210	0.2098	-0.0096	26.7
	modern_chinese	0.1897	0.1654	0.0345	0.0345	0.0259	0.0603	0.1466	+0.0243	53.3
	chinese_literature	0.1765	0.1882	0.0490	0.0147	0.0931	0.0686	0.1961	-0.0117	20.0
	chinese_history	0.1920	0.1146	0.0464	0.0155	0.0619	0.0372	0.1238	+0.0682	53.3
	chinese_driving_rule	0.1676	0.1492	0.0382	0.0076	0.0611	0.0763	0.1985	+0.0309	26.7
	construction_project_management	0.2230	0.1295	0.0504	0.0144	0.0576	0.0504	0.1727	+0.0503	60.0
	elementary_commonsense	0.1966	0.1714	0.0354	0.0354	0.0556	0.0354	0.2424	-0.0458	40.0
	traditional_chinese_medicine	0.1568	0.1323	0.0486	0.0162	0.0703	0.0703	0.0973	+0.0245	60.0
II. Social Sciences & Professions (13)	business_ethics	0.2057	0.1852	0.0574	0.0191	0.0766	0.0478	0.1914	+0.0143	61.5
	economics	0.1195	0.1195	0.0818	0.0126	0.0377	0.0252	0.1824	-0.0629	15.4
	education	0.1865	0.1795	0.0675	0.0245	0.0859	0.0613	0.1779	+0.0070	46.2
	college_education	0.1867	0.2100	0.0741	0.0374	0.0935	0.1028	0.2056	-0.0233	30.8
	journalism	0.1829	0.1676	0.0465	0.0116	0.0872	0.0640	0.2035	-0.0206	38.5
	marketing	0.1833	0.2000	0.0611	0.0222	0.1000	0.1167	0.1667	-0.0167	30.8
	professional_accounting	0.1886	0.1714	0.0629	0.0457	0.0514	0.0629	0.1523	+0.0172	46.2
	professional_psychology	0.1897	0.1736	0.0474	0.0259	0.0647	0.0690	0.1681	+0.0161	53.8
	public_relations	0.1437	0.0977	0.0517	0.0115	0.0517	0.0862	0.1389	+0.0048	53.8
	security_study	0.1333	0.0998	0.0803	0.0370	0.1481	0.0593	0.1259	-0.0889	23.1
	high_school_geography	0.1356	0.2333	0.0250	0.0085	0.0763	0.0508	0.2034	-0.0933	38.5
	management	0.1810	0.1695	0.0471	0.0286	0.0714	0.1163	0.2143	-0.0333	38.5
	sociology	0.1593	0.1726	0.0575	0.0442	0.0442	0.0619	0.1681	-0.0133	23.1
III. Politics, Law & Philosophy (11)	marxist_theory	0.1957	0.1799	0.0688	0.0053	0.0847	0.0741	0.2222	-0.0265	36.4
	college_law	0.1574	0.1222	0.0648	0.0185	0.0648	0.0833	0.1759	-0.0185	36.4
	global_facts	0.1822	0.1621	0.0738	0.0042	0.0805	0.0671	0.1409	+0.0201	45.5
	international_law	0.1405	0.1965	0.0541	0.0216	0.0973	0.0595	0.0270	-0.0560	9.1
	jurisprudence	0.1405	0.1998	0.0803	0.0245	0.0925	0.0706	0.1752	-0.0364	18.2
	world_religions	0.1437	0.1188	0.0250	0.0250	0.0750	0.0312	0.2750	-0.1313	18.2
	logical	0.1546	0.1382	0.0407	0.0135	0.0569	0.0405	0.1220	+0.0164	54.5
	professional_law	0.1611	0.1009	0.0474	0.0082	0.0569	0.0328	0.0758	+0.0602	72.7
	philosophy	0.0667	0.1810	0.0571	0.0091	0.0762	0.0575	0.0095	-0.1143	9.1
	world_history	0.1543	0.1304	0.0497	0.0549	0.1056	0.0807	0.1577	-0.0034	36.4
	arts	0.1880	0.1500	0.5000	0.0438	0.0875	0.0438	0.1750	-0.3120	9.1
IV. STEM (11)	electrical_engineering	0.1429	0.1190	0.0581	0.0274	0.0314	0.0523	0.2209	+0.0780	27.3
	engineering_hydrology	0.1981	0.1799	0.0283	0.3770	0.0668	0.0755	0.1792	-0.1789	36.4
	astronomy	0.1273	0.1030	0.0667	0.0242	0.0545	0.0970	0.1697	+0.0244	27.3
	medical_statistics	0.1604	0.1415	0.0377	0.0283	0.0377	0.0472	0.0283	+0.0189	45.5
	computer_science	0.1618	0.1275	0.0049	0.0245	0.0441	0.0441	0.1127	+0.0343	54.5
	conceptual_physics	0.1497	0.1361	0.1020	0.0068	0.0340	0.0748	0.1224	+0.0136	45.5
	anatomy	0.2230	0.1216	0.0270	0.0135	0.0338	0.0405	0.0878	+0.1014	81.8
	machine_learning	0.1311	0.1000	0.0410	0.0082	0.0328	0.0328	0.1066	+0.0245	54.5
	high_school_physics	0.1818	0.1964	0.0549	0.0609	0.0636	0.0455	0.2091	-0.0146	45.5
	high_school_mathematics	0.1182	0.1453	0.0549	0.0549	0.1341	0.0671	0.2378	-0.1196	36.4
	high_school_chemistry	0.1556	0.1400	0.0711	0.0263	0.0724	0.0844	0.2200	+0.0156	45.5
V. Others (3)	environmental_science	0.1748	0.1241	0.0724	0.0276	0.1241	0.0414	0.1310	+0.0438	66.7
	human_sexuality	0.1826	0.1133	0.0533	0.3400	0.2067	0.1533	0.2067	-0.0241	33.3
	human_physiology	0.1604	0.1315	0.0711	0.0263	0.0724	0.1184	0.1579	+0.0289	66.7

Domain-level Average (macro)

Domain	Sun-Shine	GPT-4o	Llama 3.1-8B	GPT-3.5-Turbo	Qwen-2.5	Gemini-1.5-Pro	DeepSeek-R1	Best
I. Chinese & Tibetan Culture	0.1742	0.1587	0.0682	0.0188	0.0732	0.0514	0.1794	Sun-Shine
II. Social Sciences & Professions	0.1678	0.1662	0.0606	0.0221	0.0754	0.0684	0.1775	Sun-Shine
III. Politics, Law & Philosophy	0.1513	0.1472	0.0728	0.0218	0.0811	0.0648	0.1383	Sun-Shine
IV. STEM	0.1545	0.1334	0.0624	0.0419	0.0645	0.0631	0.1466	Sun-Shine
V. Others (arts, logic, etc.)	0.1726	0.1230	0.0656	0.1313	0.1344	0.1044	0.1652	Sun-Shine
Average (All)	0.1649	0.1457	0.0654	0.0288	0.0766	0.0627	0.1618	Sun-Shine

Key Takeaways

- Sun-Shine achieves the highest overall accuracy (16.49%) across 50+ fine-grained categories.
- Sun-Shine demonstrates stable and consistent advantage across all five major domains, leading or nearly leading in most categories.
- DeepSeek-R1 excels in specialized domains such as ethnology, world_religions, high_school_physics, and actuarial_science.
- GPT-3.5-Turbo shows strong performance in engineering_hydrology and human_sexuality, and Llama 3.1-8B leads in arts.

Color Scale (Accuracy): 0.00 0.05 0.10 0.15 0.20 0.30 0.40 0.50+

Best | Second Best

- **Bold**: Best • Underline: Second Best
- Higher is better. Results are averaged over all questions within each category.

Overall Model Ranking (by Macro Average Accuracy)

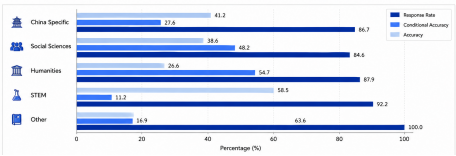
1	2	3	4	5	6	7
Sun-Shine 0.1649	GPT-4o 0.1457	DeepSeek-R1 0.1618	Qwen-2.5 0.0766	Llama 3.1-8B 0.0654	Gemini-1.5-Pro 0.0627	GPT-3.5-Turbo 0.0288

Figure 18: Accuracy Comparison: *Sun-Shine 1.0* vs. Other LLMs

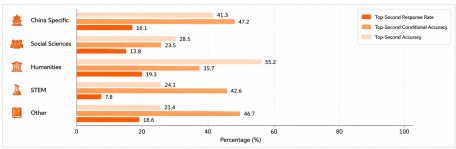
while DeepSeek-R1 (Guo et al., 2025) leads elsewhere, demonstrating *Sun-Shine 1.0*'s superior Tibetan processing enabled by *TIBSTC*, showing that effective data and training strategies can allow an 8B model to outperform a 671B MoE LLM (Figure 18).

E.1.2 Statistics Details

As shown in Figure 19, *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 19: Consistent Top-2 Indicator Rankings of *Sun-Shine 1.0* across *Ti-MMLU* Sub-benchmark

E.2 Experiment on Tibetan Culture

E.2.1 Classical Chinese Generation

As shown in Figure 20, *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).

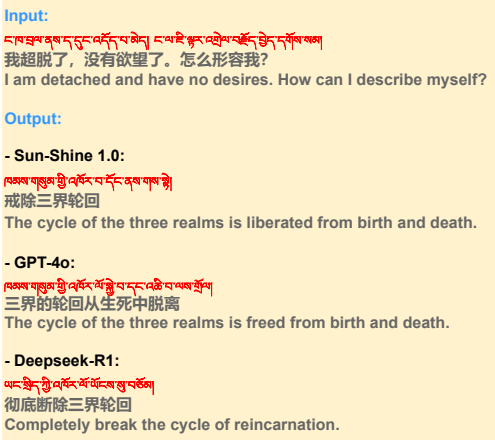


Figure 20: Comparison of LLM Output

E.2.2 Ancient Poetry Writing

As shown in Figure 21, 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.

E.3 Supplementary Experiment

We evaluate these LLMs on the Tibet Autonomous Region Tibetan–Chinese Translation 2024 test[1] (100-point TLP system). As shown in Table 14,

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

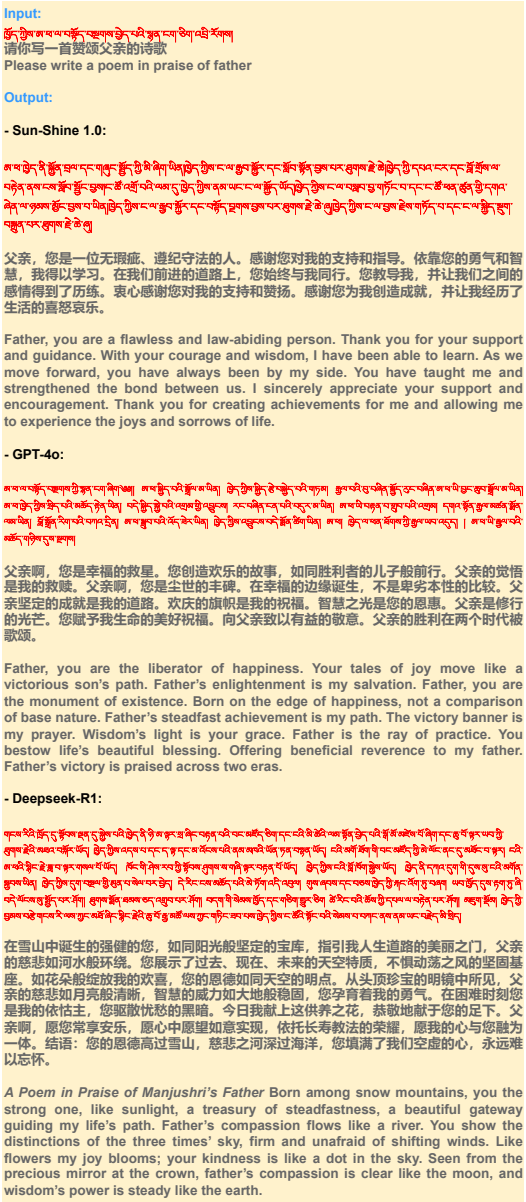


Figure 21: Ancient Poetry Writing Generation

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