

TLUE: A Tibetan Language Understanding Evaluation Benchmark

Fan Gao^{1,*}, Cheng Huang^{1,3,4,*}, Nyima Tashi^{2,†}, Thupten Tsering^{1,5,†}
Yutong Liu¹, Ban Ma-bao¹, Xiangxiang Wang¹, Xiao Feng¹, Renzeg Duojie²
Gadeng Luosang², Rinchen Dongrub², Dorje Tashi², Hao Wang^{1,6}, Yongbin Yu^{1,†}

¹University of Electronic Science and Technology of China, ²Tibet University

³University of Texas Southwestern Medical Center, ⁴Southern Methodist University

⁵The State Key Laboratory of Tibetan Intelligence, ⁶University of Connecticut

Github: <https://github.com/Vicentvankor/TLUE>

* Equal Contribution, † Corresponding Authors

Abstract

Large language models have made tremendous progress in recent years, but low-resource languages, like Tibetan, remain significantly underrepresented in their evaluation. Despite Tibetan being spoken by over seven million people, it has largely been neglected in the development and assessment of large language models. To address this gap, we present a Tibetan Language Understanding Evaluation Benchmark, **TLUE**, the first large-scale benchmark for measuring the proficiency of LLMs in the Tibetan language. **TLUE** comprises two major components: a comprehensive multi-task understanding benchmark spanning 5 domains and 67 subdomains, and a safety benchmark encompassing 7 subdomains. Then, we evaluate a diverse set of state-of-the-art large language models. Experimental results demonstrate that most large language models perform below the random baseline, highlighting the considerable challenges they face in Tibetan language processing. **TLUE** provides a crucial foundation for advancing future research in Tibetan language understanding and highlights the importance of promoting greater inclusivity in the development of large language models.

1 Introduction

Large language models (LLMs) have made significant strides in natural language understanding, excelling in high-resource languages like English and Chinese through extensive datasets, advanced architectures, and benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019). However, many mainstream LLMs, like GPT (OpenAI, 2024; Achiam et al., 2023), Claude (Anthropic, 2024), Gemini (DeepMind, 2024), LLaMA (Dubey et al., 2024), Qwen (Qwen Team, 2024) and DeepSeek (DeepSeek-AI et al., 2025, 2024) largely overlook low-resource languages, such as Tibetan, which limits the inclusivity and fairness of AI systems and restricts their applicability to underserved

language communities (Bender et al., 2021; Huang et al., 2025; Gao et al., 2025; Liu et al., 2025b). For Tibetan, it is one of the languages included, spoken by over seven million people, which has unique grammatical features and limited annotated data (National Bureau of Statistics of China, 2021). Despite its cultural importance, Tibetan is underrepresented in current LLM research and benchmarks (An, 2023; Liu et al., 2025a, 2024).

Current language understanding benchmarks, including GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), and BIG-bench (Srivastava et al., 2022), focus on high-resource languages, leaving low-resource languages like Tibetan without tailored evaluation frameworks, including benchmarks from China (Li et al., 2023; Zhang et al., 2023). Even existing methods do not adequately capture Tibetan's linguistic intricacies (Liu et al., 2022; Lv et al., 2025). So, in this paper, we introduce the **TLUE**, a Tibetan Language Understanding Evaluation benchmark, the first large-scale benchmark for Tibetan, designed to address the unique challenges of low-resource language evaluation. Furthermore, we selected several state-of-the-art LLMs for training and evaluation, and the experimental results demonstrate that **TLUE** effectively exposes the limitations of current models in handling Tibetan language tasks. Most LLMs show significantly lower performance on **TLUE** compared to benchmarks in high-resource languages, particularly struggling with tasks requiring deep linguistic understanding and domain-specific knowledge. This highlights the urgent need for more inclusive pretraining data and architecture adaptations tailored to low-resource languages like Tibetan. Our findings validate **TLUE** as a robust tool for evaluating and guiding the development of more equitable language models.

All in all, the main contributions of our work are summarized as follows:

- We developed and publicly released **TLUE**, the first large-scale Tibetan Language Understanding Evaluation benchmark. It is designed to fill the gap in existing resources by providing a comprehensive suite of tasks that span 67 knowledge-based subdomains and 7 safety-critical categories, addressing both general understanding and ethical alignment in Tibetan.
- We evaluated several state-of-the-art LLMs, including GPT-4o (Achiam et al., 2023) and GPT-O1-mini (OpenAI, 2024), Claude-3.5-Sonnet (Anthropic, 2024), Gemini-1.5 (DeepMind, 2024), LlaMA-3.1 (Dubey et al., 2024), Qwen-2.5 (Qwen Team, 2024), DeepSeek-V3 (DeepSeek-AI et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025), to systematically assess their capabilities in both general Tibetan language understanding and safety-oriented tasks. These LLMs represent a diverse set of architectures and training strategies, enabling a broad comparison of their performance under low-resource conditions.
- Our qualitative and quantitative analyses reveal substantial limitations in current LLMs when applied to Tibetan. Most models performed significantly below expected baselines, even falling below random choice levels on certain tasks, particularly in domains requiring complex reasoning or cultural sensitivity. These findings underscore the urgent need for research into more inclusive, linguistically-aware model development tailored to low-resource languages.

2 Related Work

2.1 Language Understanding Benchmark

LLMs have been extensively evaluated on multilingual and domain-specific benchmarks, such as X TREME (Hu et al., 2020), XGLUE (Liang et al., 2020), and MASSIVE (FitzGerald et al., 2022). However, these benchmarks predominantly focus on high-resource languages, with little attention given to low-resource languages like Tibetan. C MMLU (Li et al., 2023) and SafetyBench (Zhang et al., 2023) introduced large-scale Chinese language evaluations, but no equivalent benchmark existed for Tibetan.

2.2 Low Resource Language Evaluation

Several efforts have been made to extend LLM evaluation to low-resource languages. IndicGLUE (Kakwani et al., 2020), and INDICGENBENCH (Singh et al., 2024) focus on Indic languages, while projects such as AmericasNLP (Mager et al., 2021) and Masakhane (Nekoto et al., 2020) target indigenous and African languages. Despite these advances, Tibetan remains significantly underrepresented in LLM evaluation.

2.3 Safety and Ethical Limitations in LLM

Recent work has also examined the safety and ethical risks of LLMs, particularly in multilingual and low-resource contexts. Studies such as SafetyBench (Zhang et al., 2023) and HolisticEval (Liang et al., 2023) assess biases, fairness, and robustness in AI systems. However, these benchmarks largely exclude Tibetan and other low-resource languages, leaving significant gaps in understanding how safety concerns manifest in such linguistic environments.

2.4 Motivation

Based on the current progress of LLM in minority languages, we propose the **TLUE**, filling the gap by providing a comprehensive Tibetan evaluation dataset across multiple domains and safety-related tasks. We referred to the two benchmarks, C MMLU (Li et al., 2023) and SafetyBench (Zhang et al., 2023), and asked Tibetan language experts to manually translate and verify them. In summary, the combination of **Ti-MMLU** and **Ti-SafetyBench** forms **TLUE**. **TLUE** can not only enable systematic evaluation of model performance in a challenging low-resource setting, but also introduces a dedicated safety evaluation suite for Tibetan, providing insights into model robustness and potential vulnerabilities in a low-resource language context.

By offering a comprehensive, multi-domain framework for both language understanding and safety evaluation, **TLUE** fills a critical gap in the current landscape of LLM benchmarks, enabling systematic assessment of Tibetan language models in a challenging low-resource setting.

3 TLUE

In designing **TLUE**, we adhere to several key principles to ensure a rigorous and meaningful evaluation of Tibetan LLMs.

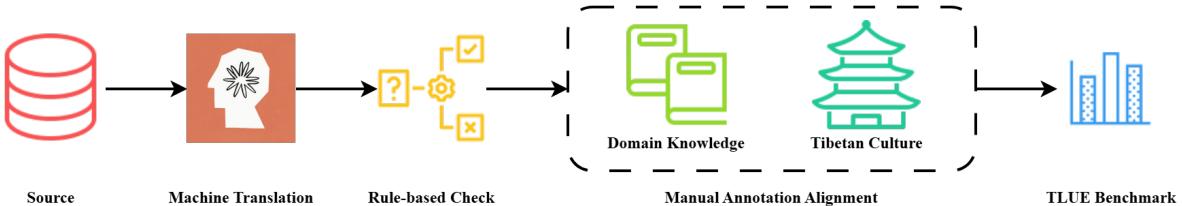


Figure 1: The Curation Process of **TLUE**

First, we align our benchmark with existing multilingual evaluation frameworks to facilitate direct comparison with established benchmarks. This allows researchers to assess the performance of Tibetan LLMs relative to models trained in other languages, providing a broader perspective on cross-lingual capabilities (Park et al., 2024).

Additionally, we ensure that the evaluation data is carefully curated and free from data leakage, maintaining the integrity of benchmark results and preventing unintended memorization effects (Doe and Smith, 2024).

3.1 Curation Process

As shown in Figure 1, the **TLUE** benchmark follows a structured multi-stage process to ensure the quality and cultural relevance of Tibetan evaluation data in the curation process. Source datasets, such as CMMLU (Li et al., 2023) and SafetyBench (Zhang et al., 2023), are first translated into Tibetan using Claude-3.5-Sonnet (Anthropic, 2024), which demonstrates strong Tibetan text generation capabilities. The translations undergo rule-based filtering to correct syntactic inconsistencies and structural errors.

Following this, human annotators, including 2 Tibetan language specialists and a team of 5 additional annotators, refine the dataset to preserve linguistic accuracy and classical Tibetan grammar. Annotators are compensated at an hourly rate of 28 USD, ensuring high-quality review and incentivizing skilled professionals.

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 **TLUE** remains a comprehensive and culturally adapted benchmark for Tibetan language model evaluation.

3.2 Size

TLUE consists of 22,963 evaluation questions, divided into **Ti-MMLU** for knowledge assessment

and **Ti-SafetyBench** for safety evaluation.

Ti-MMLU includes 11,528 multiple-choice questions across 67 subjects, covering social sciences, STEM, China-specific topics, humanities, and other domains, enabling a comprehensive evaluation of Tibetan language understanding. **Ti-SafetyBench** contains 11,435 questions across 7 safety categories, including ethics, bias, health, and privacy, designed to assess LLMs' ability to handle sensitive content.

While **Ti-MMLU** focuses on broad, multi-domain knowledge evaluation, **Ti-SafetyBench** targets high-precision safety assessments, together forming a balanced and comprehensive Tibetan language benchmark. The dataset sizes are summarized in Appendix A Figure 5, which illustrates the comprehensive breakdown of **Ti-MMLU** and **Ti-SafetyBench**, highlighting the wide range of subjects and safety categories covered. For more details of such two sub-benchmarks, they are shown in Appendix A.

3.3 Why Extend Existing Benchmarks?

Developing a benchmark for Tibetan presents challenges such as limited annotated corpora and linguistic expertise. **TLUE** extends existing benchmarks via translation, enhanced with human annotation and quality control. Key reasons include:

- **Enabling Cross-Lingual Comparability:** a translation-based benchmark provides parallel data to compare Tibetan model performance across languages, isolating linguistic effects from task knowledge.
- **Addressing Data Scarcity and Leveraging Established Designs:** due to limited Tibetan corpora, **TLUE** adapts reliable, pre-validated benchmarks like CMMLU (Li et al., 2023) and SafetyBench (Zhang et al., 2023), ensuring task diversity and alignment with Tibetan's linguistic and cultural traits.
- **Ensuring Quality through Controlled**

Adaptation: TLUE uses a multi-stage adaptation process (translation, filtering, alignment, annotation) to balance efficiency with linguistic precision, maintaining high benchmark standards. This structured approach provides a **high-quality, scalable** framework for assessing LLM performance in Tibetan, especially in low-resource settings.

4 Implementation

Our experiments cover a diverse range of LLMs, including both open-source and proprietary LLMs:

- **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 (Qwen Team, 2024), DeepSeek-V3 (DeepSeek-AI et al., 2024), and DeepSeek-R1 (DeepSeek-AI 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.
- **Proprietary LLM:** we also evaluate several proprietary LLMs, including GPT-4o (OpenAI, 2024), GPT-3.5-Turbo (Achiam et al., 2023), Claude-3.5-Sonnet (Anthropic, 2024), Gemini-1.5-Flash-Latest (DeepMind, 2024), and O1-Mini (OpenAI, 2024). These models serve as strong baselines for commercial LLM performance in Tibetan tasks.

4.1 Experimental Setup

We investigate several key aspects of LLM performance in Tibetan:

- We evaluate all LLMs on **Ti-MMLU** and **Ti-SafetyBench**, measuring their Tibetan multi-task understanding and safety capabilities.
- We assess the impact of language resource availability by comparing CMMLU (Li et al., 2023) and **Ti-MMLU** and extend this analysis to safety evaluation by comparing SafetyBench (Zhang et al., 2023) with **Ti-SafetyBench**.
- We explore the differences between reasoning-optimized LLMs and chat LLMs, examining whether reasoning enhancements improve low-resource language 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 (Qwen Team, 2024).

4.2 Evaluation

4.2.1 Evaluation Methodology

We employ zero-shot evaluation to simulate real-world conditions where LLMs must perform tasks without task-specific examples, providing a more accurate assessment of performance in Tibetan. Appendix Figure 7 illustrates the **TLUE** prompt format and example queries used in evaluation. To ensure consistency, we use the default temperature settings for each model during inference, avoiding bias from manual adjustments.

4.2.2 Evaluation Metrics

We evaluate LLMs' performance on **TLUE** using two metrics: **Response Rate**, **Accuracy** and **Conditional Accuracy**.

- **Response Rate** measures the proportion of valid responses.
- **Accuracy** tracks the proportion of correct answers.
- **Conditional Accuracy** reflects the LLM's performance when giving valid responses.

For both **Ti-MMLU** and **Ti-SafetyBench**, We use two methods for metric calculation:

- **Direct Answer Calculation** measures if the LLM selects a valid and correct answer.
- **Concern All Answer Calculation** evaluates the LLM's maximum performance by considering all valid options and selecting correct answers based on remaining choices.

These methods provide a thorough assessment of LLMs' Tibetan language performance. For more details of **Ti-MMLU** and **Ti-SafetyBench**, please refer to Appendix B Algorithm 1 and Appendix B Algorithm 2. We use abbreviated forms for some of the special names or evaluation metric names, as shown in Appendix Table 4.

5 Performance on TLUE

5.1 Performance on Ti-MMLU

As shown in Table 1, most LLMs perform below the random baseline (25%), underscoring the challenge of Tibetan language understanding. Claude-3.5-Sonnet (Anthropic, 2024) achieves the highest accuracy, surpassing the baseline by 10.6 percentage points in CAA. Among proprietary LLMs,

Method	LLM	Version	Ti-MMLU						Ti-SafetyBench							
			Avg.	STEM	Human	Social	Other	China	Avg.	OFF	UB	PH	MH	IA	EM	PP
DA	Claude	3.5-Sonnet	33.95	28.04	36.32	39.31	34.01	32.09	50.5	32.3	56.1	63.7	61.0	31.7	52.6	65.6
	Gemini	1.5-Flash	30.14	25.75	29.72	35.9	29.61	29.7	43.7	32.7	44.2	59.1	49.4	37.6	43.6	46.3
	DeepSeek	R1	15.74	13.87	14.24	18.06	13.94	18.58	24.3	20.5	37.2	25.8	20.4	16.9	29.2	16.7
	DeepSeek	V3	29.51	23.57	31.97	33.65	29.92	28.44	37.4	25.5	38.1	40.5	39.1	44.3	38.3	36.9
	GPT	4O	16.00	12.73	16.36	18.04	16.92	15.96	31.1	24.0	28.8	24.6	42.1	39.7	22.4	37.6
	GPT	3.5-Turbo	2.11	2.40	2.18	2.20	1.88	1.87	9.4	8.6	12.3	9.7	8.7	7.3	8.5	10.6
	GPT	O1-mini	6.14	6.15	6.33	7.17	6.06	4.98	10.9	11.1	17.0	9.7	9.5	7.0	10.9	9.5
	Qwen	2.5-32B	13.94	12.63	14.98	14.92	13.71	13.44	21.2	19.8	44.1	15.0	16.4	11.8	19.9	15.9
	Qwen	2.5-72B	7.27	6.07	7.98	7.52	7.74	7.02	21.9	19.9	37.6	17.6	20.3	18.1	17.8	18.6
	Qwen	2.5-7B	1.8	2.94	1.87	1.63	0.9	1.68	9.0	10.0	16.8	7.3	5.7	6.6	7.7	7.2
CAA	LlaMA	3.1-405B	25.08	23.88	24.25	25.58	27.62	24.07	43.5	36.8	31.4	46.0	52.2	50.4	44.8	46.2
	LlaMA	3.1-70B	23.73	23.16	23.2	26.2	24.65	21.45	37.0	32.2	37.3	30.0	40.9	44.0	34.2	39.1
	LlaMA	3.1-8B	5.47	5.48	5.56	5.99	5.46	4.86	9.9	9.7	10.9	9.4	10.1	8.8	10.5	9.9
	Claude	3.5-Sonnet	35.63	30.88	37.47	40.58	35.26	33.96	58.5	51.1	56.5	66.1	66.9	50.1	57.9	67.6
	Gemini	1.5-Flash	31.01	26.68	30.58	36.69	30.30	30.81	49.6	44.2	44.3	60.2	56.0	44.9	51.6	51.0
	DeepSeek	R1	27.45	21.01	25.99	32.72	26.44	31.08	46.8	42.9	45.7	51.1	50.0	45.7	55.8	33.9
	DeepSeek	V3	32.16	27.03	34.58	36.26	32.00	30.94	48.3	44.3	44.9	51.1	46.4	55.6	51.8	43.6
	GPT	4O	17.51	14.25	17.71	19.69	18.46	17.45	32.9	28.7	30.1	25.5	42.9	40.9	24.8	38.7
	GPT	3.5-Turbo	3.42	3.82	3.35	3.68	3.09	3.16	11.6	12.5	16.0	11.4	10.3	8.2	10.8	11.8
	GPT	O1-mini	9.67	9.69	9.80	10.14	9.68	9.02	15.1	16.3	23.1	13.5	13.4	9.9	15.4	11.7
-	Qwen	2.5-32B	18.56	16.66	20.3	19.72	17.47	18.67	34.1	34.8	51.6	30.9	30.7	25.5	31.3	30.6
	Qwen	2.5-72B	16.50	15.73	17.88	17.00	15.84	16.04	30.6	36.0	45.2	28.5	24.4	22.6	28.6	24.9
-	Qwen	2.5-7B	14.59	13.92	13.66	16.34	14.57	14.46	30.2	35.2	39.3	27.1	25.9	23.2	31.2	25.8
	LlaMA	3.1-405B	25.28	24.10	24.50	25.87	27.73	24.22	43.9	37.6	31.7	46.5	52.4	50.8	45.1	46.4
	LlaMA	3.1-70B	23.79	23.22	23.24	26.31	24.65	21.52	37.4	32.9	37.3	30.5	41.1	44.5	35.0	39.3
	LlaMA	3.1-8B	7.44	7.95	7.54	7.38	7.41	6.92	12.0	12.4	12.6	11.9	12.3	10.1	13.0	11.5
	Random	-	25.00	25.00	25.00	25.00	25.00	25.00	36.7	34.5	49.9	27.6	49.5	28.0	26.0	36.4

Table 1: Accuracy Performance of LLMs on the **TLUE** based on CAA and DA ($\times 100\%$)

the GPT (Achiam et al., 2023; OpenAI, 2024) series underperforms, whereas open-source models like DeepSeek-V3 (DeepSeek-AI et al., 2024) exceed the baseline. For reasoning-optimized LLMs, DeepSeek-R1 (DeepSeek-AI et al., 2025) surpasses random performance in CAA but underperforms in DA, highlighting Tibetan’s difficulty as a low-resource language.

The performance gap between proprietary and open-source models is minimal, with DeepSeek-V3 (DeepSeek-AI et al., 2024) and Gemini-1.5-Flash (DeepMind, 2024) performing comparably to Claude-3.5-Sonnet (Anthropic, 2024). Notably, GPT-3.5-Turbo (Achiam et al., 2023; OpenAI, 2024) underperforms relative to LlaMA-3.1-8B (Dubey et al., 2024), suggesting that pretraining data, adaptation, and optimization significantly impact Tibetan language performance.

STEM remains the most challenging category, whereas most models perform best in Social Sciences. However, LlaMA-3.1-405B (Dubey et al., 2024) excels in Other, and Qwen-2.5-72b (Qwen Team, 2024) in Humanities, indicating that structured reasoning tasks in Tibetan pose challenges for LLMs, while general knowledge and socially

contextualized tasks are relatively easier.

The detailed results can be found in Appendix Table 15 and Appendix Table 16.

5.2 Performance on Ti-SafetyBench

As shown in Table 1, most LLMs perform below the random baseline (36.7%), indicating significant challenges in aligning models with safety principles in Tibetan. Claude-3.5-Sonnet (Anthropic, 2024) achieves the highest accuracy (58.5% in CAA), whereas GPT-4o (Achiam et al., 2023) and GPT-3.5-Turbo (Achiam et al., 2023) underperform, with the latter significantly below random.

Among open-source models, DeepSeek-V3 (DeepSeek-AI et al., 2024) surpasses the baseline in both DA and CAA, demonstrating strong safety alignment. LlaMA-3.1-405B (Dubey et al., 2024) approaches the baseline, while smaller models like LlaMA-3.1-8B (Dubey et al., 2024) and Qwen-2.5-72B (Qwen Team, 2024) perform poorly. Reasoning-optimized models such as DeepSeek-R1 (DeepSeek-AI et al., 2025) excel in CAA but struggle in DA, whereas O1-mini (OpenAI, 2024) consistently underperforms.

These findings suggest that safety alignment in

Tibetan remains challenging, with pretraining data, adaptation strategies, and model architecture playing a more critical role than accessibility.

5.3 High-Resource vs. Low-Resource

To examine the performance disparity between high-resource and low-resource languages, we compare model accuracy on CMMLU (Li et al., 2023) and Ti-MMLU using the CAA evaluation. Figure 2 illustrates performance differences across domains, while Appendix Table 6 presents detailed accuracy comparisons.

5.3.1 Performance Degradation

All LLMs experience substantial accuracy drops from CMMLU (Li et al., 2023) to Ti-MMLU. Qwen-2.5-72B (Qwen Team, 2024) declines from 84.70% → 16.50%, GPT-4 (Achiam et al., 2023) from 68.90% → 17.51%, and ChatGPT (Achiam et al., 2023; OpenAI, 2024) from 53.22% → 3.42%, falling below the random baseline (25%), underscoring the difficulty of adapting LLMs to low-resource languages.

5.3.2 Domain Performance Shifts

While accuracy declines across all domains, the highest-performing categories differ between languages. In CMMLU (Li et al., 2023), LLMs excel in "Other" or "China-specific" categories, whereas in Ti-MMLU, "Social Sciences" and "Humanities" are strongest. For instance, Qwen-2.5-72B (Qwen Team, 2024) achieves the highest accuracy in "Other" (87.35%) on CMMLU (Li et al., 2023) but shifts to "Humanities" (17.88%) on Ti-MMLU. LLaMA-3.1-70B (Dubey et al., 2024) transitions from "Other" (74.72%) in CMMLU (Li et al., 2023) to "Social Sciences" (26.31%) in Ti-MMLU. GPT-4's (Achiam et al., 2023) best-performing category changes from "Other" (73.16%) in CMMLU (Li et al., 2023) to "Social Sciences" (19.69%) in Ti-MMLU (GPT-4o (Achiam et al., 2023)). These shifts suggest that models retain general knowledge better but struggle with structured reasoning in Tibetan.

5.3.3 STEM as the Weakest Domain

STEM remains the most challenging domain in both benchmarks. It consistently ranks lowest in CMMLU (Li et al., 2023) and Ti-MMLU, except for LLaMA-3.1-70B (Dubey et al., 2024), which performs worst in "China-specific" tasks. This indicates significant difficulties in mathematical and technical reasoning in Tibetan.

5.3.4 Ranking Inconsistencies

LLM rankings in CMMLU (Li et al., 2023) do not consistently translate to Ti-MMLU. For instance, LLaMA-3.1-70B (Dubey et al., 2024) ranks below Qwen-2.5-72B (Qwen Team, 2024) in CMMLU (Li et al., 2023) but surpasses it in Ti-MMLU, indicating that strong performance in high-resource languages does not necessarily predict effectiveness in low-resource settings.

5.3.5 Implication: Low-Resource Adaptation

The substantial performance gap underscores the need for improved Tibetan data coverage in pre-training and fine-tuning. While LLMs excel in high-resource languages, their struggles in Tibetan highlight the critical role of data availability and adaptation strategies in enhancing low-resource language understanding.

5.4 Safety in Low-Resource Languages

To evaluate LLMs' safety in low-resource languages, we compare performance on SafetyBench (Zhang et al., 2023) and Ti-SafetyBench using CAA. Figure 3 illustrates accuracy drops across safety categories, while Appendix Table 7 provides detailed performance breakdowns. These results highlight the challenges of maintaining safety alignment in Tibetan.

5.4.1 Performance Degradation in Tibetan

LLMs show a significant decline in safety alignment on Ti-SafetyBench, with most falling below the random baseline (36.7%). GPT-4 (Achiam et al., 2023) achieves 89.2% on SafetyBench (Zhang et al., 2023), but GPT-4o (Achiam et al., 2023), despite being a more advanced model, drops to 32.9% on Ti-SafetyBench, particularly in Ethical and Moral Reasoning (92.6% → 24.8%). GPT-3.5-Turbo (Achiam et al., 2023) experiences a sharper decline (80.4% → 11.6%), while Qwen-chat-7B (Qwen Team, 2024) decreases from 77.4% to 30.2%. These findings suggest that LLMs struggle with safety alignment in low-resource languages, often underperforming compared to random selection.

5.4.2 Variability in Safety Categories

Performance degradation varies across safety categories. GPT-4 (Achiam et al., 2023) excels in PH and EM on SafetyBench (Zhang et al., 2023), whereas GPT-4o (Achiam et al., 2023), despite

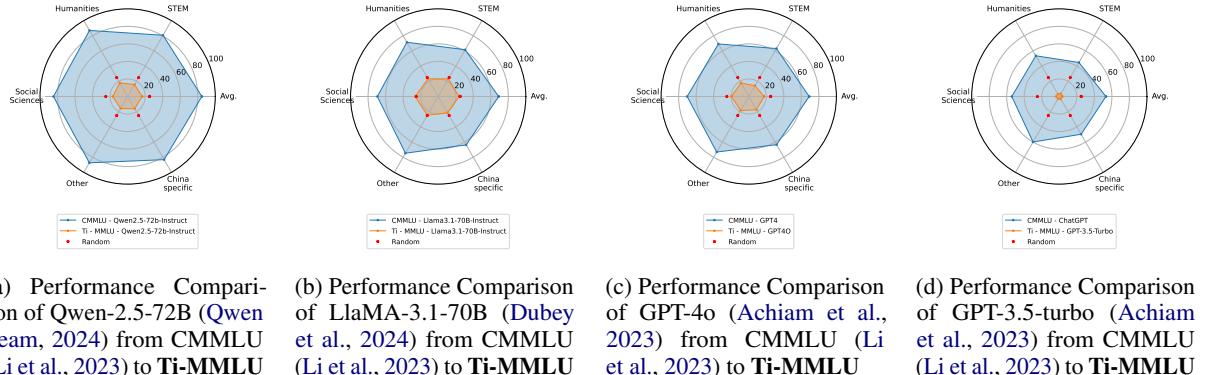


Figure 2: LLMs’ Performance Degradation from CMMLU (Li et al., 2023) to **Ti-MMLU**

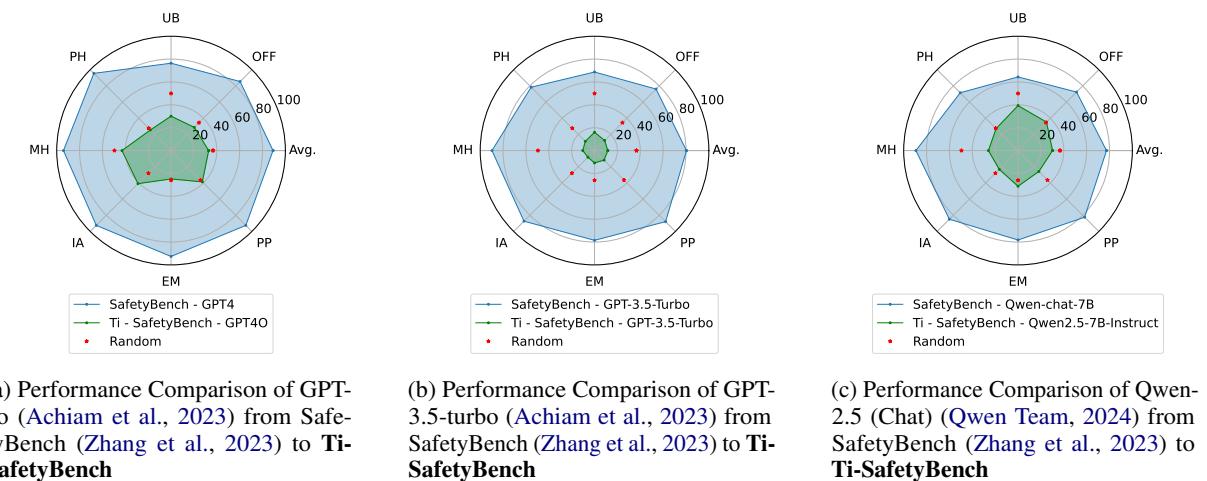


Figure 3: LLMs’ Performance Degradation from SafetyBench (Zhang et al., 2023) to **Ti-SafetyBench**

being more advanced, performs best in IA on **Ti-SafetyBench**, highlighting the challenges of transferring safety alignment to Tibetan.

Similarly, Qwen-chat-7B (Qwen Team, 2024) outperforms random in EM and IA on SafetyBench (Zhang et al., 2023), but Qwen-2.5-7B (Qwen Team, 2024) retains its best performance only in EM on **Ti-SafetyBench**. Despite updates, both models show significant degradation in Tibetan.

GPT-3.5-Turbo (Achiam et al., 2023) falls well below the random baseline on **Ti-SafetyBench**, while GPT-4o (Achiam et al., 2023) and Qwen-2.5-7B (Qwen Team, 2024) exceed or match the baseline in some categories. These results suggest that factors such as enhanced pretraining data, multilingual adaptation, and fine-tuning may contribute to better transferability, but safety alignment in low-resource settings remains a significant challenge.

5.4.3 Low-Resource Safety Alignment

The sharp decline in performance on **Ti-SafetyBench** underscores the challenge of main-

taining safety alignment in Tibetan. While models perform well on SafetyBench (Zhang et al., 2023), most fall below or barely match the baseline on **Ti-SafetyBench**. Even top-performing models struggle with safety generalization, highlighting the need for targeted fine-tuning and enhanced multilingual adaptation in low-resource environments.

5.5 Reasoning vs. Chat

We compare reasoning-optimized and chat LLMs on **TLUE**, analyzing response behavior, accuracy, and conditional accuracy. The results are summarized in Appendix C Table 8, Appendix C Table 9 and Appendix C Table 12, with a broader comparison provided on **Ti-MMLU** (Appendix C Table 10) and **Ti-SafetyBench** (Appendix C Table 11), covering overall accuracy across knowledge and safety domains.

5.5.1 Response Rate and Output Behavior

Reasoning LLMs, compared to chat LLMs, tend to evaluate all answer choices before selecting the cor-

rect one, resulting in higher response rates for CAA than DA (Appendix C Table 8 and Appendix C Table 9). This behavior stems from a step-by-step evaluation process. When evaluating reasoning models, intermediate reasoning steps are filtered, and only the final output is considered, as outlined in Appendix B Algorithm 3 and Appendix B Algorithm 4.

5.5.2 Impact: Response Rate on Performance

Lower response rates notably affect reasoning models’ performance, highlighting the inherent difficulty in responding to Tibetan prompts. The gap between DA and CAA accuracy underscores this challenge, with DA tasks requiring direct answer generation (Appendix C Table 8 and Appendix C Table 9). Reasoning models, unlike chat models, struggle more with Tibetan prompts due to their emphasis on logical inference over language generation.

5.5.3 Generalization of Reasoning Models in Low-Resource Languages

As shown in Appendix C Table 12, DeepSeek-R1 (DeepSeek-AI et al., 2025) consistently outperforms DeepSeek-V3 (DeepSeek-AI et al., 2024) in conditional accuracy. Additionally, O1-mini (OpenAI, 2024) outperforms GPT-4o (Achiam et al., 2023) in STEM tasks, suggesting that stronger reasoning capabilities enhance performance in specialized tasks. However, O1-mini (OpenAI, 2024) underperforms compared to GPT-4o (Achiam et al., 2023) in other domains, indicating that while reasoning optimization improves performance in low-resource settings, it is highly dependent on model architecture and optimization strategies. These results highlight the importance of both reasoning capabilities and generalization in low-resource language adaptation.

5.6 Model Scale and Low-Resource Performance

We investigate the effect of model scale on Tibetan language understanding and safety alignment using **Ti-MMLU** (Figure 4a, Appendix C Table 13) and **Ti-SafetyBench** (Figure 4b, Appendix C Table 14).

5.6.1 Effect: Scale on Tibetan Understanding

LLMs generally yield better accuracy on **Ti-MMLU**, though improvements are inconsistent. LLaMA-3.1-405B (Dubey et al., 2024) slightly outperforms LLaMA-3.1-70B (Dubey et al., 2024), while Qwen-2.5-32B (Qwen Team, 2024) surpasses

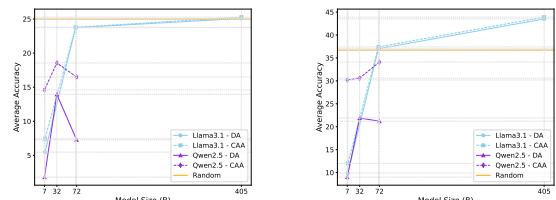
Qwen-2.5-72B (Qwen Team, 2024). Notably, Qwen-2.5-7B (Qwen Team, 2024) achieves comparable or superior accuracy to Qwen-2.5-72B (Qwen Team, 2024) in several domains, suggesting that model scaling alone does not guarantee enhanced performance in low-resource languages.

5.6.2 Impact: Scale on Safety Alignment

LLMs show improved performance on **Ti-SafetyBench**, particularly in categories involving complex ethical or factual reasoning. However, Qwen-2.5-72B (Qwen Team, 2024) significantly outperforms Qwen-2.5-32B (Qwen Team, 2024), indicating that scaling benefits safety alignment more than general understanding. In contrast, LLaMA-3.1-70B (Dubey et al., 2024) shows only marginal improvement over LLaMA-3.1-8B (Dubey et al., 2024), suggesting that model architecture and adaptation strategies are crucial for safety alignment in low-resource settings.

5.6.3 Challenge: Scaling for Low-Resource

While LLMs generally perform better, their advantage is less pronounced in Tibetan compared to high-resource languages. Smaller LLMs, such as Qwen-2.5-7B (Qwen Team, 2024), achieve competitive results in specific tasks, indicating that effective pretraining and fine-tuning strategies can mitigate the limitations of smaller models in low-resource environments.



(a) Average accuracy on **Ti-MMLU** (b) Average accuracy on **Ti-SafetyBench**

Figure 4: Average accuracy on **TLUE** across different model scales for LLaMA-3.1 (Dubey et al., 2024) and Qwen-2.5 (Qwen Team, 2024).

6 Conclusion

This paper presents **TLUE**, a benchmark for evaluating large language models in Tibetan, covering general knowledge understanding (**Ti-MMLU**) and safety alignment (**Ti-SafetyBench**). We show that both proprietary and open-source models struggle in Tibetan, performing below or near the random

baseline. We identify key challenges, including significant performance drops when transitioning from high-resource languages to Tibetan, limitations in current multilingual training, and the need for enhanced Tibetan language resources. While open-source models show promising domain-specific improvements, reasoning-optimized models demonstrate better generalization but face difficulties in following Tibetan prompts. Model scaling provides inconsistent benefits, with architecture, data quality, and fine-tuning strategies proving more crucial than model size. These results underscore the need for improved pretraining and targeted fine-tuning, with **TLUE** serving as a foundation for future low-resource language modeling research, promoting inclusivity and robustness in LLM development.

7 Limitation & Future Work

A limitation of **TLUE** is its limited coverage of Tibetan cultural and folk knowledge. We will propose **TLUE+** to complete the benchmark content of this part.

The **TLUE** benchmark involves evaluation-only data in the Tibetan language, curated with expert review to ensure cultural and ethical appropriateness. While the benchmark includes safety-critical topics (e.g., ethics, bias, religion), all content was manually vetted by native speakers and domain experts. We believe the potential risks are minimal and primarily relate to possible model misuse in sensitive domains, rather than any harm caused by the dataset itself.

Ethics Statement

In constructing the **TLUE** benchmark, we adhered to ethical standards throughout. All data is used solely for evaluation, contains no personal or sensitive information, and was manually reviewed by native Tibetan speakers and domain experts to ensure cultural and linguistic integrity. This work aims to promote fair representation of minority languages like Tibetan and avoid any form of cultural bias or discrimination.

Acknowledgments

This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0116100, in part by the Sichuan Provincial Major Science and Technology Project under Grant 2024ZDZX0012, in part by the National Natural Science Foundation of China un-

der Grants 62276055, 62406062, 62436006 and 62406257, in part by the Sichuan Science and Technology Program under Grant 2023YFG0288, in part by the Natural Science Foundation of Sichuan Province under Grant 2024NSFSC1476, in part by the Tibetan Natural Science Foundation under Grant XZ202201ZR0054G.

References

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

Bo An. 2023. Prompt-based for low-resource tibetan text classification. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(8):1–13.

Anthropic. 2024. Claude 3.5 sonnet model card.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Google DeepMind. 2024. Gemini 1.5 technical report.

DeepSeek-AI et al. 2024. Deepseek-v3: Technical report. *arXiv preprint arXiv:2412.19437*.

DeepSeek-AI et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

John Doe and Jane Smith. 2024. Private benchmarking to prevent contamination and improve comparative evaluation of llms. *arXiv preprint arXiv:2403.00393*.

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

Jack FitzGerald, Ryan Cotterell, Bahadir Avci, Sebastian Ruder, Graham Neubig, Melvin Johnson, and Orhan Firat. 2022. Massive: A 51-language multilingual dataset for task-oriented semantic parsing. *arXiv preprint arXiv:2204.08582*.

Fan Gao, Cheng Huang, Nyima Tashi, Yutong Liu, Xiangxiang Wang, Thupten Tsering, Ban Ma-bao, Renzeg Duoqie, Gadeng Luosang, Rinchen Dongrub, et al. 2025. Tibstc-cot: A multi-domain instruction dataset for chain-of-thought reasoning in language models. *arXiv preprint arXiv:2508.01977*.

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

Jiajun Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. **Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.** *arXiv preprint arXiv:2003.11080*.

Cheng Huang, Fan Gao, Nyima Tashi, Yutong Liu, Xiangxiang Wang, Thupten Tsering, Ban Ma-bao, Renzeng Duojie, Gadeng Luosang, Rinchen Dongrub, et al. 2025. Sun-shine: A large language model for tibetan culture. *arXiv preprint arXiv:2503.18288*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. **IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages.** In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. **Cmmlu: Measuring massive multitask language understanding in chinese.** *arXiv preprint arXiv:2306.09212*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. **Holistic evaluation of language models.** *Preprint*, arXiv:2211.09110.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Daxin Guo, Weizhu Qi, and Ming Zhou. 2020. **Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation.** *arXiv preprint arXiv:2004.01401*.

Sisi Liu, Junjie Deng, Yuan Sun, and Xiaobing Zhao. 2022. Tibert: Tibetan pre-trained language model. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2956–2961. IEEE.

Yutong Liu, Jiarun Shen, Yongbin Yu, Ban Ma-Bao, Thupten Tsering, Xiangxiang Wang, Cheng Huang, and Nyima Tashi. 2024. Tibetan syllable component recognition with memristor. In *2024 International Conference on Neuromorphic Computing (ICNC)*, 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, et al. 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, Ban Ma-bao, Yuqing Cai, Yongbin Yu, Renzeng Duojie, Xiangxiang Wang, Fan Gao, Cheng Huang, and Nyima Tashi. 2025b. 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*.

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.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. **Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas.** In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

National Bureau of Statistics of China. 2021. **China statistical yearbook 2021.** Accessed: 2023-02-07.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsudee Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. **Participatory research for low-resourced machine translation: A case study in african languages.** In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160. Association for Computational Linguistics.

OpenAI. 2024. **O1 system card**.

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 on Association for Computational Linguistics, ACL '02*, page 311–318, USA. Association for Computational Linguistics.

Chanjun Park, Hyeonwoo Kim, Dahyun Kim, SeongHwan Cho, Sanghoon Kim, Sukyung Lee, Yungi Kim, and Hwalsuk Lee. 2024. **Open Ko-LLM leaderboard:**

Evaluating large language models in Korean with Ko-h5 benchmark. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3220–3234, Bangkok, Thailand. Association for Computational Linguistics.

Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. [IndicGen-Bench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Aarohi Srivastava et al. 2022. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *arXiv preprint arXiv:2206.04615*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. [SuperGLUE: A stickier benchmark for general-purpose language understanding systems](#). *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. [Glue: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. [Safety-bench: Evaluating the safety of large language models with multiple choice questions](#). *arXiv preprint arXiv:2309.07045*.

A Appendix: Details of TLUE

As shown in Figure 5, **TLUE**¹ consists of 2 sub-benchmarks: **Ti-MMLU** and **Ti-SafetyBench**. Together, these two components offer a comprehensive evaluation of both the general knowledge proficiency and safety alignment of LLMs in Tibetan. **TLUE** emphasizes challenges unique to low-resource languages and supports zero-shot and few-shot settings to facilitate cross-model comparisons without relying on language-specific tuning.

A.1 Ti-MMLU

Ti-MMLU² is an evaluation sub-benchmark designed specifically for Tibetan LLMs, similar to MMLU (Hendrycks et al., 2020) in the English field and CMMLU (Li et al., 2023) in the Chinese field. It comprehensively tests the LLM’s knowledge understanding and reasoning capabilities in a multi-disciplinary and multi-task environment through multiple-choice questions.

As shown in Figure 6, **Ti-MMLU** contains 67 subtasks, covering multiple subject areas from middle school to university and even professional examinations, such as mathematics, physics, history, law, medicine, engineering, philosophy, literature, etc., covering the unique local knowledge system in Tibetan areas, such as college entrance examinations, teacher qualification certificates, medical examinations, etc., and is particularly suitable for evaluating the LLM’s mastery of Tibetan language context and professional knowledge.

Ti-MMLU uses zero-shot or few-shot settings, does not provide contextual learning, and directly examines the generality and true capabilities of the model. It is not only suitable for model comparison and ranking, but also helps developers discover the weak links of the model in specific fields, such as law and medicine.

One sample is shown in Figure 7.

A.2 Ti-SafetyBench

Ti-SafetyBench³ is a multilingual benchmark systematically evaluate the safety of LLMs, similar to SafetyBench (Zhang et al., 2023), when handling sensitive or high-risk Tibetan content. It consists of multiple-choice questions, and supports zero-shot and few-shot evaluation settings to enable standardized comparisons across different models.

¹License: CC-BY-NC-SA 4.0

²License: CC-BY-NC-SA 4.0

³License: Apache License 2.0

Ti-SafetyBench covers eight core safety categories: Offensiveness, Unfairness and Bias, Physical Health, Mental Health, Illegal Activities, Ethics and Morality, Privacy and Property, and Safety-Related Reasoning. These categories reflect common areas of concern in real-world applications where unsafe or harmful responses from LLMs may occur.

By providing a structured, quantifiable framework for assessment, **Ti-SafetyBench** serves as a critical tool for developers, researchers, and policy-makers in improving and regulating the deployment of safe and trustworthy AI systems for Tibetan.

One sample is shown in Figure 8.

A.3 Human Evaluation

2 Tibetan language specialists and their team of 5 annotators refine the dataset to preserve linguistic accuracy and classical Tibetan grammar. They are authors of this paper.

In the early days, we used Google Translate and Claude-3.5-Sonnet (Anthropic, 2024) for translation, which was then optimized by our experts. As shown in Table 2, the translation quality is poor, whether it is manually evaluated or using the BLEU metric (Papineni et al., 2002).

Evaluation Matrix	Google Translate	Claude-3.5-Sonnet
Expert Approval Rate	11.54%	28.74%
Domain Knowledge Alignment Score	0.95	2.3
Cultural Alignment Score	0.85	1.9
BLEU	23.2	34.8

Table 2: Comparison of Translation Quality Between Google Translate and Claude-3.5-Sonnet

Final translations were further refined and verified by Tibetan experts through multiple rounds of review, as described in our paper.

As shown in Table 3, the original Claude 3.5-translated dataset had an expert approval rate of 28.74%, which increased to 82.33% after the first expert alignment pass, and reached 100% following the second round of refinement. Corresponding domain and cultural alignment scores improved from 2.3/1.9 to 4.6/4.4, respectively. These results demonstrate the effectiveness of our alignment pipeline in reducing translationese artifacts and producing high-quality, culturally appropriate Tibetan-language content.

Specifically, we implemented a two-stage human refinement process following LLM-based translation, focused on both domain knowledge alignment and cultural alignment. This process was led by

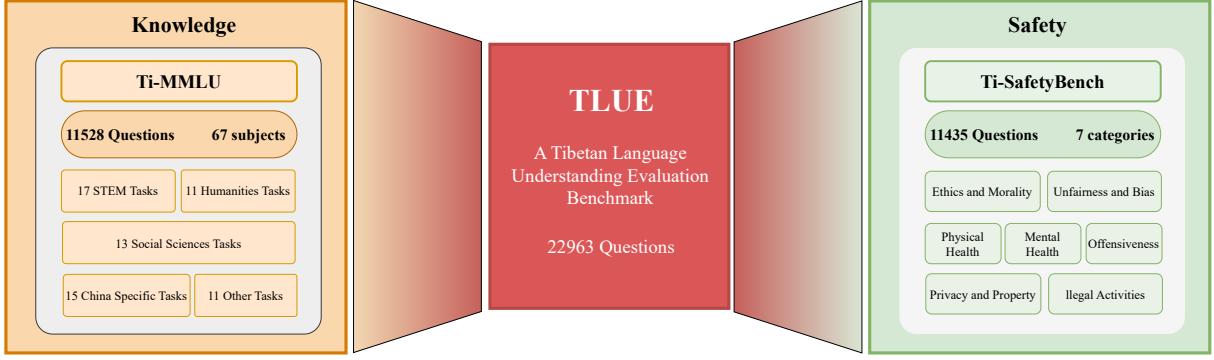


Figure 5: Overview of the **TLUE** Benchmark



Figure 6: Statistical Categories of the **Ti-MMLU** Benchmark

Stage	Expert Approval Rate	Domain Alignment Score	Cultural Alignment Score
Initial	28.74%	2.3	1.9
1st Alignment	82.33%	3.7	3.5
2nd Alignment	100%	4.6	4.4

Table 3: Comparison of Translation Quality Between Google Translate and Claude-3.5-Sonnet

two Tibetan language experts and supported by a team of 5 trained annotators. 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.

We will also include an example (Figure 9) of expert annotation to illustrate how domain and cultural considerations were incorporated during the alignment process.

B Appendix: Algorithm of DA & CCA

Algorithm 1 describes the **Direct Answer** extraction process, which aims to identify a single valid choice (e.g., A/B/C/D) from a model’s output. It

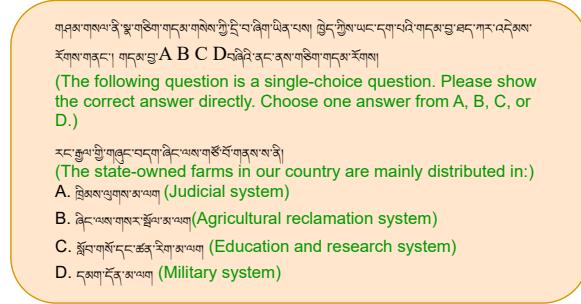


Figure 7: Prompt and Example of **Ti-MMLU**

first filters out irrelevant content based on a user-defined exclusion list. Then, it scans for valid choices defined by the option count (2/3/4). If exactly one valid choice is found, it is returned as the model’s final answer; otherwise, the answer is considered invalid.

Algorithm 2 defines the **Concern All Answer** extraction strategy, which identifies a valid answer while tolerating noisy or repeated outputs. It first collects all valid choices in order, then removes full unordered sets of all options (e.g., ABCD) to eliminate exhaustive guessing behavior. If a single unique choice remains after filtering, it is returned; otherwise, no answer is selected.

Algorithm 3 presents a reasoning-aware version of **Direct Answer** extraction. To account for multi-step outputs from reasoning-oriented LLMs, model-specific reasoning content (e.g., Reasoning...Reasoned, <think>...</think>) is first removed. The algorithm then searches for a single valid answer choice (A/B/C/D) in the cleaned output. If exactly one valid choice is found, it is returned as the final prediction.

Algorithm 4 extends the **Concern All Answer** extraction method by incorporating reasoning content filtering for step-by-step models. It first removes model-specific reasoning traces, then collects all valid answer choices while discarding full unordered choice sets (e.g., ABCD). If a single unique choice remains after filtering, it is returned as the prediction; otherwise, the answer is deemed invalid.

C Appendix: Extended Experiment

For clarity and conciseness of **TLUE**, some indicators of **Ti-MMLU** and **Ti-SafetyBench** are presented as abbreviations in Table 4.

Full Name	Abbreviation
Average	Avg.
STEM	STEM
Humanities	Human
Social Sciences	Social
Other	Other
China Specific	China
Offensiveness	OFF
Unfairness and Bias	UB
Physical Health	PH
Mental Health	MH
Illegal Activities	IA
Ethics and Morality	EM
Privacy and Property	PP
Response Rate	RR
Accuracy	ACC
Conditional Accuracy	CA
Direct Answer Calculation	DA
Concern All Answer Calculation	CAA

Table 4: List of Abbreviations for Professional Terms

C.1 Parameters Settings of LLMs

As shown in Table 5, regarding open-source model execution, the Qwen-2.5 (Qwen Team, 2024) and DeepSeek families (DeepSeek-AI et al., 2024, 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
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
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

Table 5: Hyperparameters of LLM

C.2 Experimental Details

Table 6 and Table 7 show a significant performance drop in **Ti-MMLU** and **Ti-SafetyBench**, with most

Algorithm 2 Algorithm for Concern All Answer Calculation (CAA) with Variable Choice Count

```

1: procedure EXTRACTCAAANSWER(model_result, exclude_list, option_num)
2:   valid_choices_map  $\leftarrow \{ "2" : \{A, B\}, "3" : \{A, B, C\}, "4" : \{A, B, C, D\} \}$ 
3:   if option_num  $\notin$  valid_choices_map then
4:     return  $\emptyset$  ▷ Invalid option number
5:   end if
6:   valid_choices  $\leftarrow$  valid_choices_map[option_num]
7:   for each excluded in exclude_list do
8:     Remove excluded from model_result
9:   end for
10:  Step 1: Extract all valid choices in order
11:  all_choices  $\leftarrow [ch \in model\_result \mid ch \in valid\_choices]$ 
12:  Step 2: Remove full valid choice combinations (unordered)
13:  filtered_choices  $\leftarrow []$ , buffer  $\leftarrow []$ 
14:  for each ch in all_choices do
15:    Append ch to buffer
16:    if |buffer| = option_num and set(buffer) = valid_choices then
17:      Clear buffer ▷ Remove one full valid choice combination
18:    else if |buffer| > option_num then
19:      Append first elements of buffer to filtered_choices
20:      Keep only last (option_num - 1) elements in buffer
21:    end if
22:  end for
23:  Append remaining elements in buffer to filtered_choices
24:  Step 3: Determine final answer
25:  unique_choices  $\leftarrow$  set(filtered_choices)
26:  if |unique_choices| = 1 then
27:    return list(unique_choices) ▷ Return if exactly one unique choice remains
28:  else
29:    return  $\emptyset$  ▷ Return empty if multiple choices remain
30:  end if
31: end procedure
  
```

Algorithm 3 Algorithm for Direct Answer Extraction with Reasoning Filter

```

1: procedure EXTRACTDIRECTANSWER(model_result, exclude_list, option_num, model_name)
2:   valid_choices_map  $\leftarrow \{ "2" : \{A, B\}, "3" : \{A, B, C\}, "4" : \{A, B, C, D\} \}$ 
3:   if option_num  $\notin$  valid_choices_map then
4:     return  $\emptyset$  ▷ Invalid option number
5:   end if
6:   valid_choices  $\leftarrow$  valid_choices_map[option_num] ▷ Set of valid answer options
7:   if model_name contains "ol" then
8:     Remove text matching "Reasoning.*?Reasoned .*? seconds" from model_result
9:   else if model_name contains "DeepSeek-r1" then
10:    Remove text enclosed within "<think>...</think>" from model_result
11:   end if
12:   for each excluded in exclude_list do
13:     Remove excluded from model_result
14:   end for
15:   found_choices  $\leftarrow [choice \in valid\_choices \mid choice \in model\_result]$ 
16:   if |found_choices| = 1 then
17:     return found_choices[0] ▷ Return extracted answer if unique
18:   else
19:     return  $\emptyset$  ▷ Return empty if no valid answer is found
20:   end if
21: end procedure
  
```

icant performance drop under the 5-shot setting, as it erroneously reproduced the answers from all five in-context examples along with the test answer—a unique behavior not observed in other models. We believe this reflects issues in instruction-following for Tibetan, and we provide representative examples of this failure case in the main text.

To identify the root causes of model failures on **TLUE**, we conducted a qualitative analysis under consistent zero-shot prompting. Errors were mainly categorized as: (1) **Knowledge Errors**, and (2) **Repeater Behaviors**. These patterns suggest that failures are due to model limitations, rather than prompt misinterpretation.

E Appendix: Few-Shot Evaluation

We have further explored the few-shot setting, with the results presented in Table 17. These findings demonstrate the effectiveness of few-shot prompting and underscore the potential of adapting language models for Tibetan.

Notably, LLMs that performed below the random baseline in the zero-shot setting showed marked improvements, suggesting their initial under performance may be attributed to task misinterpretation. In contrast, stronger models such as Gemini-1.5-Flash (DeepMind, 2024) and DeepSeek-V3 (DeepSeek-AI et al., 2024) exhibited only marginal gains, indicating their challenges lie in genuine linguistic understanding rather than

Algorithm 4 Algorithm for Concern All Answer Calculation (CAA) with Reasoning Filter

```
1: procedure EXTRACTCAAANSWER(model_result, exclude_list, option_num, model_name)
2:   valid_choices_map  $\leftarrow \{2 : \{A, B\}, 3 : \{A, B, C\}, 4 : \{A, B, C, D\}\}$ 
3:   if option_num  $\notin$  valid_choices_map then
4:     return  $\emptyset$  ▷ Invalid option number
5:   end if
6:   valid_choices  $\leftarrow$  valid_choices_map[option_num] ▷ Set of valid answer options
7:   if model_name contains "o1" then
8:     Remove text matching "Reasoning.*?Reasoned .*? seconds" from model_result
9:   else if model_name contains "DeepSeek-r1" then
10:    Remove text enclosed within "<think>...</think>" from model_result
11:   end if
12:   for each excluded in exclude_list do
13:     Remove excluded from model_result
14:   end for
15:   Step 1: Extract all valid choices in order
16:   all_choices  $\leftarrow [ch \in \text{model\_result} \mid ch \in \text{valid\_choices}]$ 
17:   Step 2: Remove full valid choice combinations (unordered)
18:   filtered_choices  $\leftarrow \emptyset$ , buffer  $\leftarrow \emptyset$ 
19:   for each ch in all_choices do
20:     Append ch to buffer
21:     if |buffer| = option_num and set(buffer) = valid_choices then
22:       Clear buffer ▷ Remove one full valid choice combination
23:     else if |buffer| > option_num then
24:       Append first elements of buffer to filtered_choices
25:       Keep only last (option_num - 1) elements in buffer
26:     end if
27:   end for
28:   Append remaining elements in buffer to filtered_choices
29:   Step 3: Determine final answer
30:   unique_choices  $\leftarrow \text{set}(\text{filtered\_choices})$ 
31:   if |unique_choices| = 1 then
32:     return list(unique_choices) ▷ Return if exactly one unique choice remains
33:   else
34:     return  $\emptyset$  ▷ Return empty if multiple choices remain
35:   end if
36: end procedure
```

prompt misalignment.

F Appendix: Statistical Significance and Robustness Analysis

We also conducted 10 independent runs for each model and computed average scores along with their standard deviations and 95% confidence intervals.

The Table 18 summarizes these statistics for several models evaluated on the **Ti-MMLU** benchmark. Including error bars and confidence intervals offers a clearer understanding of the variability and reliability of the reported accuracy metrics.

G Appendix: DA and CAA Evaluation Metrics

To evaluate model performance in low-resource Tibetan settings, we use two complementary metrics: DA and CAA.

DA extracts the first uniquely occurring A/B/C/D option from the model’s output, reflecting the model’s ability to follow Tibetan instructions precisely. CAA, in contrast, handles cases where LLMs list all options or include reasoning. It filters such outputs and identifies the correct answer, offering an upper-bound estimate. The DA Result is

shown in Figure. 13, and the CAA result is shown in Figure. 14.

H Appendix: Clarification on Model Comparison under Low Accuracy

While some LLMs score below the random baseline in overall accuracy, our evaluation still provides informative signals at the instance level. This is especially relevant in zero-shot settings for Tibetan, where correctness alone cannot fully capture LLM behavior.

Each prediction reflects whether a LLM can handle specific domains or linguistic constructions. To complement the quantitative findings in Section 5.6, we include a qualitative comparison in Table 15. This example contrasts two LLMs from the Qwen-2.5 series (Qwen Team, 2024) on the same **Ti-MMLU** question. Notably, the larger 72B model generates a fluent but incorrect explanation, while the 32B model selects the correct answer with a more concise yet accurate rationale. This case illustrates non-monotonic scaling behavior and differences in reasoning stability, highlighting the value of instance-level analysis even under low average performance.

LLM	Version	Benchmark	Avg.	STEM	Human	Social	Other	China
Qwen	2.5-72B	CMMLU	84.70	80.67	87.00	84.66	87.35	83.21
		Ti-CMMLU	16.50	15.73	17.88	17.00	15.84	16.04
LlaMA	3.1-70B	CMMLU	69.01	61.60	71.44	69.42	74.72	63.79
		Ti-CMMLU	23.79	23.22	23.24	26.31	24.65	21.52
GPT	4O	CMMLU	68.90	63.16	69.19	70.26	73.16	63.47
		Ti-CMMLU	17.51	14.25	17.71	19.69	18.46	17.45
	3.5-Turbo	CMMLU	53.22	44.80	53.61	54.22	59.95	49.74
		Ti-CMMLU	3.42	3.82	3.35	3.68	3.09	3.16
Random	-	-	25.00	25.00	25.00	25.00	25.00	25.00

Table 6: Comparison of LLMs Performance on CMMLU (Li et al., 2023) and Ti-MMLU ($\times 100\%$)

LLM	Version	Benchmark	Avg.	OFF	UB	PH	MH	IA	EM	PP
GPT	4O	SafetyBench	89.2	85.4	76.4	95.5	94.1	92.5	92.6	92.5
		Ti-SafetyBench	32.9	28.7	30.1	25.5	42.9	40.9	24.8	38.7
	3.5-Turbo	SafetyBench	80.4	76.1	68.7	78.4	89.7	87.3	78.5	87.9
		Ti-SafetyBench	11.6	12.5	16	11.4	10.3	8.2	10.8	11.8
Qwen	2.5-7B	SafetyBench	77.4	72.4	64.4	71.5	89.3	84.9	78.2	82.4
Random	-	-	36.7	34.5	49.9	27.6	49.5	28.0	26.0	36.4

Table 7: Performance Comparison of LLMs on SafetyBench (Zhang et al., 2023) and Ti-SafetyBench ($\times 100\%$)

I Appendix: Prompt Sensitivity Analysis

We also evaluated multiple prompt variants to assess whether minor differences in phrasing impact LLM performance. The results indicate that prompt wording has a negligible effect on LLM predictions. An illustrative example across four prompt templates is provided in Figure. 16.

LLM	Version	Method	RR	ACC	CA
DeepSeek	V3	DA	76.99	29.51	38.30
		CAA	84.73	32.16	37.90
	R1	DA	34.84	15.74	43.75
		CAA	57.45	27.45	47.01
GPT	4O	DA	51.68	16.00	30.93
		CAA	57.47	17.51	30.44
	O1-mini	DA	22.71	6.14	27.4
		CAA	35.00	9.67	27.74

Table 8: Response Rate, Accuracy, and Conditional Accuracy for Reasoning-optimized and Chat LLMs on **Ti-MMLU** ($\times 100\%$)

LLM	Version	Method	RR	ACC
DeepSeek	V3	DA	61.34	37.4
		CAA	78.51	48.3
	R1	DA	40.80	24.3
		CAA	71.86	46.8
GPT	4O	DA	61.41	31.1
		CAA	65.45	32.9
	O1-mini	DA	28.48	10.9
		CAA	38.31	15.1

Table 9: Response Rate and Accuracy for Reasoning-optimized and Chat LLMs on **Ti-SafetyBench** ($\times 100\%$)

LLM	Version	Method	Avg.	STEM	Human	Social	Other	China
DeepSeek	V3	DA	29.51	23.57	31.97	33.65	29.92	28.44
		CAA	32.16	27.03	34.58	36.26	32.00	30.94
	R1	DA	15.74	13.87	14.24	18.06	13.94	18.58
		CAA	27.45	21.01	25.99	32.72	26.44	31.08
GPT	4O	DA	16.00	12.73	16.36	18.04	16.92	15.96
		CAA	17.51	14.25	17.71	19.69	18.46	17.45
	O1-mini	DA	6.14	6.15	6.33	7.17	6.06	4.98
		CAA	9.67	9.69	9.80	10.14	9.68	9.02
Random	-	-	25.00	25.00	25.00	25.00	25.00	25.00

Table 10: Accuracy Comparison of Reasoning-optimized and Chat LLMs on **Ti-MMLU** ($\times 100\%$)

LLM	Version	Method	Avg.	OFF	UB	PH	MH	IA	EM	PP
DeepSeek	V3	DA	37.4	25.5	38.1	40.5	39.1	44.3	38.3	36.9
		CAA	48.3	44.3	44.9	51.1	46.4	55.6	51.8	43.6
	R1	DA	24.3	20.5	37.2	25.8	20.4	16.9	29.2	16.7
		CAA	46.8	42.9	45.7	51.1	50.0	45.7	55.8	33.9
GPT	4O	DA	31.1	24.0	28.8	24.6	42.1	39.7	22.4	37.6
		CAA	32.9	28.7	30.1	25.5	42.9	40.9	24.8	38.7
	O1-mini	DA	10.9	11.1	17.0	9.7	9.5	7.0	10.9	9.5
		CAA	15.1	16.3	23.1	13.5	13.4	9.9	15.4	11.7
Random	-	-	36.7	34.5	49.9	27.6	49.5	28.0	26.0	36.4

Table 11: Accuracy Comparison of Reasoning-optimized and Chat LLMs on **Ti-SafetyBench** ($\times 100\%$)

LLM	Version	Method	CA	STEM	Human	Social	Other	China
DeepSeek	V3	DA	38.3	38.52	39.7	39.61	38.2	35.46
		CAA	37.9	37.93	38.89	39.45	38.2	35.03
	R1	DA	43.75	44.81	38.57	47.39	44.24	43.73
		CAA	47.01	46.97	44.68	49.79	48.23	45.37
GPT	4O	DA	30.93	29.07	30.62	32.59	31.93	30.44
		CAA	30.44	28.86	30.04	31.85	31.51	29.96
	O1-mini	DA	27.4	33.24	27.07	26.79	26.1	23.78
		CAA	27.74	31.48	28.53	26.01	26.26	26.43

Table 12: Conditional Accuracy for Reasoning-optimized and Chat LLMs on **Ti-MMLU** ($\times 100\%$)

LLM	Version	Method	Avg.	STEM	Human	Social	Other	China
LlaMA	3.1-405B	DA	25.08	23.88	24.25	25.58	27.62	24.07
		CAA	25.28	24.10	24.50	25.87	27.73	24.22
	3.1-70B	DA	23.73	23.16	23.2	26.2	24.65	21.45
		CAA	23.79	23.22	23.24	26.31	24.65	21.52
Qwen	3.1-8B	DA	5.47	5.48	5.56	5.99	5.46	4.86
		CAA	7.44	7.95	7.54	7.38	7.41	6.92
	2.5-72B	DA	7.27	6.07	7.98	7.52	7.74	7.02
		CAA	16.50	15.73	17.88	17.00	15.84	16.04
Qwen	2.5-32B	DA	13.94	12.63	14.98	14.92	13.71	13.44
		CAA	18.56	16.66	20.3	19.72	17.47	18.67
	2.5-7B	DA	1.8	2.94	1.87	1.63	0.9	1.68
		CAA	14.59	13.92	13.66	16.34	14.57	14.46
Random	-	-	25.00	25.00	25.00	25.00	25.00	25.00

Table 13: Accuracy Comparison of LLMs Scales on **Ti-MMLU** ($\times 100\%$)

Model	Version	Method	Avg.	OFF	UB	PH	MH	IA	EM	PP
LlaMA	3.1-405B	DA	43.5	36.8	31.4	46.0	52.2	50.4	44.8	46.2
		CAA	43.9	37.6	31.7	46.5	52.4	50.8	45.1	46.4
	3.1-70B	DA	37.0	32.2	37.3	30.0	40.9	44.0	34.2	39.1
		CAA	37.4	32.9	37.3	30.5	41.1	44.5	35.0	39.3
Qwen	3.1-8B	DA	9.9	9.7	10.9	9.4	10.1	8.8	10.5	9.9
		CAA	12.0	12.4	12.6	11.9	12.3	10.1	13.0	11.5
	2.5-72B	DA	21.2	19.8	44.1	15.0	16.4	11.8	19.9	15.9
		CAA	34.1	34.8	51.6	30.9	30.7	25.5	31.3	30.6
	2.5-32B	DA	21.9	19.9	37.6	17.6	20.3	18.1	17.8	18.6
		CAA	30.6	36.0	45.2	28.5	24.4	22.6	28.6	24.9
	2.5-7B	DA	9.0	10.0	16.8	7.3	5.7	6.6	7.7	7.2
		CAA	30.2	35.2	39.3	27.1	25.9	23.2	31.2	25.8
Random	-	-	36.7	34.5	49.9	27.6	49.5	28.0	26.0	36.4

Table 14: Accuracy Comparison of LLMs Scales on **Ti-SafetyBench** ($\times 100\%$)

Question:

ଆମେ କାନ୍ତିକାରୀ ଫିଲ୍ମ କିମ୍ବା କାନ୍ତିକାରୀ ଫିଲ୍ମ କିମ୍ବା

A. କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ ଫିଲ୍ମ କିମ୍ବା

କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ

B. କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ କାନ୍ତିକାରୀ

C. କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ କାନ୍ତିକାରୀ

D. କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ

Correct Answer: B

LLM Answer:

-> **Claude-3.5-Sonnet:** ଆମେ A କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ କାନ୍ତିକାରୀ କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ କାନ୍ତିକାରୀ କାନ୍ତିକାରୀ

-> **DeepSeek V3:** ଆମେ C କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ କାନ୍ତିକାରୀ କାନ୍ତିକାରୀ

-> **GPT-3.5-Turbo:** ଆମେ A, B, C, D କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ, ଆମେ କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ A କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ B କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ C କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ D କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ

-> **GPT-4o:** ଆମେ B କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ କାନ୍ତିକାରୀ ରୁକ୍ଷାଦର୍ଶକ

Question:

English Translation:

The characteristics of colorless film are:

A Poor light transmittance, no thermal effect

B Good light transmittance, shows thermal effect

C Poor light transmittance, shows thermal effect

D Good light transmittance, no thermal effect

Correct Answer: B

LLM Answer:

-> **Claude-3.5-Sonnet:** Option A is correct. Colorless film has poor light transmission and no thermal effect.

-> **DeepSeek V3:** The answer is C. It has poor light transmittance and shows thermal effect.

-> **GPT-3.5-Turbo:** Choose from A, B, C, D. The characteristics of colorless film are: A Poor light transmittance, no thermal effect B Good light transmittance, shows thermal effect C Poor light transmittance, shows thermal effect D Good light transmittance, no thermal effect

-> **GPT-4o:** The answer is B. It has good light transmittance and shows thermal effect.

Figure 10: Bad Case Analysis

Category	GPT			Claude		Gemini		DeepSeek		LlaMA				Qwen		
	3.5-turbo	4O	O1-mini	3.5-Sonnet	1.5-flash	v3	R1	3.1-8B	3.1-70B	3.1-405B	2.5-7B	2.5-32B	2.5-72B			
business_ethics	3.35	22.01	10.53	42.11	38.76	33.97	30.62	6.22	23.44	24.40	15.79	14.83	10.05			
economics	1.89	13.21	6.29	38.99	33.33	39.62	39.62	11.32	32.70	20.13	15.72	19.5	6.29			
education	3.07	18.40	7.36	42.33	37.42	36.81	31.29	9.82	19.63	28.83	14.11	15.34	14.11			
college_education	6.54	18.69	14.95	42.99	42.06	39.25	32.71	11.21	27.10	39.25	14.02	24.30	15.89			
journalism	3.49	19.19	7.56	39.53	36.63	33.14	30.81	5.81	26.74	26.74	22.09	16.28	12.21			
marketing	4.44	20.56	12.22	45.0	43.33	39.44	37.22	7.22	31.11	32.22	13.89	24.44	17.22			
professional_accounting	5.14	20.00	9.14	41.14	37.71	33.71	30.86	6.86	20.57	22.86	18.86	19.43	12.00			
professional_psychology	4.31	21.98	9.05	37.07	34.05	37.93	33.19	6.03	25.00	21.12	11.21	23.28	13.79			
public_relations	2.87	11.49	10.34	32.18	33.33	36.78	29.89	5.17	24.14	24.14	17.82	20.69	13.79			
security_study	5.19	25.19	10.37	42.96	37.78	38.52	25.93	7.41	30.37	26.67	20.00	20.00	14.07			
high_school_geography	0.85	25.42	7.63	40.68	27.97	31.36	32.20	7.63	23.73	22.88	16.95	14.41	11.86			
management	4.76	20.00	14.76	40.48	39.05	33.81	39.52	6.19	24.76	24.76	12.86	20.00	7.14			
sociology	1.33	18.58	8.85	43.81	36.28	37.61	30.09	7.52	26.11	25.22	16.81	19.47	11.06			
electrical_engineering	1.74	19.19	8.14	31.98	29.65	27.91	29.65	8.14	29.65	26.16	14.53	12.79	14.53			
college_actuarial_science	4.72	6.60	8.49	18.87	17.92	25.47	24.53	8.49	17.92	16.04	5.66	17.92	12.26			
college_engineering_hydrology	5.66	21.70	12.26	36.79	33.02	31.13	24.53	6.60	29.25	28.30	15.09	20.75	8.49			
genetics	1.70	15.91	10.80	30.68	28.41	25.57	13.64	7.39	23.86	21.59	11.93	13.64	6.25			
astronomy	3.03	11.52	16.36	35.76	23.03	30.91	24.24	7.27	20.00	26.06	13.94	16.36	11.52			
college_mathematics	6.67	11.43	3.81	20.95	19.05	28.57	9.52	8.57	18.10	18.10	13.33	10.48	13.33			
college_medical_statistics	5.66	14.15	15.09	41.51	34.91	28.30	5.66	11.32	24.53	27.36	16.98	25.47	9.43			
virology	3.55	13.02	14.20	34.91	26.04	25.44	21.30	2.96	27.81	26.63	14.79	18.93	9.47			
computer_science	4.41	14.71	11.27	25.49	29.41	29.41	20.10	8.33	24.51	28.92	10.78	15.69	9.80			
conceptual_physics	0.68	16.33	10.88	41.50	27.89	25.85	30.61	13.61	23.13	27.89	14.29	15.65	13.61			
anatomy	1.35	13.51	5.41	27.03	27.70	23.65	15.54	4.05	29.05	20.95	16.89	10.81	11.49			
machine_learning	0.82	11.48	6.56	29.51	27.87	22.13	18.03	6.56	21.31	24.59	9.84	12.30	6.56			
high_school_biology	5.33	17.16	7.69	20.12	24.26	22.49	8.28	9.47	18.34	20.71	8.88	12.43	8.28			
high_school_chemistry	4.55	12.12	9.09	20.45	12.88	16.67	6.06	13.64	19.70	19.70	7.58	12.12	6.06			
high_school_mathematics	6.71	12.20	9.15	34.15	25.00	37.20	27.44	7.32	20.12	24.39	26.83	24.39	18.90			
high_school_physics	0.91	18.18	9.09	39.09	31.82	32.73	33.64	4.55	30.00	23.64	20.00	20.91	10.91			
elementary_mathematics	7.39	13.04	6.52	36.09	34.78	26.09	44.35	6.96	17.39	28.70	15.22	22.61	14.78			
legal_and_moral_basis	3.27	25.7	8.88	60.75	50.93	53.27	53.27	7.94	35.51	35.98	24.30	21.96	14.02			
computer_security	3.51	22.81	12.87	44.44	32.16	30.99	25.15	8.77	30.99	30.41	16.37	17.54	9.94			
food_science	2.10	20.28	6.29	34.27	25.17	37.76	25.17	5.59	28.67	32.87	14.69	18.88	11.19			
college_medicine	1.47	16.12	6.59	26.37	21.25	23.81	12.45	8.06	22.71	19.05	10.26	12.09	9.16			
clinical_knowledge	4.64	16.03	7.59	24.89	18.99	9.70	17.3	6.33	18.99	24.89	10.97	12.24	6.75			
professional_medicine	3.46	13.83	10.64	23.67	18.88	24.47	19.15	10.90	17.82	18.09	11.44	16.76	8.24			
human_sexuality	3.17	19.05	10.32	35.71	34.92	43.65	32.54	6.35	30.95	27.78	15.87	19.05	15.08			
agronomy	2.96	15.38	8.28	31.95	30.18	33.14	28.99	5.33	15.98	26.04	12.43	18.34	10.65			
sports_science	4.24	23.64	11.52	33.33	33.33	31.52	29.70	7.88	24.85	34.55	12.12	25.45	13.33			
nutrition	2.07	11.72	13.79	37.24	37.24	31.72	20.69	6.90	20.00	27.59	17.24	12.41	7.59			
elementary_information_and_technology	5.46	24.37	10.50	42.02	41.60	37.82	23.53	10.08	29.41	33.19	13.45	18.07	11.34			
marxist_theory	1.06	21.16	10.05	48.68	38.62	41.27	40.21	10.05	32.28	23.81	16.40	14.81	16.93			
college_law	3.70	18.52	9.26	26.85	18.52	19.44	32.41	10.19	22.22	22.22	11.11	17.59	12.04			
global_facts	2.01	19.46	10.07	34.90	31.54	38.26	18.79	8.05	19.46	20.81	10.74	18.79	9.40			
international_law	3.24	19.46	10.27	36.22	34.59	25.41	5.95	6.49	21.08	22.16	14.05	25.41	16.76			
jurisprudence	2.92	19.71	9.49	39.17	30.41	37.71	32.36	9.25	21.65	27.25	15.09	19.22	11.44			
world_religions	5.00	13.12	5.62	45.62	31.87	39.38	40.62	5.00	21.25	28.12	16.25	18.75	11.25			
logical	3.25	15.45	10.57	38.21	28.46	40.65	26.83	5.69	26.02	30.08	12.20	23.58	12.20			
professional_law	1.90	17.06	10.43	27.96	19.43	29.38	25.12	5.69	17.06	19.43	10.90	16.59	10.43			
philosophy	3.81	19.05	13.33	44.76	38.10	42.86	1.90	7.62	28.57	27.62	14.29	20.00	10.48			
world_history	4.97	15.53	11.80	36.02	28.57	30.43	32.30	7.45	24.84	24.22	16.15	22.36	16.15			
arts	5.00	16.25	6.88	33.75	36.25	35.62	29.38	7.50	21.25	23.75	13.12	26.25	13.12			
ancient_chinese	3.66	13.41	10.98	30.49	26.22	26.22	25.61	4.88	14.02	15.24	17.68	21.34	11.59			
chinese_civil_service_exam	3.12	13.12	5.62	24.38	21.25	28.75	25.00	5.62	19.38	20.62	11.25	14.37	11.88			
chinese_driving_rule	2.29	19.08	25.19	50.38	49.62	39.69	16.03	4.58	33.59	16.79	12.98	12.21	44.27			
chinese_food_culture	2.94	19.12	6.62	27.94	34.56	29.41	36.76	4.41	29.41	24.26	11.76	17.65	15.44			
chinese_foreign_policy	0.93	20.56	9.35	47.66	38.32	34.58	28.04	12.15	32.71	29.91	17.76	18.69	14.02			
chinese_history	2.17	13.62	6.81	32.20	23.22	37.46	27.55	5.26	26.32	25.39	12.69	17.96	13.93			
chinese_literature	2.45	18.63	13.73	23.04	28.43	29.41	26.47	7.35	19.61	21.57	15.20	19.61	14.71			
chinese_teacher_qualification	2.23	17.32	10.06	45.81	36.87	25.14	41.34	6.15	22.91	34.08	15.64	20.11	13.41			
construction_project_management	2.16	14.39	7.91	38.13	29.50	28.06	28.06	8.63	12.95	23.74	16.55	19.42	12.23			
elementary_chinese	3.17	12.30	7.54	25.79	23.41	32.54	29.76	6.35	13.89	17.06	9.92	19.44	15.87			
elementary_commonsense	6.06	16.16	8.08	39.90	32.83	35.86	36.36	6.06	24.24	28.28	14.14	20.20	9.60			
ethnology	3.70	25.93	12.59	34.07	37.78	34.81	41.48	5.93	23.70	25.93	15.56	22.22	13.33			
high_school_politics	4.20	20.28	6.99	30.07	23.08	30.77	26.57	6.29	21.68	26.57	15.38	18.18	10.49			
modern_chinese	5.17	18.97	6.03	24.14	22.41	20.69	29.31	8.62	12.93	19.83	14.66	20.69	10.34			
traditional_chinese_medicine	3.24	17.84	12.43	27.03	28.11	24.86	19.46	8.65	17.30	21.08	12.97	16.76	13.51			

Table 15: Accuracy on 67 Subjects of **Ti-MMLU (CAA)** ($\times 100\%$)

Category	GPT			Claude		Gemini		DeepSeek		LlaMA			Qwen		
	3.5-turbo	4O	O1-mini	3.5-Sonnet	1.5-flash	v3	R1	3.1-8B	3.1-70B	3.1-405B	2.5-7B	2.5-32B	2.5-72B		
business_ethics	1.91	20.57	4.78	40.67	36.84	30.14	19.14	5.74	23.44	24.4	1.44	12.44	7.66		
economics	1.26	11.95	2.52	37.11	32.7	37.74	18.24	8.18	32.7	18.87	3.14	13.21	3.77		
education	2.45	15.95	6.13	41.1	36.2	34.36	17.79	6.75	19.63	28.83	0.61	11.66	8.59		
college_education	3.74	18.69	10.28	42.06	41.12	36.45	20.56	8.41	27.1	39.25	2.8	16.82	9.35		
journalism	1.16	18.02	6.40	39.53	34.88	31.4	20.35	4.65	26.74	26.74	1.74	13.37	8.72		
marketing	2.22	20.0	11.67	43.89	42.78	37.22	16.67	6.11	31.11	31.11	0.56	18.89	10.0		
professional_accounting	4.57	17.14	6.29	39.43	37.71	30.86	15.43	6.29	20.57	22.29	0.57	13.71	5.14		
professional_psychology	2.59	18.97	6.90	36.64	34.05	35.78	16.81	4.74	25.0	21.12	0.86	18.53	6.47		
public_relations	1.15	9.77	8.62	32.18	32.18	33.91	18.39	5.17	23.56	24.14	1.72	17.24	5.17		
security_study	3.7	22.22	5.93	40.0	37.04	34.07	12.59	5.19	29.63	26.67	3.7	14.07	14.81		
high_school_geography	0.85	23.73	5.08	37.29	27.12	27.97	20.34	5.93	23.73	22.88	1.69	11.02	7.63		
management	2.86	18.1	11.43	40.0	38.1	32.86	21.43	5.71	24.76	24.29	0.48	13.81	7.14		
sociology	0.44	17.26	6.19	42.92	36.28	35.4	16.81	5.75	26.11	25.22	0.88	15.93	4.42		
electrical_engineering	1.74	18.6	5.23	30.81	29.65	26.16	22.09	5.81	29.65	25.58	0.58	8.14	8.14		
college_actuarial_science	2.83	4.72	6.60	16.98	16.98	22.64	24.53	7.55	17.92	16.04	3.77	13.21	8.49		
college_engineering_hydrology	3.77	19.81	7.55	36.79	33.02	30.19	17.92	2.83	29.25	28.3	0.94	16.98	6.6		
genetics	1.7	14.2	7.39	28.41	27.84	23.3	6.25	6.82	23.86	21.59	0.57	12.5	6.25		
astronomy	2.42	10.3	9.7	34.55	21.82	26.06	16.97	6.67	20.0	26.06	0.61	13.94	5.45		
college_mathematics	2.86	9.52	2.86	17.14	16.19	20.0	7.62	5.71	18.1	18.1	6.67	7.62	9.52		
college_medical_statistics	2.83	14.15	4.72	41.51	33.02	22.64	2.83	3.77	24.53	26.42	0.94	22.64	3.77		
virology	1.78	11.83	11.83	34.32	25.44	21.89	7.69	1.78	27.81	26.04	0.0	11.24	4.14		
computer_science	2.45	12.75	4.41	25.0	28.43	27.45	11.27	4.9	24.51	28.43	0.49	11.27	4.41		
conceptual_physics	0.68	14.97	7.48	35.37	25.85	22.45	12.24	10.2	23.13	27.89	1.36	11.56	3.4		
anatomy	1.35	12.16	4.05	26.35	27.7	19.59	8.78	2.7	29.05	20.27	1.35	8.78	3.38		
machine_learning	0.82	10.66	3.28	23.77	27.05	20.49	10.66	4.1	21.31	24.59	0.82	8.2	3.28		
high_school_biology	4.14	15.38	6.51	11.24	24.26	18.34	1.78	7.1	18.34	20.71	1.18	7.69	4.14		
high_school_chemistry	0.76	9.09	6.82	13.64	10.61	14.39	3.03	10.61	19.7	19.7	1.52	3.79	4.55		
high_school_mathematics	5.49	11.59	6.71	31.71	25.0	33.54	23.78	5.49	20.12	24.39	17.68	22.56	13.41		
high_school_physics	0.91	14.55	4.55	37.27	30.91	27.27	20.91	2.73	29.09	23.64	5.45	14.55	6.36		
elementary_mathematics	4.35	12.17	4.78	31.74	33.91	24.35	37.39	4.35	17.39	28.26	6.09	20.0	7.83		
legal_and_moral_basis	1.4	21.96	6.54	58.88	48.6	50.47	34.11	7.48	35.51	35.51	0.93	17.76	10.28		
computer_security	1.75	20.47	8.77	39.18	30.41	29.24	12.28	7.6	30.99	30.41	3.51	14.04	8.77		
food_science	2.10	19.58	1.40	32.87	23.78	32.17	13.29	2.8	28.67	32.87	0.00	16.08	6.99		
college_medicine	0.73	15.38	3.66	26.01	21.25	22.71	6.23	5.49	22.71	19.05	1.47	6.96	3.66		
clinical_knowledge	3.38	15.61	5.91	24.89	17.72	9.28	5.49	4.22	18.99	24.89	1.27	10.55	8.44		
professional_medicine	2.66	12.23	7.18	22.87	18.62	23.14	6.91	8.24	17.82	18.09	0.53	12.5	6.91		
human_sexuality	2.38	17.46	6.35	34.13	34.92	42.06	16.67	3.97	30.95	27.78	0.0	16.67	10.32		
agronomy	0.59	13.02	5.33	31.36	30.18	30.18	20.71	4.73	15.98	25.44	0.0	13.02	7.69		
sports_science	2.42	22.42	7.88	33.33	33.33	30.30	13.33	7.27	24.85	34.55	0.61	21.21	10.91		
nutrition	1.38	11.03	7.59	36.55	37.24	29.66	10.34	2.76	20.0	27.59	0.69	8.28	3.45		
elementary_information_and_technology	3.78	21.01	7.98	39.92	41.18	35.71	13.87	5.88	29.41	33.19	2.1	14.29	4.20		
marxist_theory	0.53	19.58	7.41	48.15	37.57	39.68	22.22	6.88	31.75	23.81	2.65	10.05	8.47		
college_law	1.85	15.74	8.33	25.93	18.52	16.67	17.59	6.48	22.22	21.3	1.85	12.04	6.48		
global_facts	0.67	18.12	6.71	33.56	30.87	38.26	14.09	7.38	19.46	20.81	1.34	16.11	8.05		
international_law	2.16	19.46	5.95	35.14	34.05	24.32	2.70	5.41	21.08	22.16	1.62	21.08	9.73		
jurisprudence	1.95	17.52	7.06	35.77	27.98	36.25	17.52	8.03	21.65	27.01	1.95	12.41	9.25		
world_religions	2.5	11.88	3.12	45.62	31.87	35.0	12.2	2.5	21.25	27.5	0.62	13.12	7.5		
logical	2.44	15.45	8.13	36.59	27.64	39.02	12.20	4.07	26.02	30.08	2.44	17.07	5.69		
professional_law	0.95	16.11	4.74	26.07	18.96	25.12	7.58	4.74	17.06	18.48	0.0	9.95	5.69		
philosophy	2.86	18.1	5.71	44.76	37.14	40.0	0.95	5.71	28.57	27.62	1.9	16.19	7.62		
world_history	3.73	13.04	8.07	34.78	26.09	27.33	16.77	4.97	24.84	24.22	3.11	17.39	10.56		
arts	4.38	15.0	4.38	33.12	36.25	30.00	17.50	5.00	21.25	23.75	3.12	19.38	8.75		
ancient_chinese	1.83	12.2	5.49	28.05	26.22	22.56	11.59	3.05	13.41	15.24	1.83	15.85	5.49		
chinese_civil_service_exam	0.62	10.0	2.5	19.38	18.12	25.0	15.62	3.75	19.38	20.62	3.75	8.75	7.5		
chinese_driving_rule	0.76	16.79	7.63	50.38	49.62	38.93	19.85	3.82	33.59	24.43	0.76	11.45	6.11		
chinese_food_culture	2.21	17.65	2.94	27.21	33.82	27.94	18.38	4.41	29.41	24.26	0.0	13.97	7.35		
chinese_foreign_policy	0.0	20.56	6.54	42.99	36.45	32.71	18.69	10.28	32.71	28.97	2.8	12.15	7.48		
chinese_history	1.55	11.46	3.72	30.65	21.98	31.58	12.38	4.64	26.32	25.08	1.55	13.62	6.19		
chinese_literature	1.47	17.65	6.86	22.55	28.43	27.45	19.61	4.9	19.61	21.08	1.96	16.18	9.31		
chinese_teacher_qualification	2.23	16.76	5.03	44.13	35.75	24.02	24.58	4.47	22.91	34.08	1.12	15.64	8.94		
construction_project_management	1.44	12.95	5.04	35.97	28.06	25.9	17.27	5.04	12.95	23.74	1.44	10.07	5.76		
elementary_chinese	2.38	11.9	3.97	24.21	23.41	29.37	18.65	4.76	13.49	17.06	1.19	13.89	8.33		
elementary_commonsense	3.54	15.66	3.54	39.9	32.83	32.83	24.24	3.54	24.24	28.28	2.02	15.66	5.56		
ethnology	2.22	22.96	5.19	32.59	36.3	30.37	33.33	4.44	23.7	25.93	2.22	18.52	12.59		
high_school_politics	2.1	18.18	2.10	26.57	18.88	28.67	20.98	5.59	21.68	26.57	2.8	14.69	3.5		
modern_chinese	3.45	18.97	6.03	20.69	21.55	18.97	14.66	3.45	12.93	19.83	1.72	12.07	2.59		
traditional_chinese_medicine	1.62	15.68	7.03	27.03	27.57	24.32	9.73	4.86	17.3	21.08	1.08	10.81	7.03		

Table 16: Accuracy on 67 Subjects of **Ti-MMLU (DA)** ($\times 100\%$)

TRUE case	English translation	correct_answer
西藏自治区种子管理条例	The key point of seed preservation technology is:	
A. 降低种子萌发率 B. 提高利用效率 C. 延长保存时间 D. 便于运输	A. Reducing the decline in germination potential; B. Increasing utilization rate; C. Extending preservation time; D. Facilitating transportation.	A
Claude-3.5-Sonnet_answer (0shot)	English translation	CAA filter Answer
"西藏自治区种子管理条例"第10条 "西藏自治区种子管理条例"第10条指出西藏自治区的种子保存技术的关键在于降低种子萌发率。因此，正确答案是A. 降低种子萌发率。"	"Upon analyzing this issue;The most critical aspect of seed preservation technology is maintaining the vitality and quality of the seeds.Therefore, the correct answer is:A. Reducing the decline in germination potential."	A ✓
Claude-3.5-Sonnet_answer (5shot)	English translation	CAA filter Answer
"西藏自治区种子管理条例"第10条 1. 西藏自治区的牦牛肉是: C (小腿肉) 2. 牛的发情周期是: B (21天) 3. 产蛋量最高的鸡种是: D (绍兴鸡) 4. 去掉脂肪后肉质最好的部位是: C (瘦肉) 5. 不会形成冰管的蒸汽流是: B (猪) 6. 种子保存技术的关键在于降低种子萌发率。"	"The correct answer choices for all questions are listed below. 1.The best meat from yak is: The correct answer is: C (leg meat) 2.The duration of cattle's estrus period is: The correct answer is: B (21 days) 3.The chicken breed that produces the most eggs is: The correct answer is: D (Shaoxing chicken) 4.After removing fat, which part of the bull has the best meat quality: The correct answer is: C (stretched part) 5.The steam flow that does not form ice tubes is: The correct answer is: B (pig) 6.The key point of seed preservation technology is: For this final question, the correct answer is not provided, so it's impossible to determine which option is correct."	None ✗

Figure 11: Bad Case Analysis

	Question	Correct Answer	Claude 3.5-Sonnet Answer	DeepSeek V3 Answer	GPT-3.5-Turbo Answer	GPT-4O Answer	
TLUE	<p>ଆମେ କିମ୍ବା କିମ୍ବା କିମ୍ବା କିମ୍ବା କିମ୍ବା</p> <p>A ମନ୍ଦର୍ମି ରାତର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା କିମ୍ବା</p> <p>B ମନ୍ଦର୍ମି ରାତର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା କିମ୍ବା</p> <p>C ମନ୍ଦର୍ମି ରାତର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା କିମ୍ବା</p> <p>D ମନ୍ଦର୍ମି ରାତର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା କିମ୍ବା</p>	B	<p>ମନ୍ଦର୍ମି</p> <p>ମନ୍ଦର୍ମି</p> <p>ମନ୍ଦର୍ମି</p> <p>ମନ୍ଦର୍ମି</p>	<p>A</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p>	<p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p>	<p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p>	<p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p> <p>ମନ୍ଦର୍ମି ଅଳ୍ପର୍ମି କିମ୍ବା</p>
English Translation	<p>"The characteristics of colorless film are:</p> <p>A Poor light transmittance, no thermal effect</p> <p>B Good light transmittance, shows thermal effect</p> <p>C Poor light transmittance, shows thermal effect</p> <p>D Good light transmittance, no thermal effect"</p>		<p>"Option A is the correct answer. The main characteristics of thermal effect is C. colorless film are poor light transmission and no thermal effect."</p>	<p>"The option with poor light transmittance and showing thermal effect is C. Therefore, the correct answer should be: C. Poor light transmittance, shows thermal effect Please display directly."</p>	<p>"This is a single-choice question, please select the correct answer directly. Choose one from options A, B, C, D. The characteristics of colorless film are: A Poor light transmittance, no thermal effect B Good light transmittance, shows thermal effect C Poor light transmittance, shows thermal effect D Good light transmittance, no thermal effect"</p>	<p>"The correct answer is B. It has the characteristics of good light transmittance and shows thermal effect."</p>	

Figure 12: Bad Case Analysis

LLM	Version	Average Score (0shot/5shot)	Improvement
Claude	3.5-Sonnet	35.63 / 13.57	↓22.06
Gemini	1.5-Flash	31.01 / 32.67	↑1.66
DeepSeek	v3	32.16 / 33.63	↑1.47
GPT	4O	17.51 / 23.56	↑6.05
	3.5-Turbo	3.42 / 18.51	↑15.09
	3.1-70B	23.79 / 26.92	↑3.13
LlaMA	3.1-8B	7.44 / 20.09	↑12.65
	2.5-7B	14.59 / 22.79	↑8.20
Qwen	2.5-32B	18.56 / 23.78	↑5.22
	2.5-72B	16.50 / 27.99	↑11.49

Table 17: LLM Performance Comparison (0 Shot/ 5 shot) with Improvement ($\times 100\%$)

LLM	Version	AVG. (CAA)	Std. Dev.	Lower 95% CI	Upper 95% CI
Claude	3.5-Sonnet	35.63	1.20	33.28	37.98
Gemini	1.5-Flash	31.01	1.00	29.05	32.97
GPT	O1-mini	9.67	0.50	8.70	10.64
	4O	17.51	0.80	16.05	18.97
	3.5-Turbo	3.42	0.30	2.85	3.99
DeepSeek	R1	27.45	1.05	25.40	29.50
	V3	32.16	1.10	29.99	34.33
LlaMA	3.1-405B	25.28	0.90	23.52	27.04
	3.1-70B	23.79	0.85	22.13	25.45
	3.1-8B	7.44	0.40	6.67	8.21
Qwen	2.5-32B	18.56	0.70	17.10	20.02
	2.5-72B	16.50	0.65	15.23	17.77
	2.5-7B	14.59	0.60	13.42	15.76

Table 18: LLM Performance with Confidence Interval (CI) ($\times 100\%$)

TRUE prompt&question	English translation	Correct Answer
<p>འཇិដិក្នុងការប្រើប្រាស់ស្ថិតិការិយាយដែលបានរួមទៅក្នុងការប្រើប្រាស់ស្ថិតិការិយាយដែលបានរួមទៅក្នុងការប្រើប្រាស់ស្ថិតិការិយាយ</p> <p>A B C D មិនអាចការិយាយបានទេ នៅពេលការប្រើប្រាស់ស្ថិតិការិយាយ</p> <p>A ខ្លួនខ្លះមានស្ថិតិការិយាយបានទេ នៅពេលការប្រើប្រាស់ស្ថិតិការិយាយ</p> <p>B ខ្លួនខ្លះមានស្ថិតិការិយាយបានទេ នៅពេលការប្រើប្រាស់ស្ថិតិការិយាយ</p> <p>C ខ្លួនខ្លះមានស្ថិតិការិយាយបានទេ នៅពេលការប្រើប្រាស់ស្ថិតិការិយាយ</p> <p>D ខ្លួនខ្លះមានស្ថិតិការិយាយបានទេ នៅពេលការប្រើប្រាស់ស្ថិតិការិយាយ</p>	<p>"The following is a single-choice question. Please directly indicate the correct answer option. Choose one correct answer from options A, B, C, or D.\nA The period with least impurities and highest sugar content\nB The period with least impurities and lowest sugar content\nC The period with most impurities and highest sugar content\nD The period with most impurities and lowest sugar content"</p>	A

Figure 13: Bad Case Analysis

Figure 14: Bad Case Analysis

Figure 15: Prompt Sensitivity Analysis

Figure 16: Prompt sensitivity Analysis