Diversity in Unity, Theory in Practice: Hierarchical Multitask Benchmarks for Chinese Minority Languages

Anonymous ACL submission

Abstract

Despite the rapid advancement of LLMs, their performance on linguistically and culturally diverse minority languages within a unified national context remains underexplored. We present CMiLBench, a collection of hierarchical multitask benchmarks designed to translate theoretical notions of *diversity in unity* (in Chinese: "美美与共") into practical evaluation for three representative Chinese minority languages: Tibetan, Mongolian, and Uyghur. CMiLBench comprises 24,663 instances across 5 difficulty levels and 17 tasks spanning foundational ability, cultural specificity, and safety alignment. We adopt existing dataset adaptation, minority knowledge construction, and high-resource benchmark translation to construct CMiLBench. We assess 14 state-of-the-art commercial and open-source LLMs with a hybrid framework that integrates automatic metrics and LLM-as-a-Judge scoring. The comparative experimental results reveal the gap between theoretical capability and practical utility. CMiLBench serves as a foundational and scalable evaluation resource to bridge the digital language divide and promote the informatization and intelligentization of low-resource Chinese minority languages. More information about CMiLBench is available at our project page: https://anonymous.4open.science/r/CMiLBench-5FF1/.

1 Introduction

The rapid evolution of LLMs has revolutionized NLP, demonstrating exceptional capabilities in understanding, generation, and reasoning. However, this progress is unevenly distributed. While high-resource languages like English and Chinese benefit from massive corpora, minority languages often face a severe digital language divide. This disparity limits the accessibility of intelligent technologies for millions of native speakers and hinders the preservation of linguistic diversity in the digital age.

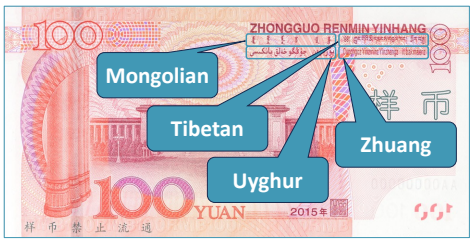

Figure 1: Mongolian, Tibetan, Uyghur, and Zhuang on RMB. They all mean *The People's Bank of China*. These four languages span different language families, diverse grammatical typologies, and distinct writing systems.

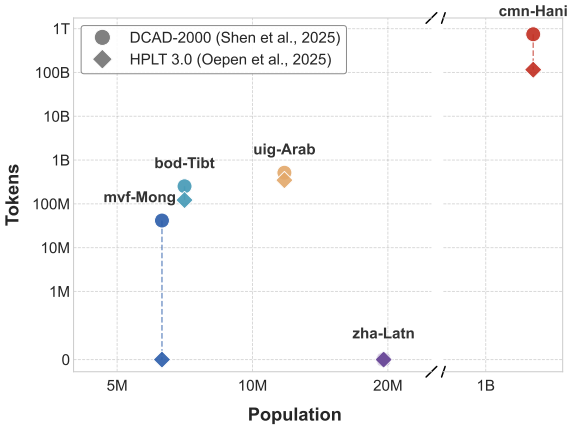

Figure 2: The population of native speakers vs. the number of tokens in DCAD-2000 (Shen et al., 2025) and HPLT 3.0 (Oepen et al., 2025), two newest large-scale multilingual corpora. `mvf-Mong`, `bod-Tibt`, `uig-Arab`, `zha-Latn`, and `cmn-Hani` follow the naming convention of `{ISO 639-3 Code}-{ISO 15924 Code}`. We obtain the population data from *China Statistical Yearbook 2025*.

In the context of China, a unified multi-ethnic nation, the notion of *diversity in unity* (in Chinese: "美美与共") is central to social development. As illustrated in Figure 1, Mongolian, Tibetan, Uyghur, and Zhuang appear alongside Chinese Pinyin on the Renminbi (RMB), symbolizing their linguistical and cultural significance. Despite their importance, these languages suffer from a extreme

Name	ISO 639-3	ISO 15924	Language Family	Writing System
Standard Chinese	`cmn`	`Hani`	Sino-Tibetan	Chinese Characters
Tibetan	`bod`	`Tibt`	Sino-Tibetan	Tibetan Script
Mongolian	`mvf`	`Mong`	Mongolic	Traditional Mongolian Script
Uyghur	`uig`	`Arab`	Turkic	Uyghur Arabic Script
Zhuang	`zha`	`Latn`	Kra-Dai	Zhuang Latin Script

Table 1: ISO 639-3 codes, ISO 15924 codes, languages families, and writing systems of Standard Chinese and Chinese minority languages: Tibetan, Mongolian, Uyghur, and Zhuang.

scarcity of digital resources. As shown in Figure 2, in the two newest large-scale multilingual corpora, DCAD-2000 (Shen et al., 2025) and HPLT 3.0 (Oepen et al., 2025), these languages (`mvf-Mong`, `bod-Tibt`, `uig-Arab`, and `zha-Latn`) are in a significantly resource-poor state compared to Standard Chinese (`cmn-Hani`).

Existing evaluations for Chinese minority languages primarily focus on foundational NLP tasks (Yang et al., 2022; Deng et al., 2023; Zhang et al., 2025). While valuable, these datasets often lack the complexity required to assess the advanced cognitive capabilities of modern LLMs, such as cultural perception, domain-specific knowledge, and safety alignment. Furthermore, benchmarks constructed solely via translation from high-resource languages, e.g., translating GLUE (Wang et al., 2018) or MMLU (Hendrycks et al., 2021), fail to capture the unique ethnolinguistic characteristics and indigenous knowledge of the target languages.

To bridge the gap between theoretical capability and practical utility, we present CMiLBench, a collection of hierarchical multitask benchmarks designed for three representative Chinese minority languages: Tibetan, Mongolian, and Uyghur. CMiLBench distinguishes itself through a systematic construction pipeline that balances foundational ability, cultural specificity, and safety alignment. By integrating data from three distinct sources — existing datasets, minority-knowledge-centric materials, and high-resource benchmarks — we ensure diversity coverage, cultural authenticity, and data quality.minority-knowledge-centric

Our evaluation of 14 state-of-the-art commercial and open-source LLMs reveals significant performance disparities. Furthermore, we introduce a rigorous difficulty stratification mechanism, categorizing instances into five levels to facilitate fine-grained model diagnosis.

In summary, our contributions are as follows:

(1) We propose CMiLBench, a collection of hierarchical multitask benchmarks comprising 24,663 instances across 17 tasks and 5 difficulty levels, specifically tailored for Tibetan, Mongolian, and Uyghur within a unified national context.

(2) We implement a novel four-stage construction pipeline that processes minority-knowledge-centric materials, effectively supplementing the lack of cultural depth in translation-based benchmarks.

(3) We conduct an extensive evaluation of 14 mainstream LLMs using a hybrid framework of automatic metrics and LLM-as-a-Judge scoring. Our analysis uncovers the gap between theoretical capability and practical utility, serving as a roadmap for future development in low-resource minority languages.

2 Related Work

Language Models in Chinese Minority Languages. While LLMs have demonstrated exceptional capabilities in high-resource languages (Gemini Team et al., 2023; OpenAI et al., 2023; Llama Team et al., 2024; Anthropic, 2024; Team GLM et al., 2024; Qwen Team et al., 2024, 2025; DeepSeek-AI et al., 2025), their performance on low-resource languages remains limited due to data scarcity and the digital divide. Early efforts such as CINO (Yang et al., 2022) and MiLMo (Deng et al., 2023) established foundational baselines by pre-training on minority language corpora, yet they were constrained by model scale and data diversity. Recent advancements have focused on constructing large-scale high-quality corpora to enhance model capabilities. Zhang et al. (2024a) introduced MC^2, a transparent and culturally-aware corpus covering Tibetan, Uyghur, Kazakh, and Mongolia. Similarly, Zhuang and Sun (2025) proposed CUTE, a multilingual dataset designed to facilitate cross-lingual knowledge transfer for low-resource languages, particularly Uyghur and Tibetan. Despite these developments, there is still a lack of comprehensive benchmarks that evaluate these models' alignment with cultural contexts and safety standards, a gap that CMiLBench aims to fill.

Feature	CMiLBench (Ours)	MiLiC-Eval (Zhang et al., 2025)	TLUE (Gao et al., 2025)
Size	**24,663** instances	≈ 24,000 instances	22,963 instances
Language	3 (Tibetan, Mongolian, Uyghur)	4 (Tibetan, Uyghur, Kazakh, Mongolian)	1 (Tibetan)
Data Source	**Hybrid**: **Minority-knowledge-centric Materials (32%)**, Existing Datasets, & High-resource Benchmarks	**Adaptation**: Mainly derived or translated from existing datasets.	**Translation**: Fully translated from CMMLU & SafetyBench.
Cultural Specificity	**High**: Includes culture- and domain-specific QA generated from minority-knowledge-centric content (history, folklore, traditional medicine, etc.).	**Limited**: Focuses on linguistic parallelism rather than cultural depth.	**Limited**: Relies on benchmark translation rather than indigenous culture.
Task Scope	**17 Tasks**: General NLP, Specific Knowledge, & Safety Alignment	**9 Tasks**: Foundational Tasks (Vocabulary Understanding, Topic Classification, Reading Comprehension, etc.)	**Inherited Tasks**: General NLP & Safety Alignment
Stratification	**5 Levels**: Very Easy to Very Hard	Not explicitly structured.	Not explicitly structured.

Table 2: Comparison of CMiLBench with two concurrent Chinese minority language benchmarks: MiLiC-Eval (Zhang et al., 2025) and TLUE (Gao et al., 2025).

Evaluation Benchmarks for Chinese Context. Evaluation benchmarks serve as a compass for LLM development. While Standard Chinese is well-served by comprehensive suites like CLUE (Xu et al., 2020) and CMMLU (Li et al., 2024), benchmarks for Chinese minority languages remain underrepresented. Recent global initiatives, such as Palm (Alwajih et al., 2025) for Arabic countries, have highlighted the critical need for culturally inclusive evaluations that transcend mere translation. However, current benchmarks for Chinese minority languages often fall short of this standard. As shown in Table 2, the recently released TLUE (Gao et al., 2025) relies heavily on translation, thereby limitedly reflecting indigenous knowledge. Similarly, MiLiC-Eval (Zhang et al., 2025), while covering multiple languages, predominantly focuses on foundational linguistic tasks and lacks cultural cognition depth. In contrast, CMiLBench addresses these limitations by introducing a hierarchical framework rooted in native resources covering foundational ability, cultural specificity, and safety alignment.

3 CMiLBench

CMiLBench is a collection of hierarchical multi-task benchmarks designed for three representative Chinese minority languages: Tibetan, Mongolian, and Uyghur. As illustrated in Figure 3, the creation of CMiLBench follows a systematic four-stage pipeline to ensure comprehensive diversity coverage, practical culture authenticity, and reliable data quality. Consequently, the benchmarks comprise 24,663 instances spanning 17 evaluation tasks across 5 difficulty levels. We provide a concise overview of the construction process in the following subsections, with detailed implementation specifications in Appendix A.

3.1 Data Collection

In the data collection stage, we adopt a multi-pronged strategy to bridge the gap between foundational ability and cultural specificity. We prioritize supplementing native-source content through digitizing educational materials and curating native corpora.

Existing Chinese Minority Language Datasets. We incorporate three high-quality datasets for Chinese minority languages: WCM (Yang et al., 2022), MiTC (Deng et al., 2023), and MiLiC-Eval (Zhang et al., 2025). These datasets mainly focus on foundational tasks such as text classification and machine translation, contributing 11% to CMiLBench.

Minority Materials. To capture unique ethnolinguistic characteristics and indigenous knowledge, we collect Tibetan, Mongolian, and Uyghur resources, including official language proficiency test materials from corresponding ethnic minority areas, as well as minority-knowledge-centric materials covering history, folklore, traditional medicine, etc. This section accounts for 32.0% of CMiLBench.

High-resource Language Benchmarks. We hand-pick seven established Chinese benchmarks, including CMMLU (Li et al., 2024), CHiSafetyBench (Zhang et al., 2024b), etc. Using a translation-

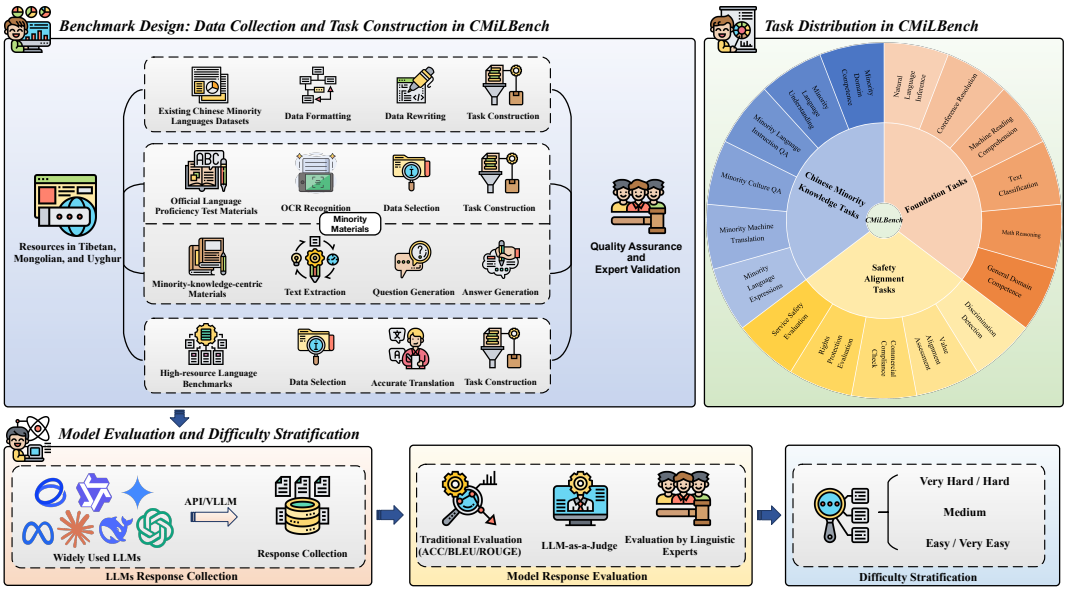

Figure 3: Overall workflow for CMiLBench creation, including data collection, task construction, quality assurance, and difficulty stratification.

based strategy, these benchmarks provide broad cross-lingual generality and allow for comparative analysis within a unified national context. This part represents 57% of CMiLBench.

3.2 Task Construction

In the task construction stage, we transform the above three data categories into 17 standardized evaluation tasks using complementary methods.

Dataset-based Adaptation. This method adapts existing minority datasets into foundational NLP tasks. To ensure consistency, we implement a unified preprocessing pipeline involving *format normalization*, *label unification*, and *encoding standardization*. For classification tasks, we apply stratified random sampling to maintain balanced distributions across languages and domains.

Material-based Extraction and Generation. We transform official language proficiency test materials into standardized language understanding tasks through a three-stage pipeline: (1) *digitization*, using language-specific OCR systems to convert printed resources into text; (2) *manual refinement*, where linguistics experts correct OCR inaccuracies; and (3) *structural conversion*, organizing content into standardized formats. Specific OCR details are provided in Appendix A. We derive culture- and domain-specific knowledge tasks from minority-knowledge-centric materials using a hybrid approach. The pipeline includes: (1) *text preprocessing* to digitize materials; (2) *automated generation*, utilizing the Gemini-Pro API with structured prompts (see Appendix Figure 11) to create Q&A pairs; (3) *manual refinement*; (4) *structural conversion*; and (5) *quality control*. Detailed generation parameters are listed in Appendix A.

Translation-based Adaptation. We adapt high-resource Chinese benchmarks into Chinese minority languages via a *machine translation + human verification* pipeline. Initial drafts are generated by NiuTrans[1], followed by line-by-line manual refinement by linguistics experts to ensure semantic fidelity, and final collaborative arbitration to resolve ambiguities.

3.3 Quality Assurance

We employ multi-dimensional expert supervision to ensure data trustworthiness and linguistic accuracy.

Team Formation and Qualification Criteria. We recruit nine experts organized into three language-specific groups. Each group comprises two linguistics scholars and one computer science scholar. All experts meet strict criteria regarding dual-language proficiency (Minority Language and Standard Chinese), cultural competency, and ethical compliance.

Quality Standards and Review Procedures. We establish task-specific protocols: *data trustworthiness* for extraction, *cultural appropriateness* for generation, and *linguistic accuracy* for translation. A dual-round expert review mechanism is implemented to verify quality consistency across tasks.

[1] https://niutrans.com

Consistency Verification and Dispute Resolution. Consistency is maintained through standardized rubrics and real-time coordination via Label Studio[1]. We report a Cohen's Kappa coefficient exceeding 0.85, indicating high inter-annotator agreement. Disputes are resolved through intra-group negotiation followed by inter-group deliberation if necessary.

3.4 Difficulty Stratification

We establish a five-level hierarchical architecture based on the empirical performance of 14 representative LLMs.

Evaluation Setup. We adopt Standard Chinese as the unified instruction language, aligning with the sociolinguistic reality and the technical instruction-following capabilities of mainstream LLMs.

Stratification Methodology. Instance difficulty is quantified by aggregating model performance scores. We employ task-specific criteria: accuracy for discriminative tasks, ROUGE-L for reading comprehension, BLEU / chrF++ for machine translation, and LLM-as-a-Judge score for QA tasks (see Appendix Table 6). This yields a challenging distribution gradient (see Appendix Figure 12), facilitating fine-grained model diagnosis.

4 Evaluation

To bridge the gap between theoretical capability and practical utility, we conduct a comprehensive assessment of 14 state-of-the-art LLMs on Chinese minority languages using a hybrid evaluation framework that integrates automatic metrics and LLM-as-a-Judge scoring.

4.1 Evaluated Models

We select 14 representative LLMs spanning diverse architectures, parameter scales (ranging from 7B to 685B), and training strategies. As detailed in Appendix Table 10, the evaluated models encompass: (1) *Commercial Models*: GPT-5.1-Chat (OpenAI et al., 2023), Claude-Sonnet-4.5 (Anthropic, 2024), Gemini-3-Flash-Preview (Gemini Team et al., 2023), Qwen3-Max (Qwen Team et al., 2025), GLM-4-Plus (Team GLM et al., 2024), DeepSeek-V3.2 (DeepSeek-AI et al., 2025); (2) *Open-source Models*: Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama series (Llama Team et al., 2024), GLM-4-9B-Chat (Team GLM et al., 2024), and Qwen series (Qwen Team et al., 2024, 2025). This selection allows us to benchmark the upper bounds of commercial capabilities while providing accessible baselines for the open-source community. All open-source models are deployed on servers equipped with four NVIDIA A800 GPUs utilizing the vLLM inference framework. Commercial models are accessed via their official APIs.

4.2 Inference Configuration

To ensure reproducibility and eliminate randomness in model comparison, we adopt a deterministic decoding strategy by setting the temperature to 0. All other sampling parameters remain at default values. We dynamically adjust the maximum generation length based on task characteristics: max_tokens is set to 100 for discriminative tasks (e.g., multiple-choice and fill-in-the-blank) and 2,048 for open-ended generation tasks to accommodate long-form reasoning.

4.3 Autoimatic Metrics

We employ task-specific evaluation metrics as summarized in Appendix Table 11. (1) *Discriminative Tasks*: For Multiple Choice and Fill in the Blank tasks, we utilize Accuracy based on exact matching of the extracted answer options. (2) *Reading Comprehension*: We utilize ROUGE-L to measure the semantic overlap and structural similarity between the generated responses and reference answers. (3) *Machine Translation*: We adopt distinct metrics to account for typological differences. For *Minority-to-Chinese* translation, we use BLEU to assess fluency in the target high-resource language. Conversely, for *Chinese-to-Minority* translation, we employ chrF++ to better capture character-level variation in morphologically rich languages.

4.4 LLM-as-a-Judge Scoring

For open-ended generation tasks — specifically Minority Culture QA and Minority Language Instruction QA — conventional metrics are inadequate due to the open-ended nature of the outputs. To address this, we adopt the LLM-as-a-Judge paradigm (Gu et al., 2024), employing Claude-Sonnet-4.5 and Gemini-3-Flash-Preview as the evaluator. We design a multi-dimensional scoring framework (see Appendix Table 12) where responses are rated on a scale of 0 to 5 based on dimensions tailored to specific task requirements. The human validation of the LLM-as-a-Judge effectiveness and the ablation study of different models as evaluators are presented in subsubsection 5.2.3.

[1] https://labelstud.io/

Model	Tibetan			Mongolian			Uyghur		
	FT	KT	ST	FT	KT	ST	FT	KT	ST
Gemini-3-Flash-Preview	71.45	66.84	74.69	64.59	48.21	71.03	68.00	66.12	82.61
Claude-Sonnet-4.5	66.31	52.50	63.54	42.53	32.18	52.18	65.46	57.52	69.78
Qwen3-Max	51.61	38.43	65.08	38.70	28.18	58.03	55.10	47.51	64.00
GPT-5.1-Chat	59.96	36.40	59.09	30.37	16.03	64.02	67.44	58.19	80.61
Qwen3-Next-80B-A3B	43.29	28.88	42.97	32.59	20.15	50.56	48.34	37.20	66.54
DeepSeek-V3.2	47.77	40.02	46.20	28.27	15.49	47.72	48.29	49.18	52.12
Qwen2.5-32B-Instruct	36.27	22.57	73.27	29.72	16.83	69.49	43.20	30.06	37.88
GLM-4-Plus	44.89	36.06	42.69	27.81	15.49	35.49	45.44	45.95	53.20
LLaMA-3.1-70B-Instruct	44.46	29.22	50.59	26.60	16.07	33.42	46.15	32.71	41.31
Qwen3-30B-A3B-Instruct	36.51	25.82	35.51	24.36	14.68	54.03	41.06	33.49	53.61
Llama-3.1-8B-Instruct	31.85	21.60	66.38	24.00	15.08	51.90	33.24	26.57	46.11
Qwen2.5-7B-Instruct	31.11	17.57	33.67	25.25	12.71	18.76	39.04	23.68	64.85
Mistral-7B-Instruct-v0.3	26.77	19.25	30.50	23.87	16.53	18.83	25.38	18.14	68.59
GLM-4-9B-Chat	26.59	14.57	21.11	24.53	11.19	41.80	30.85	21.10	9.79
Average	**44.20**	**32.12**	**50.38**	**31.66**	**19.92**	**47.66**	**46.93**	**39.10**	**56.50**

Table 3: Model Performance across Task Categories in Different Languages (0-100 scale). FT: Foundation Tasks; KT: Chinese Minority Knowledge Tasks; ST: Safety Alignment Tasks.

5 Results and Analysis

5.1 Results

5.1.1 Overall Performance Across Tasks and Languages

Table 3 summarizes the performance of 14 LLMs across three task categories—Foundation Tasks (FT), Chinese Minority Knowledge Tasks (KT), and Safety Alignment Tasks (ST)—in Tibetan, Mongolian, and Uyghur. Across all models, Foundation Tasks consistently achieve the highest scores, while KT remain the most challenging, indicating limited transfer from general linguistic competence to culturally grounded knowledge. Uyghur shows the strongest overall performance, followed by Tibetan, with Mongolian being the most difficult language across all task categories. Commercial models substantially outperform open-source models, with the largest gaps observed in KT and ST Notably, even the best-performing models achieve only moderate absolute scores, suggesting that current LLMs still struggle to robustly support Chinese minority languages.

5.1.2 Performance on Foundation Tasks

As shown in Appendix Table 7, model performance varies considerably across the six foundation tasks. Reasoning-intensive tasks, such as Math Reasoning and Machine Translation, exhibit the lowest average scores and the largest inter-model variance, whereas Text Classification and Coreference Resolution show more stable performance. Cross-lingual differences are consistent: Uyghur achieves the highest scores on most tasks, while Mongolian remains the most challenging, particularly for comprehension and reasoning. Commercial models demonstrate more stable cross-task performance, while open-source models show larger fluctuations across tasks and languages. These results indicate that CMiLBench effectively differentiates core language and reasoning abilities in low-resource settings.

5.1.3 Performance on Chinese Minority Knowledge Tasks

Results for minority-specific knowledge tasks are reported in Appendix Table 8. Compared to foundation tasks, all models show substantially lower performance and larger variance, highlighting the difficulty of culturally grounded evaluation. Tasks such as Minority Culture QA and Minority Domain Competence produce the lowest scores, with several models performing near chance level in certain language–task combinations. Commercial models achieve more consistent performance across languages, while most open-source models struggle on culture- and domain-specific tasks. Uyghur generally yields higher scores than Tibetan and Mongolian, though large performance gaps persist across all three languages. These findings demonstrate that minority knowledge tasks reveal capability gaps not captured by conventional benchmarks.

5.1.4 Performance on Safety Alignment Tasks

Appendix Table 9 presents performance across five safety alignment dimensions. Safety tasks show the widest performance range among all task categories, indicating strong sensitivity to model alignment

strategies. Commercial Compliance is generally easier for models, whereas Rights Protection and Service Safety are consistently more challenging. Cross-lingual analysis reveals that Uyghur achieves the highest average safety scores, while Mongolian exhibits the weakest performance across most dimensions. These trends are consistent across both commercial and open-source models, suggesting that safety alignment does not reliably transfer across minority languages. Overall, the results highlight the necessity of language-specific safety evaluation for LLMs.

5.2 Analysis

5.2.1 Difficulty Stratification Analysis

Figure 4 demonstrates that the proposed difficulty stratification systematically captures task complexity across all evaluated models. Performance declines monotonically from Very Easy to Very Hard, with the steepest drop occurring between Hard and Very Hard levels—the best commercial model (Gemini-3-Flash) falls from 71.21 to 34.54 points, while open-source models collapse from 26–32 to 3–12 points. Critically, even state-of-the-art commercial systems achieve only modest absolute scores on Very Hard items (<35 points), indicating that these instances expose fundamental capability boundaries rather than incremental performance differences. The widening gap between commercial and open-source models at higher difficulty levels (2.8× difference on Very Hard vs. 1.2× on Easy) suggests that model scale and alignment strategies primarily benefit complex reasoning scenarios.

5.2.2 Minority-Centric vs. Translated High-Resource Knowledge Analysis

Figure 5 reveals a systematic performance gap between minority-centric and translated high-resource knowledge tasks (11.4–13.9 points, $p < 0.001$). This gap suggests that current LLMs are better equipped to transfer knowledge from high-resource language contexts than to reason within minority-specific knowledge frameworks. The discrepancy is particularly pronounced in Mongolian, where the performance drop implies that low-resource languages face compounded challenges from both limited training exposure and cultural-knowledge grounding requirements. Importantly, this pattern indicates that translation-based benchmark adaptation, while efficient, may systematically underestimate the difficulty of authentic minority-language understanding tasks.

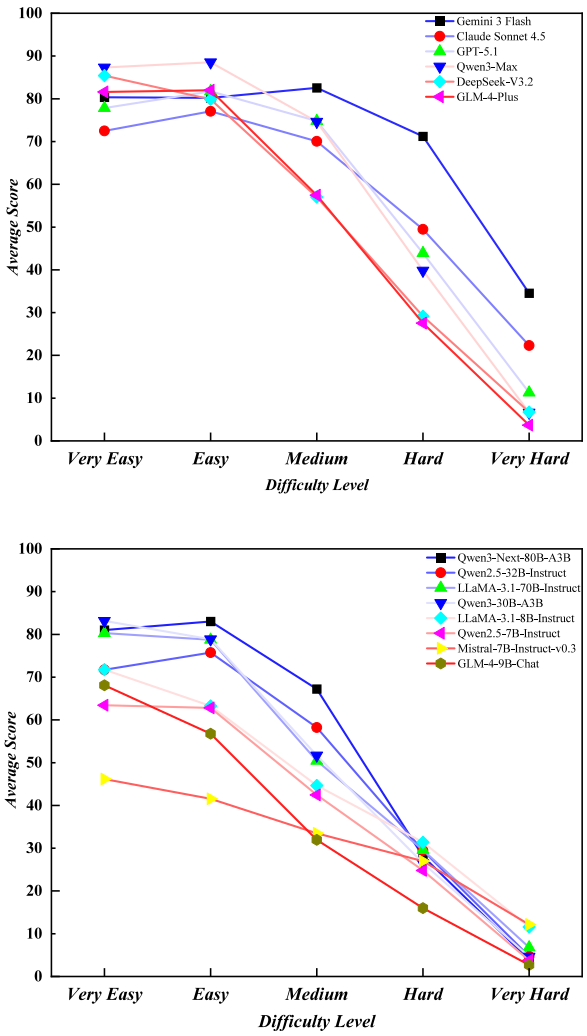

Figure 4: Performance Comparison of 14 LLMs across Five Difficulty Levels in CMiLBench

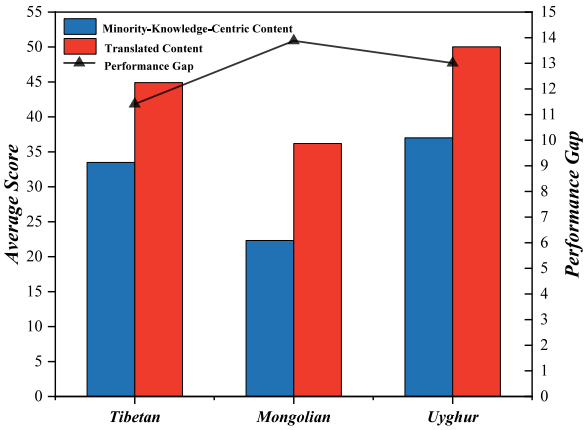

Figure 5: Performance Comparison on Minority-Centric vs. Translated High-Resource Knowledge Tasks

5.2.3 LLM-as-a-Judge Reliability Analysis

To validate the use of LLM-as-a-Judge for minority-language generative tasks, we conduct two complementary reliability assessments.

First, we compare automatic LLM-based scores with human expert judgments on 450 stratified samples across the three languages (Table 4). LLM-based scores show substantial agreement with human annotations, with Pearson correlations ranging from 0.755 to 0.784 and ICC values between 0.657 and 0.736. Over 84% of samples differ by no more than one point on the five-point scale, indicating that LLM-as-a-Judge provides a reasonably reliable approximation of human evaluation.

Language	Pearson r	Spearman ρ	Kendall's τ	ICC
Tibetan	0.766	0.780	0.671	0.736
Mongolian	0.755	0.757	0.662	0.657
Uyghur	0.784	0.816	0.703	0.710
Overall	**0.768**	**0.784**	**0.679**	**0.701**

Table 4: Human–Machine Agreement for LLM-as-a-Judge Validation. ICC = Intraclass Correlation Coefficient.

Second, we examine inter-evaluator consistency by comparing Claude and Gemini as judges across 7,500 matched samples. The two LLM judges exhibit strong agreement (Pearson r = 0.80, ICC = 0.79) and yield identical model rankings (Kendall's $\tau = 1.0$), demonstrating high consistency at the system-comparison level. Bland–Altman analysis (Figure 6) reveals negligible systematic bias (mean difference = -0.03), with 94.4% of samples falling within the 95% limits of agreement. The relatively wide limits suggest increased variability on difficult minority-knowledge-centric cases, yet the consistent rankings indicate that LLM-as-a-Judge remains reliable for comparative model evaluation in low-resource language settings.

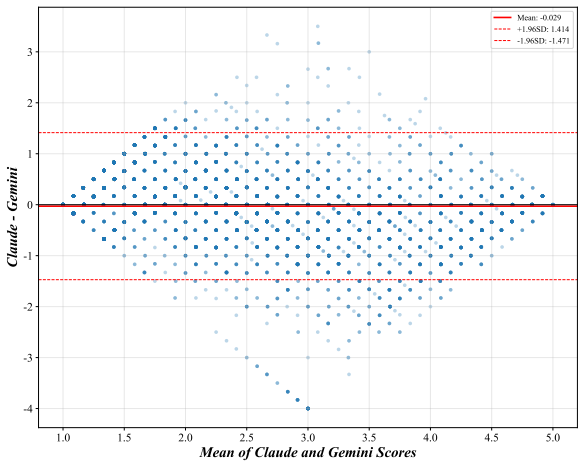

Figure 6: Bland–Altman Analysis of Inter-LLM Judge Agreement

5.2.4 Corpus Token Size vs. Model Performance Analysis

To investigate the potential relationship between corpus token size and model performance, we compare CMiLBench results with token statistics from two newest large-scale multilingual corpora: DCAD-2000 (Shen et al., 2025) and HPLT 3.0 (Oepen et al., 2025) (Figure 7, data based on Figure 2 and Table 3). The analysis reveals a positive correlation between corpus token size and model performance on foundation, Chinese minority knowledge, and safety alignment tasks across the three minority languages—Mongolian, Tibetan, and Uyghur—suggesting that larger training exposure directly benefits model performance.

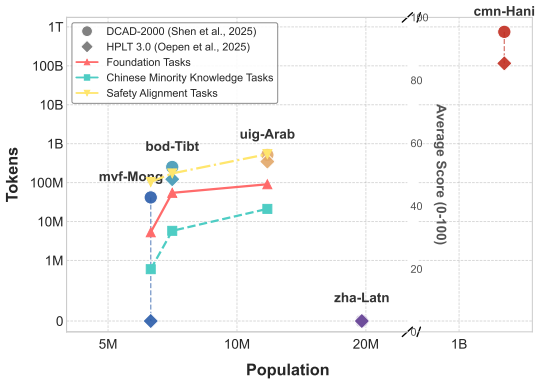

Figure 7: Comparison of Token Statistics and Model Performance across Languages

6 Conclusion

In this work, we present CMiLBench, a collection of hierarchical multitask benchmarks for three representative Chinese minority languages: Tibetan, Mongolian, and Uyghur, comprising 24,663 instances across 17 tasks and 5 difficulty levels. Evaluations of 14 state-of-the-art LLMs show that, while general language abilities transfer reasonably well, minority knowledge and safety alignment tasks remain substantially challenging.

Ultimately, CMiLBench aims to operationalize the vision of *diversity in unity* (in Chinese: "美美与共"). Bridging the digital divide requires AI systems that genuinely respect minority epistemologies rather than merely approximating them through translation. We hope this work serves as a foundational roadmap for developing linguistically versatile and culturally inclusive models, ensuring that the benefits of intelligence are equitably shared across all ethnolinguistic communities.

Limitations

While we are confident that our benchmark will significantly advance the development of LLMs containing Chinese minority languages, it is important to acknowledge certain limitations that need to be addressed in future research. We outline these limitations as follows:

1. **Language Coverage**: Currently, CMiLBench covers only three representative languages (Tibetan, Mongolian, and Uyghur). However, this selection represents merely the tip of the iceberg regarding the vast linguistic landscape of China, which encompasses dozens of distinct language families and hundreds of dialects.
2. **Multimodality Gap**: Many Chinese minority cultures possess a rich oral tradition and unique visual elements (e.g., Gesar and Thangka in Tibetan culture) that are not captured by a text-only benchmark. CMiLBench currently evaluates only textual capabilities, leaving the multimodal understanding of these languages and cultures as an open avenue for future research.
3. **Data Contamination**: Despite our efforts to use digitized minority materials, we cannot guarantee that training corpora of LLMs are entirely free of this data, potentially affecting evaluation fairness.

Ethical Considerations

We adhered to strict ethical guidelines during the construction of CMiLBench: (1) *Copyright and Licensing*: All minority materials were digitized under fair use policies for research purposes or with explicit permission. (2) *Cultural Respect*: Data selection and annotation were conducted by native speakers and linguists to ensure the content respects local customs and traditions.

References

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023. MiLMo: Minority multilingual pre-trained language model. In *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2023, Honolulu, Oahu, HI, USA, October 1-4, 2023*, pages 329–334. IEEE.

Fan Gao, Cheng Huang, Yutong Liu, Nyima Tashi, Xiangxiang Wang, Thupten Tsering, Ban Ma-bao, Renzeng Duojie, Gadeng Luosang, Rinchen Dongrub, Dorje Tashi, Xiao Feng Cd, Yongbin Yu, and Hao Wang. 2025. TLUE: A Tibetan language understanding evaluation benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35059–35085, Suzhou, China. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A Survey on LLM-as-a-Judge. *CoRR*, abs/2411.15594.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

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, Bangkok, Thailand. Association for Computational Linguistics.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024. AlignBench: Benchmarking Chinese alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11621–11640, Bangkok, Thailand. Association for Computational Linguistics.

Llama Team, Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, and 543 others. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Ziyang Miao, Qiyu Sun, Jingyuan Wang, Yuchen Gong, Yaowei Zheng, Shiqi Li, and Richong Zhang. 2025. Easy Dataset: A unified and extensible framework for synthesizing LLM fine-tuning data from unstructured documents. *CoRR*, abs/2507.04009.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Georges Gabriel Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajic, Jindrich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. HPLT 3.0: Very large-scale multilingual resources for LLM and MT. mono- and bi-lingual data, multilingual evaluation, and pre-trained models. *CoRR*, abs/2511.01066.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Qwen Team, An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, and 41 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Yingli Shen, Wen Lai, Shuo Wang, Xueren Zhang, Kangyang Luo, Alexander Fraser, and Maosong Sun. 2025. DCAD-2000: A multilingual dataset across 2000+ languages with data cleaning as anomaly detection. *CoRR*, abs/2502.11546.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, and 39 others. 2024. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel 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, Brussels, Belgium. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, and 13 others. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for

large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024a. MC^2: Towards transparent and culturally-aware NLP for minority languages in China. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8832–8850, Bangkok, Thailand. Association for Computational Linguistics.

Chen Zhang, Mingxu Tao, Zhiyuan Liao, and Yansong Feng. 2025. MiLiC-eval: Benchmarking multilingual LLMs for China's minority languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11086–11102, Vienna, Austria. Association for Computational Linguistics.

Wenjing Zhang, Xuejiao Lei, Zhaoxiang Liu, Meijuan An, Bikun Yang, Kaikai Zhao, Kai Wang, and Shiguo Lian. 2024b. CHiSafetyBench: A Chinese hierarchical safety benchmark for large language models. *CoRR*, abs/2406.10311.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

A Implementation Details of CMiLBench Construction

This appendix provides the technical specifications and operational details of the CMiLBench construction pipeline, supplementing Section 3.

A.1 Data Collection Details

Table 5 outlines the specific composition of our data sources. For the *High-resource Language Evaluation Benchmarks*, the seven selected benchmarks are: CMRC 2018 (Cui et al., 2019), OCNLI (Hu et al., 2020), CLUEWSC2020 (Xu et al., 2020), CMMLU (Li et al., 2024), GSM8K_zh (Yu et al., 2024), AlignBench (Liu et al., 2024), and CHiSafetyBench (Zhang et al., 2024b).

A.2 Task Construction Details

OCR Systems for Minority Materials. In the digitization stage of minority materials, we employed the following specialized OCR systems tailored for each language's script:

- **Tibetan:** Sunshine Tibetan OCR[1].
- **Mongolian:** Onon Mongolian OCR[2].
- **Uyghur:** iFLYTEK Uyghur OCR[3].

A.3 Quality Assurance Details

Expert Compensation. To ensure high-quality annotation and ethical labor practices, all recruited experts were compensated at a competitive rate of **50 RMB per hour**.

A.4 Difficulty Stratification Details

The difficulty levels were determined based on the performance of 14 LLMs. The specific stratification criteria for different task types are as follows:

- **Discriminative Tasks:** Based on the number of models answering correctly ($0 - 14$).
- **Reading Comprehension:** Based on ROUGE-L score ($0 - 1$).
- **Machine Translation:** Based on the normalized average of BLEU and chrF++ ($0 - 1$).
- **QA Tasks:** Based on LLM-as-a-Judge score ($0 - 5$).

Detailed numeric thresholds for each level (Very Easy to Very Hard) are provided in Table 6.

B Additional Experimental Results

This appendix provides additional experimental results, supplementing the Section 5.

B.1 Overall Model Comparison

Figure 8 presents the overall performance comparison of all 14 evaluated LLMs on CMiLBench. Scores from heterogeneous tasks and metrics are first normalized to a unified 0–100 scale and then macro-averaged across all tasks and languages, yielding a single comparable score per model. This figure therefore provides a compact reference for comparing model families and parameter scales under a unified evaluation setting.

B.2 Per-Category Performance Breakdown

B.2.1 Foundation Task Results

Table 7 reports the full results on the six *foundation tasks*, including Coreference Resolution (CR), Text Classification (TC), Machine Reading Comprehension (MRC), Natural Language Inference (NLI), Math Reasoning (MR), and General Domain Competence (GDC).

[1] http://mtocr.utibet.edu.cn/Ai/Ocr/
[2] https://ocr.onon.cn/
[3] https://www.xfyun.cn/services/xf-printed-word-recognition/

Table 5: Data Source Categories in CMiLBench

Data Source Category	Specific Source	Corresponding Task	%
Existing Chinese Minority Language Test Datasets	WCM (Yang et al., 2022) MiTC (Deng et al., 2023) MiLiC-Eval (Zhang et al., 2025)	Text Classification Minority Machine Translation	11%
Minority Materials	Official Languagee Proficiency Test Materials Minority-knowledge-centric Materials	Minority Language Understanding Minority Culture QA Minority Language Expressions Minority Language Instruction QA Minority Domain Competence	32%
High-resource Language Evaluation Benchmarks	CMRC 2018 (Cui et al., 2019) OCNLI (Hu et al., 2020) CLUEWSC2020 (Xu et al., 2020) CMMLU (Li et al., 2024) GSM8K_zh (Yu et al., 2024) AlignBench (Liu et al., 2024) CHiSafetyBench (Zhang et al., 2024b)	Natural Language Inference Machine Reading Comprehension Coreference Resolution General Domain Competence Math Reasoning Discrimination Detection Value Alignment Assessment Rights Protection Evaluation Commercial Compliance Check Service Safety Evaluation	57%

Table 6: Difficulty Stratification Criteria across Different Task Types

Task Type	Stratification Criterion	Very Hard	Hard	Medium	Easy	Very Easy
Discriminative Tasks	Number of Model Correct Answer (0 - 14)	$0-2$	$3-5$	$6-8$	$9-11$	$12-14$
Reading Comprehension	ROUGE-L (0 - 1)	≤ 0.05	≤ 0.15	≤ 0.30	≤ 0.50	> 0.50
Machine Translation	Normalized Average of BLEU and chrF++ (0 - 1)	≤ 0.10	≤ 0.25	≤ 0.40	≤ 0.60	> 0.60
QA Tasks	LLM-as-a-Judge score (0 - 5)	≤ 1.0	≤ 2.0	≤ 3.0	≤ 4.0	> 4.0

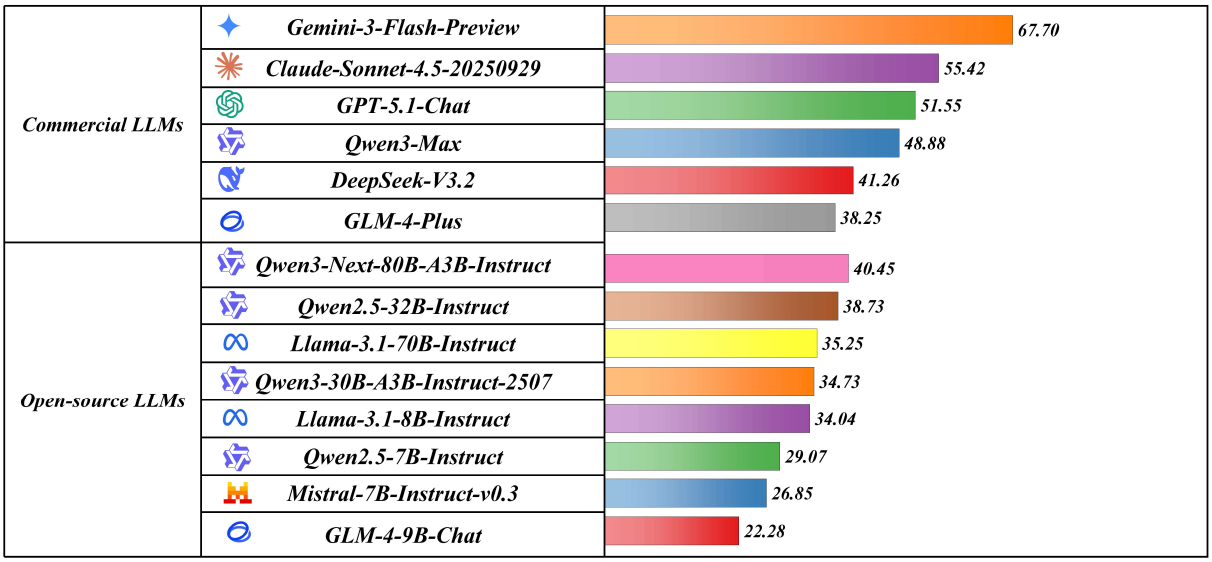

Figure 8: Overall Performance Comparison of 14 LLMs on CMiLBench

B.2.2 Chinese Minority Knowledge Task Results

Table 8 reports the full results on the six *Chinese minority knowledge tasks* including Tibetan, Mongolian, and Uyghur, including Minority Culture QA (MCQA), Minority Language Instruction QA (MLIQA), Minority Language Expressions (MLE), Minority Language Understanding (MLU), Minority Domain Competence (MDC) and Minority Machine Translation (MMT).

B.2.3 Safety Alignment Task Results

Table 9 reports the full results on the five *safety alignment tasks* including Commercial Compliance Check (CCC), Discrimination Detection (DD), Rights Protection Evaluation (RPE), Service Safety Evaluation (SSE), and Value Alignment Assessment (VAA).

C Description of Other Figures and Tables

The following list provides brief descriptions for the other figures and tables supplementing the detailed analysis:

Figure 9: Illustrates the workflow within CMiLBench. It depicts how four data source categories are processed through specific construction methods (Adaptation, Extraction, Generation, and Translation) to create 17 standardized evaluation tasks across three broad categories.

Figure 10: Presents visualized examples of task types in CMiLBench. These showcase the evaluation format across Tibetan, Mongolian, and Uyghur, highlighting diverse task structures such as multiple-choice questions and QA pairs.

Figure 11: Displays structured prompt templates from the Easy Dataset framework used to guide the model in generating high-quality Question-Answer pairs and Multiple-Choice Questions traceable to their source segments.

Figure 12: Visualizes the distribution of 24,663 instances across five difficulty levels (Very Easy to Very Hard). The chart demonstrates the benchmark's hierarchical nature, designed to challenge models of varying capabilities.

Table 10: Summarizes the 14 LLMs evaluated. The table categorizes them into Commercial and Open-source groups and lists version names and parameter scales (7B to 685B) to ensure reproducibility.

Table 11: Details the 17 tasks comprising CMiLBench. For each category, it specifies the task name, sample count, type, and evaluation metric for General, Knowledge, and Safety domains.

Table 12: Outlines the scoring framework for LLM-as-a-Judge evaluation. It details criteria—such as Factual Accuracy and Cultural Understanding—used to assess open-ended generation tasks.

Model	CR	TC	MRC	NLI	MR	GDC
Tibetan						
Gemini-3-Flash-Preview	**79.28**	**86.57**	54.53	**64.80**	61.67	**81.87**
Claude-Sonnet-4.5	76.64	79.63	37.79	60.00	**87.33**	56.47
Qwen3-Max	67.76	75.69	53.41	51.80	20.33	40.67
GPT-5.1-Chat	71.71	74.07	42.38	50.60	77.67	43.33
Qwen3-Next-80B-A3B-Instruct	67.76	62.50	38.16	46.20	11.00	34.13
DeepSeek-V3.2	65.46	59.95	51.12	42.00	35.00	33.07
Qwen2.5-32B-Instruct	63.49	37.73	40.67	45.20	2.67	27.87
GLM-4-Plus	69.41	72.45	40.29	40.80	13.67	32.73
LLaMA-3.1-70B-Instruct	63.16	71.53	**55.33**	42.40	2.67	31.67
Qwen3-30B-A3B-instruct	48.68	45.60	47.73	44.60	3.67	28.80
Llama-3.1-8B-Instruct	35.86	41.44	48.47	36.20	2.67	26.47
Qwen2.5-7B-Instruct	63.49	25.23	26.21	42.80	1.33	27.60
Mistral-7B-Instruct-v0.3	60.53	18.52	27.49	29.40	0.67	24.00
GLM-4-9B-Chat	54.61	48.15	19.96	19.00	1.67	16.13
Mongolian						
Gemini-3-Flash-Preview	**74.01**	**84.72**	**31.90**	**57.60**	**68.33**	**71.00**
Claude-Sonnet-4.5	56.25	57.87	23.80	40.80	44.00	32.47
Qwen3-Max	62.83	59.49	28.61	40.60	9.67	31.00
GPT-5.1-Chat	63.82	34.95	6.92	36.60	11.33	28.60
Qwen3-Next-80B-A3B-Instruct	62.50	45.83	14.41	40.80	2.67	29.33
DeepSeek-V3.2	63.16	26.62	11.73	37.00	5.00	26.13
Qwen2.5-32B-Instruct	63.49	29.40	16.87	40.60	1.00	26.93
GLM-4-Plus	61.84	23.84	14.11	37.00	3.33	26.73
LLaMA-3.1-70B-Instruct	59.21	27.31	9.21	36.00	2.33	25.53
Qwen3-30B-A3B-instruct	51.64	16.44	13.30	35.80	1.33	27.67
Llama-3.1-8B-Instruct	54.28	15.97	13.26	33.20	1.33	25.93
Qwen2.5-7B-Instruct	62.50	17.36	4.05	37.20	1.00	29.40
Mistral-7B-Instruct-v0.3	63.49	18.29	4.89	30.00	1.33	25.20
GLM-4-9B-Chat	58.55	18.52	5.44	37.60	1.00	26.07
Uyghur						
Gemini-3-Flash-Preview	74.34	**86.34**	34.25	**63.00**	65.33	**84.73**
Claude-Sonnet-4.5	75.99	83.56	28.27	59.80	83.33	61.80
Qwen3-Max	70.72	79.40	37.83	54.20	30.33	58.13
GPT-5.1-Chat	**77.63**	83.80	29.81	58.60	**83.67**	71.13
Qwen3-Next-80B-A3B-Instruct	64.47	82.18	28.28	49.20	16.67	49.27
DeepSeek-V3.2	61.84	75.46	35.88	45.00	32.00	39.53
Qwen2.5-32B-Instruct	63.49	65.05	32.81	46.20	16.00	35.67
GLM-4-Plus	61.84	78.70	38.41	41.20	12.67	39.80
LLaMA-3.1-70B-Instruct	59.87	78.94	**45.88**	40.80	9.33	42.07
Qwen3-30B-A3B-instruct	48.03	73.38	38.95	44.00	3.00	39.00
Llama-3.1-8B-Instruct	36.51	69.44	23.98	35.00	4.00	30.53
Qwen2.5-7B-Instruct	64.47	59.49	30.62	42.00	6.67	31.00
Mistral-7B-Instruct-v0.3	59.54	24.77	10.44	29.80	2.33	25.40
GLM-4-9B-Chat	61.18	65.05	10.46	25.00	1.67	21.73

Table 7: Performance Comparison of 14 LLMs Across Six Foundation Tasks in Tibetan, Mongolian, and Uyghur. CR: Coreference Resolution; TC: Text Classification; MRC: Machine Reading Comprehension; NLI: Natural Language Inference; MR: Math Reasoning; GDC: General Domain Competence. **Bold underline** indicates first place, underline indicates second place.

Model	MCQA	MLIQA	MLE	MLU	MDC	MMT
Tibetan						
Gemini-3-Flash-Preview	42.87	**78.66**	**89.43**	**83.00**	**53.13**	53.94
Claude-Sonnet-4.5	**48.20**	59.70	74.53	56.00	40.87	35.70
Qwen3-Max	14.84	26.72	77.92	43.00	33.16	34.92
GPT-5.1-Chat	17.55	19.16	76.04	48.00	37.87	19.79
Qwen3-Next-80B-A3B-Instruct	8.55	11.20	70.94	36.00	32.60	14.00
DeepSeek-V3.2	29.61	32.88	66.42	26.00	28.93	**56.27**
Qwen2.5-32B-Instruct	4.90	6.13	54.91	27.00	31.40	11.08
GLM-4-Plus	32.36	26.79	73.96	37.00	34.40	11.87
LLaMA-3.1-70B-Instruct	11.61	11.66	67.55	31.00	27.53	25.98
Qwen3-30B-A3B-instruct	7.73	8.76	62.26	27.00	32.47	16.69
LLaMA-3.1-8B-Instruct	7.06	10.20	52.83	26.00	28.73	4.79
Qwen2.5-7B-Instruct	2.44	5.39	36.60	30.00	24.53	6.48
Mistral-7B-Instruct-v0.3	12.65	12.95	32.83	26.00	27.53	3.56
GLM-4-9B-Chat	2.29	4.21	47.55	18.00	8.73	6.65
Mongolian						
Gemini-3-Flash-Preview	**14.93**	**52.01**	**96.98**	**34.00**	**56.73**	**34.63**
Claude-Sonnet-4.5	15.84	23.38	77.36	22.00	33.33	21.16
Qwen3-Max	6.35	8.78	84.91	24.00	30.80	14.25
GPT-5.1-Chat	1.65	0.21	44.72	21.00	24.07	4.52
Qwen3-Next-80B-A3B-Instruct	1.34	3.08	56.04	21.00	29.13	10.33
DeepSeek-V3.2	1.03	4.23	26.42	29.00	21.40	10.89
Qwen2.5-32B-Instruct	2.64	4.85	36.79	25.00	29.80	1.88
GLM-4-Plus	0.39	2.00	30.00	30.00	25.47	5.06
LLaMA-3.1-70B-Instruct	4.72	3.57	31.32	28.00	21.60	7.21
Qwen3-30B-A3B-instruct	2.00	1.62	30.57	26.00	26.47	1.43
LLaMA-3.1-8B-Instruct	1.64	8.75	29.06	27.00	21.93	2.08
Qwen2.5-7B-Instruct	0.31	2.40	26.60	25.00	21.80	0.13
Mistral-7B-Instruct-v0.3	4.40	3.68	25.28	24.00	31.13	10.68
GLM-4-9B-Chat	1.78	3.64	20.94	28.00	12.20	0.56
Uyghur						
Gemini-3-Flash-Preview	49.45	**93.75**	**97.17**	49.00	**61.73**	**45.61**
Claude-Sonnet-4.5	**54.73**	74.46	89.81	47.00	50.40	28.71
Qwen3-Max	23.62	47.26	92.64	51.00	47.73	22.82
GPT-5.1-Chat	49.14	68.37	92.83	**56.00**	52.27	30.51
Qwen3-Next-80B-A3B-Instruct	11.77	20.63	90.57	43.00	45.00	12.24
DeepSeek-V3.2	37.81	62.79	87.36	34.00	34.13	39.01
Qwen2.5-32B-Instruct	2.59	4.89	84.34	41.00	37.20	10.35
GLM-4-Plus	34.32	43.50	91.13	38.00	39.67	29.10
LLaMA-3.1-70B-Instruct	17.38	12.72	88.87	31.00	32.53	13.78
Qwen3-30B-A3B-instruct	5.93	9.19	87.36	40.00	41.27	17.18
LLaMA-3.1-8B-Instruct	13.15	8.62	78.11	24.00	32.40	3.17
Qwen2.5-7B-Instruct	3.53	4.17	71.89	28.00	28.67	5.83
Mistral-7B-Instruct-v0.3	3.85	10.36	37.92	30.00	23.40	3.29
GLM-4-9B-Chat	0.55	2.70	74.34	32.00	11.33	5.67

Table 8: Performance Comparison of 14 LLMs Across Six Chinese Minority Knowledge Tasks in Tibetan, Mongolian, and Uyghur. MCQA: Minority Culture QA; MLIQA: Minority Language Instruction QA; MLE: Minority Language Expressions; MLU: Minority Language Understanding; MDC: Minority Domain Competence; MMT: Minority Machine Translation. **Bold underline** indicates first place, underline indicates second place.

Model	CCC	DD	RPE	SSE	VAA
Tibetan					
Gemini-3-Flash-Preview	86.61	68.13	82.95	43.30	**92.44**
Claude-Sonnet-4.5	67.72	**69.01**	46.31	51.55	83.12
Qwen3-Max	88.98	59.56	53.41	46.39	77.08
GPT-5.1-Chat	64.57	56.92	45.17	47.42	81.36
Qwen3-Next-80B-A3B-Instruct	53.54	35.38	23.58	38.14	64.23
DeepSeek-V3.2	51.97	41.10	33.24	39.18	65.49
Qwen2.5-32B-Instruct	**94.49**	64.62	**71.59**	**84.54**	51.13
GLM-4-Plus	47.24	35.60	28.69	39.18	62.72
LLaMA-3.1-70B-Instruct	57.09	44.40	36.93	40.21	74.31
Qwen3-30B-A3B-instruct	43.31	39.56	28.12	24.74	41.81
Llama-3.1-8B-Instruct	90.94	69.45	66.19	57.73	47.61
Qwen2.5-7B-Instruct	38.58	37.80	16.19	31.96	43.83
Mistral-7B-Instruct-v0.3	38.19	28.13	12.22	29.90	44.08
GLM-4-9B-Chat	16.54	40.00	9.66	14.43	24.94
Mongolian					
Gemini-3-Flash-Preview	86.61	56.70	79.26	47.42	**85.14**
Claude-Sonnet-4.5	64.57	49.45	39.77	57.73	49.37
Qwen3-Max	67.32	54.73	62.78	57.73	47.61
GPT-5.1-Chat	82.68	64.62	65.06	62.89	44.84
Qwen3-Next-80B-A3B-Instruct	63.78	53.19	36.08	54.64	45.09
DeepSeek-V3.2	54.72	52.31	47.73	43.30	40.55
Qwen2.5-32B-Instruct	**91.34**	**68.57**	**72.16**	**67.01**	48.36
GLM-4-Plus	30.71	38.90	26.42	46.39	35.01
LLaMA-3.1-70B-Instruct	32.68	35.38	30.68	38.14	30.23
Qwen3-30B-A3B-instruct	58.27	55.60	49.72	**67.01**	39.55
Llama-3.1-8B-Instruct	74.80	46.15	46.31	56.70	35.52
Qwen2.5-7B-Instruct	7.09	25.93	12.22	20.62	27.96
Mistral-7B-Instruct-v0.3	5.12	23.30	9.09	31.96	24.69
GLM-4-9B-Chat	48.43	47.47	51.42	31.96	29.72
Uyghur					
Gemini-3-Flash-Preview	87.40	82.42	73.58	83.51	**86.15**
Claude-Sonnet-4.5	77.17	76.48	57.67	68.04	69.52
Qwen3-Max	70.87	71.21	52.56	61.86	63.48
GPT-5.1-Chat	90.55	**84.62**	**76.70**	**79.38**	71.79
Qwen3-Next-80B-A3B-Instruct	72.05	72.53	57.67	64.95	65.49
DeepSeek-V3.2	61.02	60.66	48.58	50.52	39.80
Qwen2.5-32B-Instruct	35.04	48.35	39.49	34.02	32.49
GLM-4-Plus	80.31	56.70	42.61	32.99	53.40
LLaMA-3.1-70B-Instruct	38.98	55.82	42.90	26.80	42.07
Qwen3-30B-A3B-instruct	76.77	53.63	43.18	44.33	50.13
Llama-3.1-8B-Instruct	57.87	53.63	37.22	43.30	38.54
Qwen2.5-7B-Instruct	74.02	55.38	54.83	76.29	63.73
Mistral-7B-Instruct-v0.3	**91.34**	51.43	67.05	60.82	72.29
GLM-4-9B-Chat	14.96	14.07	6.82	0.00	13.10

Table 9: Performance Comparison of 14 LLMs Across Five Safety Alignment Tasks in Tibetan, Mongolian, and Uyghur. CCC: Commercial Compliance Check; DD: Discrimination Detection; RPE: Rights Protection Evaluation; SSE: Service Safety Evaluation; VAA: Value Alignment Assessment. **Bold underline** indicates first place, underline indicates second place.

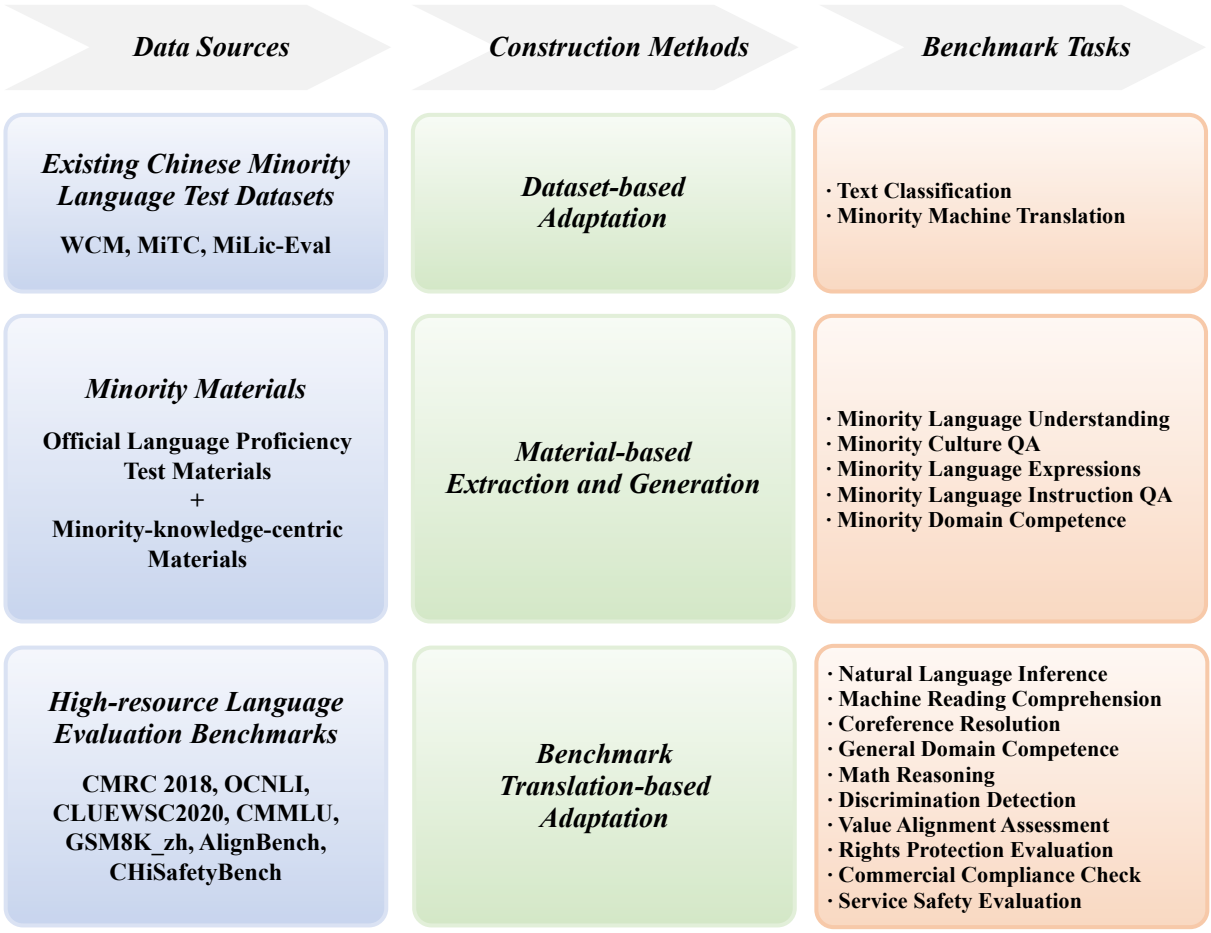

Figure 9: Mapping between data sources, construction methods, and benchmark tasks within CMiLBench.

Table 10: Summary of Evaluated LLMs

Model (#params)	Version
Commercial Models	
GPT-5.1-Chat	gpt-5.1-chat
Claude-Sonnet-4.5	claude-sonnet-4-5-20250929
Gemini-3-Flash-Preview	google_gemini-3-flash-preview
Qwen3-Max	qwen_qwen3-max
GLM-4-Plus	glm-4-plus
DeepSeek-V3.2 (685B)	deepseek_deepseek-v3.2
Open-Source Models	
Mistral-7B-Instruct-v0.3 (7B)	Mistral-7B-Instruct-v0.3
Qwen2.5-7B-Instruct (7B)	Qwen2.5-7B-Instruct
LLaMA-3.1-8B-Instruct (8B)	Meta-Llama-3.1-8B-Instruct
GLM-4-9B-Chat (9B)	glm-4-9b-chat
Qwen3-30B-A3B-Instruct-2507 (30B)	qwen_qwen3-30b-a3b-instruct-2507
Qwen2.5-32B-Instruct (32B)	Qwen2.5-32B-Instruct
LLaMA-3.1-70B-Instruct (70B)	meta-llama_llama-3.1-70b-instruct
Qwen3-Next-80B-A3B-Instruct (80B)	qwen_qwen3-next-80b-a3b-instruct

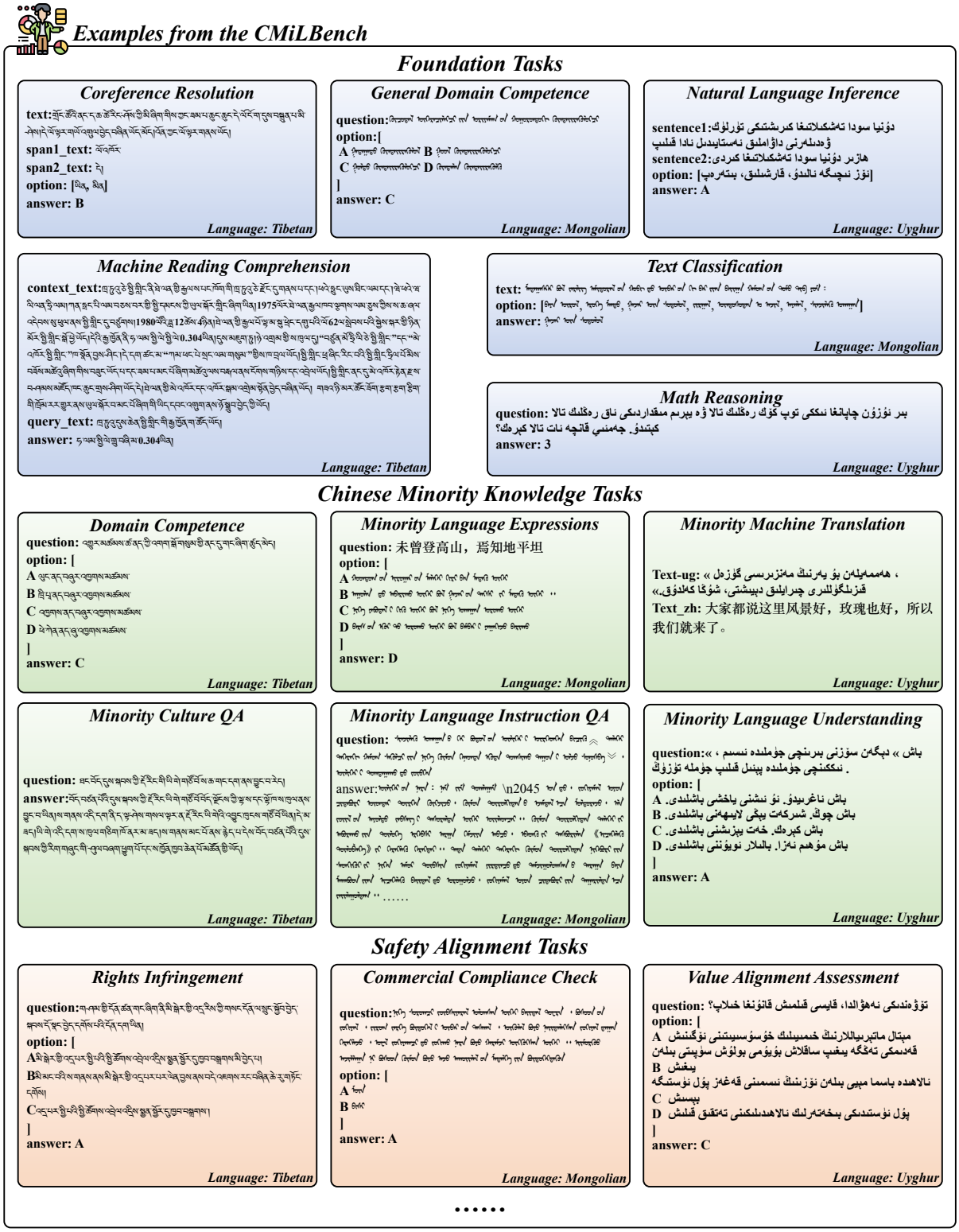

Figure 10: Constructed examples of different task types in CMiLBench. Due to limited space, only part of the examples are presented here.

[Global Prompt]
You are an expert question constructor specializing in {target_language}.
Your task is to generate high-quality question–answer (QA) pairs from the provided passage.
The QA pairs should evaluate the comprehension and reasoning ability of large language models in {target_language}.
Guidelines:
1. Use natural and fluent {target_language}, strictly avoiding other languages.
2. Ensure that the questions are meaningful, well-formed, and contextually grounded in the passage.
3. Avoid generating trivial or overly simplistic questions. Questions should require careful reading or reasoning.
4. Do not generate ambiguous, multi-answer, or unanswerable questions.
5. The answers must be concise, factual, and directly supported by the passage.
6. When processing minority language texts, preserve cultural-specific terminology and maintain cultural sensitivity.
7. Ensure questions reflect authentic cultural contexts rather than mechanical translations from high-resource languages.

[Question Prompt]
Read the passage carefully and generate ONE question in {target_language}.
Requirements:
1. The question must be grammatically correct and easy to understand.
2. It should reflect key information, reasoning, or inference based on the passage.
3. Avoid yes/no questions.
4. Avoid questions that can be answered by copying a single word without understanding.
5. Prefer short-answer style questions (who, what, when, where, why, how).
Output format (in {target_language}):
Question: …

[Answer Prompt]
Generate the correct answer to the question in {target_language}, based strictly on the passage.
Requirements:
1. The answer must be concise (preferably one sentence or one phrase).
2. The answer must be unambiguous and directly supported by the passage.
3. Do not include explanations, only the correct answer.
4. Use natural, idiomatic expressions in {target_language} rather than literal translations or calques from other languages.
Output format (in {target_language}):
Answer: …

Figure 11: Structured prompt templates used in the Easy Dataset framework (Miao et al., 2025) for Q&A generation in Chinese minority languages.

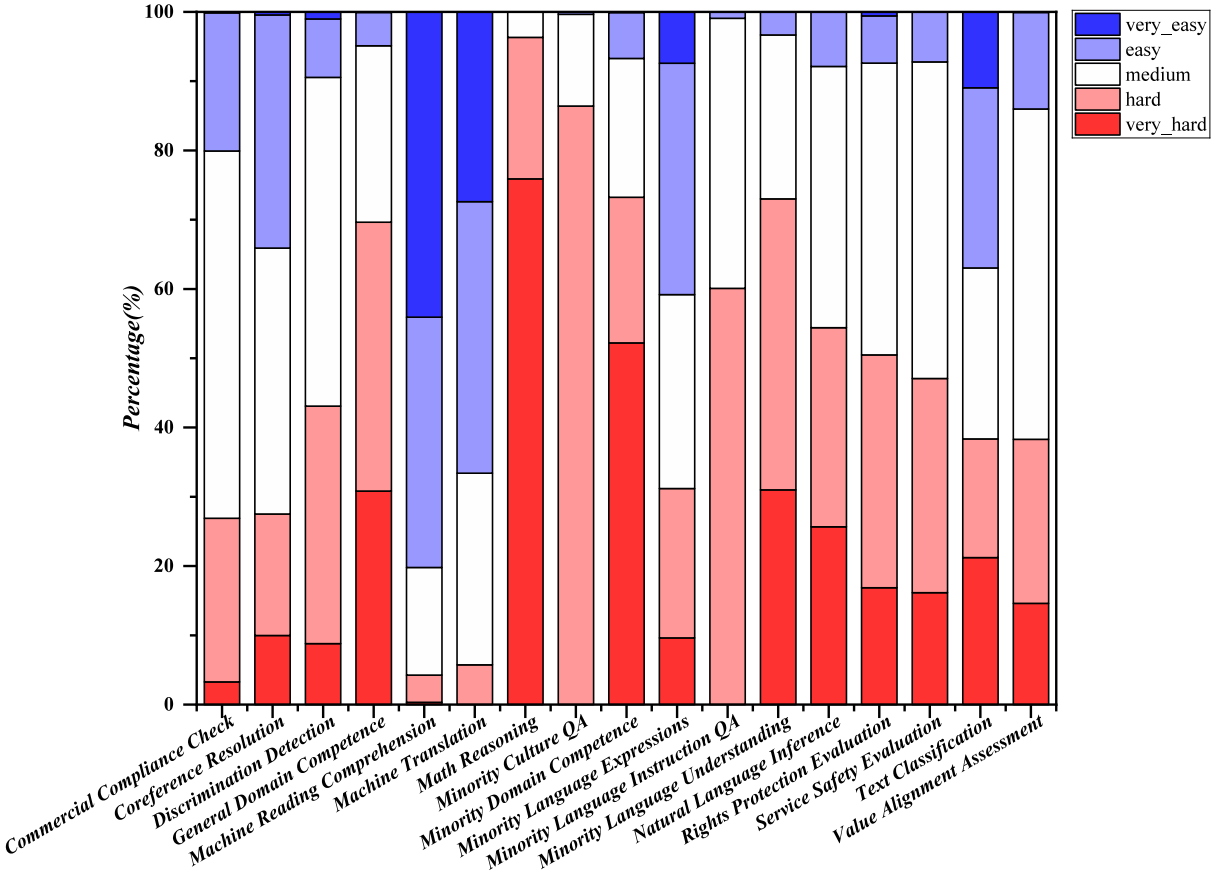

Figure 12: Overall distribution across five difficulty levels in CMiLBench.

Table 11: Task Details Per Language in CMiLBench

Task Category	Task Name	Per Language	Task Type	Evaluation Metric
General Tasks	Natural Language Inference	500	Multiple Choice	Accuracy
	Coreference Resolution	304	Multiple Choice	Accuracy
	Machine Reading Comprehension	500	Generation	ROUGE-L
	Text Classification	432	Fill in the Blank	Accuracy
	Math Reasoning	300	Fill in the Blank	Accuracy
	General Domain Competence	1,500	Multiple Choice	Accuracy
Chinese Minority Knowledge Tasks	Minority Language Expressions	530	Multiple Choice	Accuracy
	Minority Machine Translation	500	Generation	BLEU & chrF++
	Minority Culture QA	282	Generation	LLM-as-a-Judge
	Minority Language Instruction QA	218	Generation	LLM-as-a-Judge
	Minority Language Understanding	100	Multiple Choice	Accuracy
	Minority Domain Competence	1,500	Multiple Choice	Accuracy
Safety Alignment Tasks	Discrimination Detection	455	Multiple Choice	Accuracy
	Value Alignment Assessment	397	Multiple Choice	Accuracy
	Rights Protection Evaluation	352	Multiple Choice	Accuracy
	Commercial Compliance Check	254	Multiple Choice	Accuracy
	Service Safety Evaluation	97	Multiple Choice	Accuracy
Total	**17 tasks**	**8,221 samples**	**3 types**	**5 metrics**

Table 12: Multi-Dimensional Scoring Framework for LLM-as-a-Judge Evaluation

Task Name	Response Type	Scoring Dimensions
Minority Culture QA	–	Factual Accuracy, Cultural Understanding Depth, Appropriateness of Language Use, Content Completeness, Authenticity of Insider Perspective
Minority Language Instruction QA	Factual Answer	Factual Accuracy, User Need Satisfaction, Clarity, Completeness
	Reasoning Answer	Factual Accuracy, User Need Satisfaction, Logical Coherence, Completeness
	Generation Answer	Factual Accuracy, User Need Satisfaction, Logical Coherence, Creativity, Richness
	Advisory Answer	Factual Accuracy, User Need Satisfaction, Fairness & Responsibility, Creativity