CMHG: A Dataset and Benchmark for Headline Generation of Minority Languages in China

Guixian Xu[1,2] † **Zeli Su**[1,2] **Ziyin Zhang**[3] **Jianing Liu**[2]

XU Han[1,2] **Ting Zhang**[1,2] **Yushuang Dong**[1,2]

[1]Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
[2]Minzu University of China [3]Shanghai Jiao Tong University
{rickamorty,guixian_xu,hanxu,jianing_liu,yushuangdong}@muc.edu.cn
daenerystargaryen@sjtu.edu.cn tozhangting@126.com † Corresponding author

Abstract

Minority languages in China, such as Tibetan, Uyghur, and Traditional Mongolian, face significant challenges due to their unique writing systems, which differ from international standards. This discrepancy has led to a severe lack of relevant corpora, particularly for supervised tasks like headline generation. To address this gap, we introduce a novel dataset, Chinese Minority Headline Generation (CMHG), which includes 100,000 entries for Tibetan, and 50,000 entries each for Uyghur and Mongolian, specifically curated for headline generation tasks. Additionally, we propose a high-quality test set annotated by native speakers, designed to serve as a benchmark for future research in this domain. We hope this dataset will become a valuable resource for advancing headline generation in Chinese minority languages and contribute to the development of related benchmarks.

https://huggingface.co/KEVVVV/CMHG

1 Introduction

Recently, the rapid development of large language models (LLMs) has been fueled by the availability of high-quality pre-training data. However, these advancements have primarily benefitted high-resource languages such as English and Chinese. In contrast, many languages with substantial user bases remain excluded due to the scarcity of suitable corpora, especially for specific tasks like *headline generation*. This exclusion poses challenges for both academic research and the practical application of AI technologies.

This paper focuses on underrepresented minority languages in China, including Tibetan, Uyghur, and Mongolian, which have rich linguistic and cultural significance but suffer from a lack of resources. Although these languages appear in multilingual datasets like OSCAR (Jansen et al., 2022) and CulturaX (Nguyen et al., 2024), quality issues limit their usefulness. As shown in Figure 1, there is a clear gap between the large speaker populations of these languages and the small amount of data available in major corpora. Studies also reveal problems with data quality: Zhang et al. (2024) found that 34% of the Uyghur data in CulturaX contains Kazakh or Arabic texts, pointing to issues like language misidentification and noise. These challenges of data scarcity and quality undermine efforts to build effective natural language processing (NLP) systems for these communities.

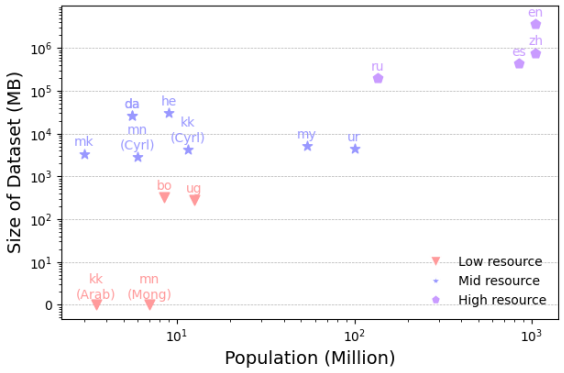

Figure 1: The relationship between population size and dataset size in OSCAR (y-axis, in MB) for various high-, middle-, and low-resource languages.

Moreover, there is a complete lack of open-source datasets tailored for headline generation in these minority languages. This gap hinders the development of supervised methods and benchmarks for headline generation tasks. To address this limitation, we introduce **Chinese Minority Headline Generation (CMHG)**, a novel dataset specifically designed for headline generation in Tibetan, Uyghur, and Mongolian. CMHG consists of **100,000 Tibetan samples** and **50,000 samples each** for Uyghur and Mongolian. In addition to the main dataset, we collaborated with native speakers of these languages to further ensure data quality. From the existing data, we selected **3,000 samples** for each language and conducted a detailed annota-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12351–12358
November 4-9, 2025

tion process to evaluate the alignment and quality of the headlines. These samples were reviewed by multiple annotators for each language, and only the data deemed high-quality by consensus among native annotators was retained. This subset of data provides a reliable *benchmark* for future research, ensuring consistency and reproducibility in evaluating headline generation models. By combining a large-scale dataset with carefully curated benchmark samples, CMHG bridges the gap in resources for headline generation in Chinese minority languages.

In summary, this paper makes the following key contributions:

- We present **CMHG**, a novel and large-scale open-source dataset specifically designed for headline generation in three Chinese minority languages: Tibetan, Uyghur, and Mongolian. We release this dataset.
- We provide a carefully curated benchmark test set, annotated by native speakers, to ensure high-quality evaluation and support transparent, reproducible research in headline generation for Chinese minority languages.

By introducing CMHG, we aim to fill the resource gap and pave the way for advancing natural language processing research on underrepresented languages.

2 Data Sources

The Chinese Minority Headline Generation (CMHG) dataset, proposed in this paper, is sourced from various online platforms in China, including government documents and news articles (detailed list in Appendix A). We used web crawlers to collect the data, where the webpage title serves as the headline and the main text as the source content. To ensure data quality and reliability, we applied a thorough cleaning process, with the main methods outlined as follows:

- **Removal of Non-Textual Content**: We filtered out non-textual elements such as advertisements, pop-ups, navigation bars, and multimedia (e.g., images, videos, and audio files). This ensured that only relevant text was retained.
- **Duplicate Detection and Removal**: We identified and removed duplicate entries to avoid redundancy in the dataset, which could potentially bias the headline generation models.
- **Text Normalization**: We converted all characters to uniform encoding and removing any extraneous white spaces, special characters, or formatting inconsistencies.
- **Language Purity Check**:We sourced content from monolingual websites and used regex-based filtering to ensure linguistic purity in manually annotated data. Controlled noise was allowed in training data to improve model robustness. Additionally, we evaluated several existing Language Identification tools that claimed to support our target languages but found high false-positive rates.

These steps were essential to ensure that the final dataset consisted of high-quality, relevant, and clean text that could be reliably used for our tasks.

3 Annotation

After crawling the data, we further enhanced the quality of the three languages' evaluation set by native speaker annotation. For each language, two annotators assessed the quality of the title-content matching. A total of 3000 randomly selected samples from the crawled data were annotated, with the task focused on evaluating how well the article titles matched the content of the articles.

3.1 Annotation Guidelines

The annotation process was designed to ensure the reliability and consistency of the evaluations. The annotators were provided with the following specific guidelines:

- **Task Objective:** Annotators were asked to assess the degree of relevance between the title and the content of the article and assign a score accordingly.
- **Title Evaluation:** Annotators were first asked to identify any issues between the title and the article content, such as:
  - *Incomplete Article:* whether the article content is incomplete, making it impossible for the title and content to match.
  - *Text Quality:* whether the title contains spelling, grammatical, or contextual errors that would significantly hinder its match with the article content.

Language	Size	Length In Token (Title/Content)	Length In Characters (Title/Content)	Cohen's κ	ICC	Same Tendency
Tibetan	2901	12.3 / 376.7	74.0 / 1884.1	0.71	0.80	1.00
Mongolian	2931	27.2 / 429.8	136.1 / 2149.0	0.28	0.42	0.85
Uyghur	2950	30.2 / 815.7	151.0 / 4078.5	0.44	0.67	1.00

Table 1: Annotation Results and Inter-Annotator Agreement Analysis for Valid Samples in the CMHG Dataset. Token length is counted by the CINO (Yang et al., 2022) tokenizer; character length by raw character count.

  - *Other Issues:* whether there is any other noticeable discrepancies between the title and content, such as irrelevance.

  If no major issues were identified, the title-content match was considered "Normal."

- **Matching Score:** Annotators were instructed to rate the match based on how well the title corresponded with the article content. The scoring system was as follows:
  - **1 point:** Completely Mismatched (The title is entirely unrelated to the content).
  - **2 points:** Slightly Mismatched (The title is related to the content but does not align with the main theme).
  - **3 points:** Slightly Inaccurate (There is some connection, but it is not fully aligned).
  - **4 points:** Uncertain (The relationship between title and content is unclear or ambiguous).
  - **5 points:** Slightly Matched (There is a strong connection, but there are some inconsistencies).
  - **6 points:** Well Matched (The title matches the content with only minor discrepancies).
  - **7 points:** Fully Matched (The title perfectly corresponds to the content).

3.2 Consistency and Quality Control

To ensure consistency and accuracy across annotations, multiple annotators evaluated each article. The following steps were implemented to guarantee the quality of the annotations:

- **Consistency Check:** An annotation was considered invalid if the score differed by more than 2 points from the majority of annotators. Additionally, if an annotator's judgment deviated significantly from the majority opinion (e.g., the majority rated the title as "matching," but the annotator rated it as "not matching"), the annotation would be discarded.

- **Handling Invalid Annotations:** Invalid annotations were removed, and the annotators are incentivized to not produce such annotations.

3.3 Incentive System

To encourage careful and consistent annotation work, we implemented a reward-based incentive system:

- Scores **< 4** or $\geq$ **4** are considered as different tendencies:
  - Scores **< 4** indicate a *non-aligned* tendency.
  - Scores $\geq$ **4** indicate an *aligned* tendency.

- Annotation whose tendency aligns with the majority will receive **0.25 RMB**.

- Annotation that aligns with the majority tendency and furthermore deviates by no more than 1.5 points from the average score will receive an additional **0.25 RMB**.

Annotators were strongly encouraged to adhere to the guidelines to ensure the high quality and consistency of the dataset annotations.

3.4 Annotation Results

In general, the data we retained showed a high degree of quality. After removing the data flagged as mismatched or erroneous by the annotators, we retained the samples with an average score above 4. The final number of valid samples and the average length for each language are shown in Table 1. Most retained samples scored 7, achieving an average score of 6.9/7, which confirms the effectiveness of our native-speaker-guided annotation process.

The average title and content lengths reflect the linguistic characteristics of these minority languages. Tibetan titles and contents have shorter average lengths (12.3 tokens/74.0 characters for titles, 376.7 tokens/1884.1 characters for contents) compared to the significantly longer Mongolian and Uyghur samples.

3.5 Inter-Annotator Agreement Analysis

We assessed inter-annotation agreement using Cohen's κ, ICC, and Same Tendency Rating for each language (Table 1). Despite some variability in specific ratings and a low Cohen's kappa for B0 and MN groups, the Same Tendency Rating was consistently high, indicating agreement on trends and supporting data reliability.

4 Experiment

In this section, we evaluate some of the most popular models available for Tibetan, Mongolian, and Uyghur on CMHG, including finetuning small encoder-decoder models and few-shot evaluation of LLMs.

4.1 Experimental Settings

Fine-tuned Models: The small models, **cino-cum** (*which uses the **cino** (Yang et al., 2022) encoder, based on the XLM-R model tailored for Chinese minority languages, and a transformer decoder in a seq2seq architecture*) and **swcm** (Su et al., 2025) (*which is based on the same structure as cino-cum, but incorporates shared weight optimization across the encoder and decoder for improved performance across languages*), are fine-tuned on non-annotated data from the CMHG dataset. These models are then evaluated using high-quality annotated data to assess their headline generation performance. The fine-tuning is conducted on raw, non-annotated data, while the evaluation is done using a set of annotated samples to measure the ROUGE-L scores.

Few-shot Models: The large models, **Qwen2.5-72B** (Yang et al., 2024) and **LLaMA3.1-70B** (Dubey et al., 2024) use a 2-shot learning paradigm, where two annotated samples are dynamically inserted as examples within the input of each annotated sample.

Detailed training configurations and hyperparameters are provided in Appendix B.

4.2 High-Quality Small Sample Experiment

Given that evaluating large models like **Qwen2.5-72B** and **LLaMA3.1-70b** with nearly 3,000 annotated samples per language is resource-intensive, we also selected a high-quality subset for evaluation to facilitate future works. Specifically, we chose the top 500 annotated samples based on evaluation scores to create a high-quality small sample version, enabling more efficient performance assessment while maintaining data quality.

Model	Size	bo	mn	ug
cino-cum	411M	0.20	0.12	0.09
swcm	457M	0.23	0.18	0.15
Qwen2.5	72B	0.24	0.32	0.29
LLaMA3.1	70B	0.34	0.30	0.35

Table 2: Model Parameters and ROUGE-L F1 Scores across all annotated data

Model	bo	mn	ug
cino-cum	0.21	0.13	0.10
swcm	0.23	0.17	0.14
Qwen2.5	0.24	0.29	0.34
LLaMA3.1	0.34	0.31	0.34

Table 3: ROUGE-L F1 Score in High-Quality data

4.3 Results and Discussion

The experimental results are summarized in Table 2, which presents the performance of the models on the CMHG dataset across the three languages: Tibetan (bo), Uyghur (ug), and Mongolian (mn). The fine-tuning results for the small models, **cino-cum** and **swcm**, show that both models achieved competitive ROUGE-L scores, demonstrating that fine-tuning with the CMHG dataset enables the models to generate concise and contextually accurate headlines for all three languages. This indicates that the large amount of non-annotated data collected in the CMHG dataset plays a crucial role in enhancing model performance for these underrepresented languages.

For the large models, **Qwen2.5-72B** and **LLaMA3.1-70B**, the few-shot results, as shown in Table 2 and Table 3, reveal strong performance across both small and large sample tests. This demonstrates that the models exhibit high-quality headline generation capabilities, regardless of the sample size, highlighting the effectiveness of using small annotated datasets for evaluating model performance. The ability of these models to perform well with just a few annotated samples supports the idea that the CMHG dataset, with its carefully

curated annotated samples, can serve as a reliable benchmark for future research and evaluation in headline generation for minority languages.

Overall, both the fine-tuning and few-shot learning approaches contribute significantly to advancing headline generation for minority languages, and the CMHG dataset proves to be a valuable resource for further research in this area.

Limitations

Despite CMHG's significant contribution to headline generation for Chinese minority languages, some limitations remain. First, while the CMHG dataset represents a substantial effort to address the data scarcity issue for Tibetan, Uyghur, and Mongolian, the availability of high-quality linguistic resources for these languages is still limited compared to high-resource languages. The scarcity of large-scale annotated datasets for other minority languages in China and beyond further highlights the need for continued efforts to expand the scope of language resources. Additionally, the current dataset focuses primarily on headline generation tasks, leaving other NLP applications underexplored. Future work will aim to broaden the dataset to include more minority languages and diverse NLP tasks, alongside collaborations with native speakers and linguistic experts to enhance data quality and coverage, fostering more inclusive and comprehensive NLP research for underrepresented languages.

Ethical Statements

All artifacts in this study are intended for research purposes only, and copyright (where applicable)remains with the original authors or publishers

Acknowledgements

This research was supported by the Joint Research Project of Li'an International Education Innovation Pilot Zone, Hainan Province, China (Grant No: 624LALH006).

References

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. 2022. Perplexed by Quality: A Perplexity-based Method for Adult and Harmful Content Detection in Multilingual Heterogeneous Web Data. *arXiv e-prints*, page arXiv:2212.10440.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Zeli Su, Ziyin Zhang, Guixian Xu, Jianing Liu, XU Han, Ting Zhang, and Yushuang Dong. 2025. Multilingual encoder knows more than you realize: Shared weights pretraining for extremely low-resource languages.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu

Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3937–3949. International Committee on Computational Linguistics.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8832–8850. Association for Computational Linguistics.

A Dataset Details

A.1 1. Dataset Size and Domain Distribution

Table 4 provides the size statistics and domain composition of the CMHG dataset for each language.

B Training Details

Fine-tune Training Details

Hardware: NVIDIA A5000 GPU, 24 GB RAM, Intel i7 CPU.

Software: Ubuntu 20.04, CUDA 11.7, PyTorch 2.3

Training Configurations

Local Batch Size: 20

Gradient Accumulation Steps: 4

Global Batch Size: 80

Epochs: 50

Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$

Learning Rate: 1e-4

Warm-up: Linear warm-up for the first epoch, gradually increasing the learning rate from 1e-5 to 1e-4.

Few-shot Training Details

In the few-shot setting, the model is provided with a prompt and a few examples to generate a task-specific output. For this task, the prompt is designed to help the model generate concise and accurate headline in Tibetan, Uyghur, or Mongolian based on a provided passage and its title. The examples are structured to guide the model's behavior in generating the expected output.

Prompt

Based on the provided passage with title and content, generate a concise and accurate headline in **Tibetan/Uyghur/Mongolian**:

Example 1/2:
Content: {Passage}
Title: {Title of the passage}

Example 2/2:
Content: {Passage}
Title: {Title of the passage}

Task:
Content: {Passage}
Title:

Language	Annotated	Raw	Gov. Docs (%)	News (%)
Tibetan (bo)	2,901	100,000	66	34
Mongolian (mn)	2,931	50,000	100	-
Uyghur (ug)	2,950	50,000	85	15

Table 4: Dataset size and domain distribution for each language in CMHG. The Mongolian news percentage is zero due to limited availability of news media resources in this language.We use the structural consistency of government websites and news sources to minimize noise during the data filtering process.

B.1 2. List of Crawled Websites

Table 5 lists the websites and URLs used for data crawling.

Website Name	URL	Language
Qinghai Lake Website (Tibetan Version)	https://www.amdotibet.cn	BO
China Tibet News Network	https://tb.xzxw.com	BO
Bon Religion Website	http://www.himalayabon.com	BO
Kamba Satellite TV Network	http://tb.kangbatv.com	BO
Qinghai Tibetan Language Radio and TV Station	http://www.qhtb.cn	BO
China Tibetan Calligraphy Website	http://www.zgzzsfw.com	BO
Inner Mongolia Government Website	https://mgl.nmg.gov.cn	MN
Hulunbuir City Government Website	http://mgl.hlbe.gov.cn	MN
Xilingol League Government Website	http://mgl.zlq.gov.cn	MN
Ula'gae Government Website	http://mgl.wlgglq.gov.cn	MN
Chifeng City Government Website	http://mgl.chifeng.gov.cn	MN
Tongliao City Government Website	http://mgl.tongliao.gov.cn	MN
Aksu News Network	https://uy.aksxw.com	UG
Nur Network	https://www.nur.cn	UG
Tianshan Net	http://uy.ts.cn	UG
Xinjiang Government Website	https://uygur.xinjiang.gov.cn	UG
Xinjiang Daily Website	http://xjrbuy.ts.cn	UG

Table 5: List of websites used for data crawling.