



T-LLaMA: a Tibetan large language model based on LLaMA2

Hui Lv^{1,2} · Chi Pu¹ · La Duo³ · Yan Li¹ · Qingguo Zhou¹ · Jun Shen⁴

Received: 30 December 2023 / Accepted: 19 August 2024 / Published online: 19 December 2024
© The Author(s) 2024

Abstract

The advent of ChatGPT and GPT-4 has generated substantial interest in large language model (LLM) research, showcasing remarkable performance in various applications such as conversation systems, machine translation, and research paper summarization. However, their efficacy diminishes when applied to low-resource languages, particularly in academic research contexts like Tibetan. In this study, we trained Tibetan LLaMA (T-LLaMA), a model based on efficient pre-training technology for three downstream tasks: text classification, news text generation and automatic text summarization. To address the lack of corpus, we constructed a Tibetan dataset comprising 2.2 billion characters. Furthermore, we augmented the vocabulary of LLaMA2 from META AI by expanding the Tibetan vocabulary using SentencePiece. Notably, the text classification task attains a state-of-the-art (SOTA) accuracy of 79.8% on a publicly available dataset Tibetan News Classification Corpus. In addition, manual review of 500 generated samples indicates satisfactory results in both news text generation and text summarization tasks. To our knowledge, T-LLaMA stands as the first large-scale language model in Tibetan natural language processing (NLP) with parameters in the billion range. We openly provide our trained models, anticipating that this contribution not only fills gaps in the Tibetan large-scale language model domain but also serves as foundational models for researchers with limited computational resources in the Tibetan NLP community. The T-LLaMA model is available at <https://huggingface.co/Pagewood/T-LLaMA>.

Keywords Large language model · Tibetan · Low-resource languages · Text classification

Introduction

Large language model (LLM) has achieved remarkable success in the field of Natural Language Processing (NLP), particularly since the introduction of ChatGPT [1], which has generated extensive research interest. These models have demonstrated exceptional performance across various tasks,

including text generation, machine translation, sentiment analysis, and question answering.

However, a notable observation is that the majority of large language models have predominantly concentrated on languages with abundant corpora, such as English and Chinese [2, 3]. Tibetan, spoken by over 4.5 million people, lags behind in research on large language models. This challenge stems from the scarcity of Tibetan language corpora and limited engagement from researchers in the Tibetan NLP domain. In previous research, to leverage large language models in scenarios with limited corpora, researchers have demonstrated the effectiveness of fine-tuning pre-trained multilingual models [4]. Fortunately, the open-sourcing of the LLaMA2 by META AI [5] has obviated the need to initiate training from scratch. Additionally, the introduction of Low-Rank Adaptation (LoRA) for LLM, as proposed by Hu et al. [6], allows us to fine-tune large models with 7 billion parameters using significantly less computational power. These technological advancements open avenues for training LLMs for low-resource languages, particularly Tibetan, laying the groundwork for the informatization of the Tibetan language.

✉ Yan Li
liyan_2007@lzu.edu.cn

✉ Qingguo Zhou
zhouqg@lzu.edu.cn

¹ School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China

² School of Electrical Engineering, Northwest Minzu University, Lanzhou 730000, China

³ The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Qinghai Normal University, Xining 810000, China

⁴ School of Computing and Information Technology, University of Wollongong, Wollongong, NSW, Australia

In this paper, we aim to bridge the gap in Tibetan NLP by leveraging large language models. To achieve this, we employed LoRA technology to fine-tune a model named Tibetan LLaMA (T-LLaMA), utilizing the pre-existing 7 billion parameters of LLaMA2 [5]. The training data comprised a Tibetan corpus containing 2.2 billion characters. In order to assess the impact of vocabulary size on model performance, we employed SentencePiece to construct vocabularies of varying sizes [7]. These vocabularies were subsequently integrated into the LLaMA2 lexicon, and their effects on performance were assessed through a text classification task. Moreover, we broadened our investigation by examining the influence of different rank configurations on model performance. Finally, our model evaluation encompassed three fundamental NLP tasks: text classification, news text generation, and automatic text summarization.

For text classification, we utilized the Tibetan News Classification Corpus (TNCC) [8]. T-LLaMA achieved a state-of-the-art accuracy of 79.8%, showcasing exceptional performance. We explored the impact of varying max sequence lengths on text classification accuracy, noting an improvement as sequence length increased. Moving on to news text generation, we transformed proprietary news data into a dataset and manually reviewed 500 generated articles for grammatical correctness and content relevance to headlines. Results showed that the majority of articles were grammatically correct, with 81% exhibiting content relevance. Finally, we fine-tuned the T-LLaMA model for text summarization using LoRA. The evaluation, conducted on 500 randomly chosen test samples, demonstrated the model's effectiveness in generating accurate and concise summaries.

These results collectively underscore the effectiveness of T-LLaMA as a powerful LLM for the Tibetan language. It marks the introduction of the first Tibetan large-scale model with parameters in the billion range. However, challenges persist, including the lack of a substantial prompt engineering corpus for downstream tasks such as dialogue, reasoning, and translation. We acknowledge the need for a more extensive collaborative effort within the research community to address these challenges and propel the development of Tibetan large language models. We encourage and aspire to see more researchers join this community in the future, contributing their expertise and efforts to overcome the hurdles and advance the capabilities of Tibetan large language models.

Related work

Large language models

The recent evolution of NLP has been marked by a transformative paradigm shift, propelled by the advent of LLM.

Notable models such as GPT-3 [9], PaLM [10] and LLaMA [5] exemplify the capabilities of pushing the boundaries of machine understanding and text generation. Large language models have showcased exceptional performance across a spectrum of complex NLP tasks, including text classification [11], question answering (QA) [12, 13], text generation [14, 15] and text summarization [16].

However, the demanding computational resources required for training large models pose a challenge, hindering full-scale model pre-training and fine-tuning. Consequently, various efficient parameter fine-tuning methods have been proposed, including adapter tuning [17], prefix tuning [18], prompt tuning [19], and LoRA [6].

Despite the strides made by LLM in NLP, progress in low-resource languages, such as Tibetan and Southeast Asia languages, has been notably sluggish [20, 21]. Furthermore, the performance of multilingual large models on Tibetan, to the best of our knowledge, remains unsatisfactory. According to our analysis and practical usage of ChatGPT [1], LLaMA [5], and Baichuan2 [22], it is evident that these models exhibit notable performance disparities in Tibetan when compared to high-resource languages like English and Chinese. This highlights an existing gap in the application of these advanced models to languages with limited resources, underlining the importance of our efforts in addressing these challenges within the Tibetan NLP domain.

Tibetan pre-trained language models

Pre-trained language models such as BERT [23], ALBERT [24] and GPT [25, 26] are widely recognized for their effectiveness in enhancing a model's language understanding capabilities. Language-specific pre-trained models have proven to be particularly effective in this regard [27, 28]. Within the Tibetan NLP community, Li undertook the training of a pre-trained model based on ALBERT, finetuning it on self-constructed text classification and summarization datasets, yielding satisfactory results [29]. Sun et al., by constructing a vocabulary based on SentencePiece [7], trained a pre-trained model named TiBERT [30] to excel in Tibetan NLP downstream tasks. Additionally, Yang et al. introduced a Chinese minority pre-trained language model (CINO) based on XLM-R [31], achieving satisfactory results in downstream tasks for minority languages by enriching the vocabulary of the original XLM-R tokenizer.

More recently, a method for Tibetan text classification, utilizing pre-trained models and based on prompt, has been developed [32]. Similarly, Zhou et al. explored three efficient fine-tuning methods based on the TNCC dataset, achieving preliminary success [33]. While these works have made strides in pre-trained models and prompt learning, the exploration of billion-parameter LLM in the Tibetan



context remains largely uncharted. This gap has motivated our efforts in this field.

Our model builds upon LLaMA2, a large language model offering various models ranging from 7 to 65 billion parameters. Through fine-tuning the LLaMA2 model with 7 billion parameters based on our own corpus and leveraging LoRA technology [6], our model has demonstrated satisfactory results.

T-LLaMA

In this study, our approach involves the initial optimization of the Tibetan vocabulary, which is subsequently integrated into the LLaMA2 tokenizer. Following this, we proceed to train a fine-tuned model utilizing LoRA technology. Finally, we assess the performance of the model through evaluations on downstream Tibetan classification tasks, news text generation tasks and text summarization tasks. The code employed in this paper is based on the implementation provided in Ref. [4].

Introduction of LLaMA2

LLaMA, introduced by META AI in 2023 [2], stands out as a large language model featuring four models with varying scales, spanning from 7 to 65 billion parameters. The initial version of LLaMA, constructed using the Transformer deep learning framework [34], adheres to scaling laws [35] by training on a more extensive dataset while employing smaller models. Notably, LLaMA-13B, one of these models, outperformed GPT-3 (175B) [9] in the majority of benchmark tests.

LLaMA exhibits exceptional proficiency in various tasks, including text generation, machine translation, sentiment analysis, and automated question answering, providing robust support for a wide spectrum of NLP applications. In a bid to enhance both the quality and quantity of the training data, the researchers introduced LLaMA2 [5], which effectively doubled the context length of the model. LLaMA2 has demonstrated state-of-the-art (SOTA) performance across multiple tasks, solidifying its position among numerous open-source models.

Low-Rank adaptation

The training of large language models is widely recognized as a resource-intensive process. The endeavor necessitates an extensive corpus, exemplified by the 45 TB data [36] volume utilized in the training of models like ChatGPT. Furthermore, it demands a substantial number of GPUs to provide the necessary computational power, coupled with a considerable amount of time to complete the training process. Given

these challenges, the prospect of training a large language model for Tibetan seems nearly impossible.

In response to this predicament, Edward Hu et al. introduced the Low-Rank Adaptation (LoRA) approach [6]. LoRA posits that over-parametrized models learned during training inherently exist within a lower intrinsic dimension. This innovative approach empowers researchers to train models by optimizing much smaller low-rank matrices, significantly alleviating computational demands and dataset requirements.

The adoption of LoRA technology has made it feasible for us to train a large language model for Tibetan, a low-resource language where such training was previously challenging due to resource constraints. This breakthrough not only broadens the scope of language modeling but also facilitates applications in Tibetan-specific NLP tasks.

Data preparation

We constructed a dataset consisting of approximately 2.2 billion characters, amalgamating content from diverse sources such as websites, various books, journal papers, textbooks, and historical documents. This dataset spans a broad spectrum of content, covering fields such as encyclopedic knowledge, legal documents, history, language, economics, and more. The distribution of content across these categories is outlined in Table 1.

Vocabulary optimization

Tokenization is a critical step in the domain of pre-trained models and large language models. Notably, among various tokenization models, WordPiece [19] and SentencePiece [7] stand out as prominent choices. WordPiece excels in breaking down words into smaller components, facilitating the handling of intricate words and languages. In contrast, SentencePiece offers notable flexibility, accommodating various languages and tokenization granularities, including whole words, subwords, or characters, all of which can be referred to as tokens.

The Tibetan language, following a phonetic writing system, structures its syllables through a combination of

Table 1 Data distribution of the corpus

Document source	Size (GB)
Ancient literature	1.3
Website news	2.8
Blog and Weibo	0.26
Books	1.77
Journal Article	0.12
Teaching material	0.002
Total	6.27

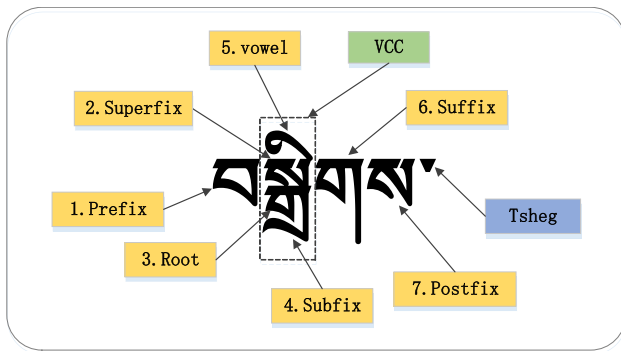


Fig. 1 Tibetan syllable structure

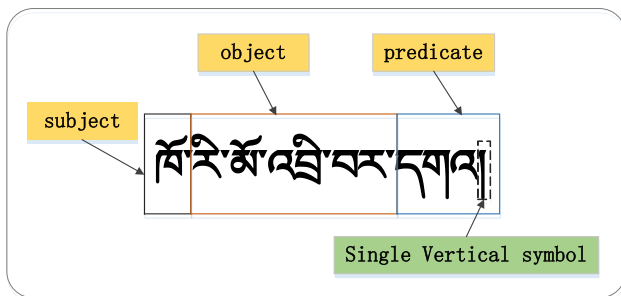


Fig. 2 Tibetan Sentence Structure

vertical and horizontal, separated by the separator . (Tsheg). The combination of upper and lower characters is also called Vertical Combination Character (VCC). Each syllable is composed of one or more characters [37], as illustrated in Fig. 1 for the syllable “བཞུགས་” which consists of components like Prefix “བ”, Superfix “མ”, Root “ཁ”, Subfix “ེ”, Vowel “འ”, Suffix “ག” and Postfix Suffix “ས” named according to their positions. These components collectively form Tibetan characters. In contrast to the subject-verb-object structure found in Chinese and English sentences, Tibetan sentence structure follows a pattern where the subject precedes the object, with the predicate coming after. Figure 2 illustrates the fundamental structure of a Tibetan sentence, exemplified by “ཁོ་མོ་འདྲི་བ་དགའ་” (He likes to draw). It mentions the components of subject, object, and predicate. Each sentence concludes with “།” (shed) as a standard grammatical requirement.

In the context of Tibetan script intricacies, SentencePiece emerges as a superior choice due to its adaptability in handling variations in multicomponent characters and adjustments in word segmentation granularity. Its ability to flexibly adapt to the unique requirements of Tibetan script positions SentencePiece as a crucial tool in achieving effective tokenization.

In LLaMA2, a significant part of the training data is in English, with a vocabulary consisting of only 315 Tibetan

tokens. This has resulted in significant disparities in LLaMA2's comprehension and generation capabilities in Tibetan compared to English and Chinese. To address this issue, we utilized SentencePiece to generate vocabularies of varying sizes on our corpus. Subsequently, we fine-tuned the model separately on each of the expanded vocabularies derived from the original vocabulary of LLaMA2 to choose a better model.

However, during the direct generation of the vocabulary with SentencePiece, two issues surfaced. Firstly, some words ended with “།”, which serves as a punctuation symbol in Tibetan, making it inappropriate. Secondly, certain character components were segmented as individual tokens in the vocabulary, increasing coverage but posing challenges for word segmentation. To address these issues, we removed clearly inappropriate cases during the vocabulary generation process.

Experiments and Results

Fine-tune of LLaMA2

LoRA has proven to be a highly efficient tool for fine-tuning pre-trained models. Different ranks can influence the model's performance post fine-tuning. To identify the optimal model, we conducted separate fine-tuning experiments with ranks $r=1$, $r=8$, and $r=16$. These fine-tuning tasks were executed on 8 NVIDIA A100 GPUs, with each task completing within one day. Figure 3 illustrates the loss curves of the three models with different ranks. The experimental results indicate that as the rank increases, the rate of loss reduction exhibits a slight increment. To ascertain the influence of different ranks on model performance, we conducted tests on the text classification task with various rank configurations. Our results suggest that increasing the rank does not significantly enhance the model's performance, as shown in Table 2. From this series of experiments, we obtained the highest accuracy of 77.17% at a rank of $r=8$, which is highlighted in Table 2 (in bold). This model was then selected for further validation.

Furthermore, in large language models, the size of the vocabulary is another crucial factor that significantly impacts the model's performance. To assess the influence of vocabulary size on model performance, we utilized SentencePiece to generate separate vocabularies containing 2000, 4000, 6000, 8000, 10,000, and 12,000 tokens, respectively. These vocabularies were then merged with the original vocabulary of the LLaMA2 model and tested in the text classification task. The results showcase the effectiveness of the additional Tibetan vocabulary, and we achieved the best performance with an additional vocabulary size of 8000, as highlighted in Table 3 (in bold).



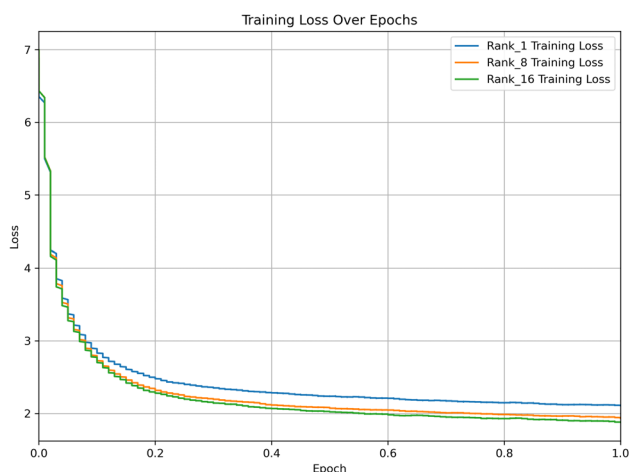


Fig. 3 The training loss during the fine-tune task of different rank

Table 2 TNCC text classification results with different rank configuration

rank	Accuracy (%)	Macro-F1 (%)
1	76.52	74.94
8	77.17	75.9
16	75.98	74.05

Table 3 TNCC text classification results with varied extra Tibetan token counts

Extra Tibetan token counts	Accuracy (%)	Macro-F1 (%)
0	68.3	68.3
2000	78.61	78.64
4000	77.96	77.91
6000	78.94	79.11
8000	79.8	79.77
10,000	76.98	77.05
12,000	78.72	78.73

Our experiments unveiled a lack of linear improvement in performance with the increasing vocabulary size, ranging from 2000 to 12,000. Instead, we observed fluctuations in the model's performance. Notably, even after fine-tuning the text classification task on the original LLaMA2, the accuracy failed to meet satisfactory levels. However, following the enlargement of the vocabulary, significant performance enhancement was evident, even with a vocabulary size of 2000. And the vocabulary expanded by 8000 Tibetan tokens attained the highest performance in the fine-tuned text classification task. Despite

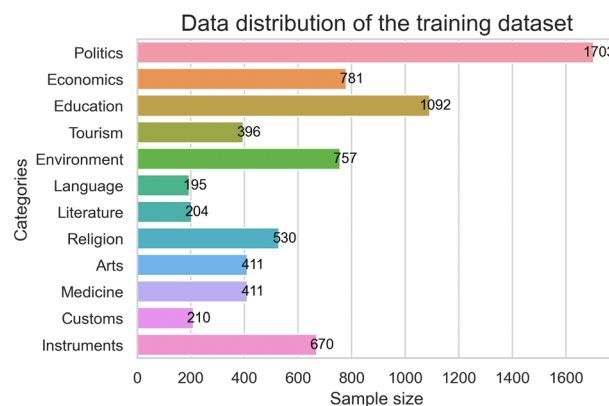


Fig. 4 The data distribution of the Tibetan classification dataset

the limitations inherent in using performance from a single task as a basis, we opted for the vocabulary size of 8000 for subsequent experiments, given the results of the text classification task.

Text classification

Text classification, a fundamental task in NLP, finds applications in sentiment analysis, spam detection, topic categorization, and more. To evaluate the performance of our trained T-LLaMA model, we selected text classification as the natural language understanding task. The chosen dataset is TNCC, a publicly available dataset as referenced in [38], comprising 12 classes, with data distribution depicted in Fig. 4.

The fine-tuning for this task was accomplished using LoRA within 2 h on our 8 NVIDIA A100 GPUs. The classification model, based on LLaMA2, prepared for fine-tuning in downstream tasks by adding a sequence classification head on top of LLaMA2. The code implementation leveraged the AutoModelForSequenceClassification module provided by Hugging Face.

Our training parameter settings are outlined in Table 4. We conducted three sets of experiments on different ranks of T-LLaMA, with each set containing three experiments with varying max sequence lengths.

Our evaluation of text classification results on the TNCC dataset, when compared to publicly available benchmarks, highlights the superior performance of our model in this task (Table 5). The experimental results indicate that an increase in the max sequence length generally leads to improved performance, while an increase in rank does not consistently result in performance enhancement. TLLaMA exhibited optimal performance with a configuration of $r=8$ and $msl=1536$, achieving an impressive accuracy rate of 79.8%, as highlighted in Table 5 (in bold). Furthermore, a comprehensive analysis of the confusion matrix is visualized

Table 4 Train parameters of text classification tasks

Parameters	Value
LoRA rank of T-LLaMA	8
Max sequence length	512, 1024, 1536
Rank	2
lora_alpha	32
Num of epochs	20
Learning rate	1e-4
Batch size per device	1

Table 5 Text classification results of TNCC, r means the rank of T-LLaMA and msl means the max sequence length of input text

Models	Accuracy (%)	Macro-F1 (%)
Transformers	28.63	28.79
CNN(syllable)	61.51	57.34
TextCNN	61.71	61.53
DPCNN	62.91	61.17
TextRCNN	63.67	62.81
BERT-base-Tibetan	–	51
CINO-base	73.1	70
CINO-large	76.3	73.7
TiKEM	74.46	72.61
T-LLaMA $r=8$, $msl=512$	77.17	75.90
T-LLaMA $r=8$, $msl=1024$	79.02	77.35
T-LLaMA $r=8$, $msl=1536$	79.8	79.77

in Fig. 5. The results underscore the efficacy of T-LLaMA in outperforming existing models in text classification tasks, particularly in low-resource language scenarios like Tibetan.

From the confusion matrix, we observe notable precision and recall values for specific classes. For instance, Class Politics (0) demonstrates a precision of 82.27% and a recall of 80.29%. Despite having the largest sample size in the dataset, Politics (0) does not exhibit the lowest performance among all categories. Conversely, classes Economics (1) and Tourism show precision of 67.96% and 66.67%, respectively, making them influential categories in evaluating model performance. This is attributed to sample imbalance, as evident from the data distribution in Fig. 4. Misclassifications for class Economics (1) are primarily directed towards Politics (0), potentially stemming from the close ties between politics and the economy, as political news often includes elements of economic data.

According to these experiments, T-LLaMA exhibits strong performance in low-resource Tibetan text classification tasks, particularly on the TNCC dataset. The model's effectiveness, especially with optimized settings, highlights its potential for addressing challenges in

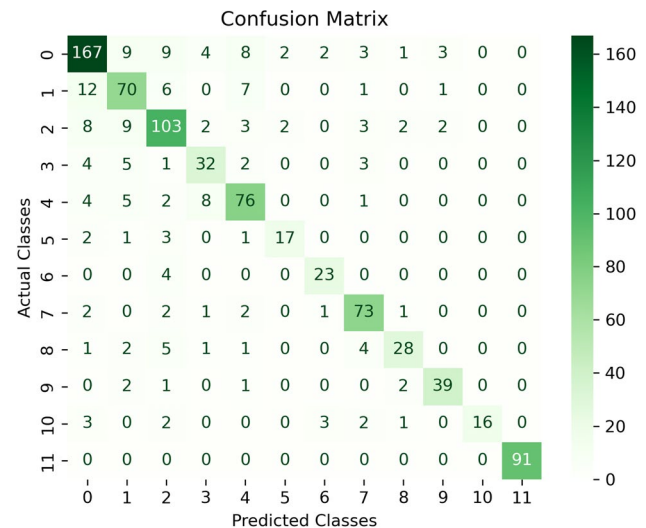


Fig. 5 The confusion matrix of the text classification task of the model with $r=8$ and $msl=1536$. The numerical labels ranging from 0 to 11 correspond to specific categories: Politics, Economics, Education, Tourism, Environment, Language, Literature, Religion, Arts, Medicine, Customs, and Instruments

resource-constrained language scenarios. The successful use of LoRA technology underscores its practical utility. Opportunities for improvement include addressing data imbalances and further exploration of fine-tuning strategies. Looking forward, T-LLaMA shows promise for applications in low-resource language contexts, with ongoing opportunities for model refinement and enhancement.

News text generation

The ability of LLM to generate text on specific topics is a crucial aspect for various applications such as content creation and automated writing. In this study, we sought to further validate the effectiveness of T-LLaMA by conducting experiments on news text generation based on news titles. We established a news text generation corpus using our extensive proprietary news dataset, consisting of a billion Tibetan characters. Filtering out shorter news articles (less than 1024 characters), we retained 80% of the dataset, utilizing the titles as inputs and the corresponding contents as outputs.

In our experiment, we explored two types of prompts inspired by the structure of Stanford Alpaca for prompt tuning using LoRA. The first prompt featured an instruction section describing the task, with the title and content as input and output, respectively (Fig. 6a). The second prompt took a different approach, utilizing the title as a direct instruction and the content as the output, with null input (Fig. 6b). And both prompts underwent prompt tuning with LoRA, proving effective in this task.



Fig. 6 The two different prompts we used in our experiments

Prompt1:{

Instruction:གཤམ་གྱི་གསར་འགྱུར་ཁ་བྱང་ལ་གཞིགས་ནས་གསར་འགྱུར་གྱི་ནང་དོན་བྱུང་མོགས་།,
Input: {The title of the news},
Output: {The content of the news}

(a)

Prompt2: {

Instruction: {The title of the news},
Input: "
Output: {The content of the news}

(b)

Considering that the primary beneficiaries of our developed Tibetan language model are users with constrained computational resources, we recognized the necessity of model quantization during the validation phase. Quantization refers to the process of reducing the bit precision of model parameters to minimize storage requirements and computational overhead. By quantizing the model to lower bit precision, we can enhance inference efficiency on resource-constrained systems. This becomes particularly crucial in scenarios where resources are limited, such as when conducting inference on low-performance GPUs or CPUs. Thus the model's performance was evaluated after quantization using llama.cpp [39]. We randomly selected

500 news headlines from the test dataset and generated corresponding news content for manual evaluation using the second prompt.

The manual evaluation revealed that the model exhibited minimal grammatical errors and demonstrated strong alignment between the generated content and the headlines. Table 6 showcases two sets of samples with closely matched generated content and headlines. Conversely, Table 7 presents two sets with poorer alignment. Out of the 500 samples, 51 were generated as empty lines. Almost all remaining data exhibited correct grammar, with only one exception. Despite the labor-intensive nature of manual evaluation, it remains the most accurate assessment method. In terms of content relevance, 81% of the samples closely matched the

Table 6 The generated content exhibits a strong alignment with the headline

Title	Generated Content
རྒྱལ་སའི་ཁྱེད་བཅོས་དང་སྒྲངས་པས་སི་ཁྲོན་བོད་ཁུལ་དུ་ཁྱེད་ལྷན་བཅོས་བྱེད་སྒྲི་ཐེལ་རྒྱ་ཡིན་པ། (Medical volunteers from the capital will go to Tibetan areas in Sichuan Province to carry out public welfare medical activities.)	3ཚེས28ཉིན། རྒྱལ་སའི་གསེང་གི་སྒྲི་ཐེལ་མ་ཞིག་གིས་སི་ཁྲོན་ཁིང་ཆེན་དཀར་མངོན་ཁུལ་མི་དམངས་ཁྱེད་ཁང་ནས་ཁྱེད་བཅོས་ལན་ལག་ཆོག་རྒྱུ་གི་ཁོངས་གནས་ཁྱེད་ཁང་དང་པོར་ཁྱེད་ལ་རྒྱུ་གི་འོས་འགན་ཁྱེད་བཅོས་བྱེད་སྒྲི་ཐེལ་ཞིང་། ཐོངས་འདྲིའི་བྱེད་སྒྲིའི་བརྒྱུད་དོན་གཤམ་པོ་ནི་ཁྱེད་པས་ནད་ཐོག་ལག་ལེན་དང་། བདེ་ཐང་འགན་སྒྲུབ་དང་། མིན་ཐང་འགན་ལེན་བཅས་ཁྱེད་ནས་སི་ཁྲོན་བོད་ཁུལ་དུ་ཁྱེད་ལྷན་པས་དོན་ལ་ཁྱེད་ཀྱིས་ཀྱིས་ཡོད་སྒྲི་ཐེལ་འདྲིའི་གཤམ་པོ་རྒྱུ་དེ་ཡིན་པ་རེད། (On March 28, a student of Capital Medical University carried out a free medical treatment activity for small doctors in the first hospital of Ganzi Prefecture People's Hospital of Sichuan Province, to promote the public welfare of Tibetan areas in Sichuan Province.)
རྒྱལ་ཁྱིམ་གྱི་ཚྷོགས་ཀྱིས་དབྱིན་ཇིའི་མེ་རྒྱུན་ཇི་སི་ཇི་ཤོང་ཁྱེད་དུ་བྱུང་བའི་འབར་གས་འཆབ་ཞོལ་དོན་རྒྱུན་ལ་ཁ་རྒྱུང་དག་པོ་བཏང་པ། (The international community has strongly condemned the bombing attack in Manchester, England.)	ཞིན་གྲ་གསར་འཕྲུལ་ཁང་གི་སྒྲིག་འཛིན་སྐད་དབྱིན་ཇི་ཏ་འཛམ་ས་འདི་མཇང་བཞིན་པའི་རྒྱལ་ཁྱིམ་གྱི་བརྒྱུན་པའི་སྐད་ཆོར་ཆོས་ཉི་ཤུ་འི་ཉིན་དབྱིན་ཇིའི་ཁུལ་སི་ཇི་ཤོང་ཁྱེད་ཀྱིས་ནས་གནས་སྡོད་ཐེལ་དོན། རྒྱལ་ཁྱིམ་གྱི་ཚྷོགས་ཀྱིས་དབྱིན་ཇིའི་མེ་རྒྱུན་ཇི་སི་ཇི་ཤོང་ཁྱེད་ཀྱིས་དང་ལས་དུ་བྱུང་བའི་འབར་གས་འཆབ་ཞོལ་དོན་རྒྱུན་ལ་ཁ་རྒྱུང་དག་པོ་བཏང་བ་དང་། དབྱིན་ཇི་ཚྷོགས་ཀྱི་མི་ལྷན་ཤི་ཁྱས་བཅོས་པའི་ལྷང་རྒྱུས་པ་རེད། (According to Xinhua News Agency, the international channel, which is visiting the UK, reported from the UK Callister city on the 20th that the international community strongly condemned the explosion attack in the streets of Manchester, England, and expressed condolences for the casualties caused by British personnel.)

Table 10 A case of our model performing text summarization[illegible]

To validate the effectiveness of our model, we employed the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics for the evaluation of text summarization task [40]. ROUGE constitutes a set of automatic evaluation metrics designed to gauge the quality of text summaries, with a primary emphasis on word overlap between the generated summary and the reference summary. The various sub-metrics within ROUGE include ROUGE-1 (single-word overlap), ROUGE-2 (bigram overlap), and ROUGE-L (length of the longest common subsequence). The results of the evaluation on a subset of 500 samples from the test set are presented in Table 9. Table 10 illustrates a case of our model performing text summarization. The preliminary results of our experiments indicate that our model has achieved satisfactory performance in this task.

Conclusion

In this study, we presented T-LLaMA, a large language model fine-tuned for Tibetan language processing, with a focus on text classification, news text generation and text summarization. The experiments conducted on the TNCC showcased TLLaMA’s robust performance in text classification, surpassing existing models with an accuracy of up to 79.8%. Moreover, T-LLaMA demonstrated its proficiency in news text generation, leveraging a proprietary dataset and two distinct prompts for prompt tuning with LoRA. Manual evaluations of the generated content revealed minimal grammatical errors and a strong alignment with news headlines. And we have also achieved satisfactory results in the text summarization task.

The incorporation of LoRA technology proved instrumental in fine-tuning T-LLaMA, enabling effective training

in resource-constrained environments. The model’s exceptional performance highlights the significant potential of large language models in advancing natural language processing tasks for minority languages. Although we have achieved good results in several downstream tasks using large language models, there is currently a lack of specific evaluation methods for Tibetan natural language processing in comparison to the established evaluation approaches for mainstream large language models.

Looking ahead, our focus is on constructing datasets that cover a wider array of downstream tasks, allowing the model to demonstrate its effectiveness across various Tibetan natural language processing challenges. Additionally, there are promising directions for future research, including delving into hyperparameter tuning, enhancing model interpretability, and exploring applications in additional low-resource language tasks. T-LLaMA’s notable achievements in text classification, news text generation, and text summarization position it as a valuable asset for the advancement of Tibetan NLP. Furthermore, its contributions extend to the broader landscape of other low-resource languages, signifying its importance in driving progress in these linguistic domains. We anticipate that the release of the T-LLaMA model will significantly contribute to the advancement of the Tibetan NLP community.

Acknowledgements This work was partially supported by National Natural Science Foundation of China under Grant No. U22A20261 and 61402210, National Key R & D Program of China under Grant No. 2020YFC0832500, National Natural Science Foundation of China under Grant No. 62266037, the Fundamental Research Funds for the Central Universities under Grant No. lzujbky2022-kb12Science and Technology Plan of Qinghai Province under Grant No.2020-GX-164, and Natural Science Foundation of Gansu Province under Grant No. 22JR5RA186.

Data availability The classification data used in this study are publicly available from the Tibetan News Classification Corpus (TNCC) at <https://github.com/FudanNLP/Tibetan-Classification>.

Declarations

Conflict of interest Corresponding authors declare on behalf of all authors that there is no conflict of interest. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- OpenAI. (2022) ChatGPT [GPT-3.5]. <https://openai.com/chatgpt>
- Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Laroix T, Lample G (2023) Llama: Open and efficient foundation language models. arXiv preprint. arXiv: 2302.13971
- Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, Zhu T (2023) Qwen technical report. arXiv preprint. arXiv: 2309.16609
- Cui Y, Yang Z, Yao X (2023) Efficient and effective text encoding for chinese llama and alpaca. arXiv preprint. arXiv: 2304.08177
- Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Scialom T (2023) Llama 2: Open foundation and fine-tuned chat models. arXiv preprint. arXiv: 2307.09288
- Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Chen W (2021) Lora: Low-rank adaptation of large language models. arXiv preprint. arXiv: 2106.09685
- Kudo T, Richardson J (2018) Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint. arXiv: 1808.06226
- Qun N, Li X, Qiu X, Huang X (2017) End-to-end neural text classification for tibetan. In: Chinese computational linguistics and natural language processing based on naturally annotated big data, pp 472–480
- Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Amodei D (2020) Language models are few-shot learners. In: Advances in neural information processing systems, pp 1877–1901
- Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Fiedel N (2023) Palm: scaling language modeling with pathways. J Mach Learn Res 24(240):1–113
- Sun X, Li X, Li J, Wu F, Guo S, Zhang T, Wang G (2023) Text classification via large language models. arXiv preprint. arXiv: 2305.08377
- Kasai J, Sakaguchi K, Le Bras R, Asai A, Yu X, Radev D, Inui K (2024) RealTime QA: what's the answer right now?. In: Advances in neural information processing systems, vol 36
- Kamalloo E, Dziri N, Clarke CL, Rafiei D (2023) Evaluating open-domain question answering in the era of large language models. arXiv preprint. arXiv: 2305.06984
- Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Wen JR (2023) A survey of large language models. arXiv preprint. arXiv: 2303.18223
- Xiao L, Shan X, Chen X (2023) PatternGPT: a pattern-driven framework for large language model text generation. In: Proceedings of the 2023 12th International Conference on Computing and Pattern Recognition, pp 72–78
- Xie Q, Bishop JA, Tiwari P, Ananiadou S (2022) Pre-trained language models with domain knowledge for biomedical extractive summarization. Knowl-Based Syst 252:109460
- Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, Gelly S (2019) Parameter-efficient transfer learning for NLP. In: International Conference on machine learning, pp 2790–2799
- Li XL, Liang P (2021) Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint. arXiv: 2101.00190
- Lester B, Al-Rfou R, Constant N (2021) The power of scale for parameter-efficient prompt tuning. arXiv preprint. arXiv: 2104.08691
- Hedderich MA, Lange L, Adel H, Strötgen J, Klakow D (2020) A survey on recent approaches for natural language processing in low-resource scenarios. arXiv preprint. arXiv: 2010.12309
- Nguyen XP, Zhang W, Li X, Aljunied M, Tan Q, Cheng L, Bing L (2023) SeaLLMs—large language models for Southeast Asia. arXiv preprint. arXiv: 2312.00738
- Yang A, Xiao B, Wang B, Zhang B, Bian C, Yin C, Wu Z (2023) Baichuan 2: open large-scale language models. arXiv preprint. arXiv: 2309.10305
- Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv: 1810.04805
- Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: a lite bert for self-supervised learning of language representations. arXiv preprint arXiv: 1909.11942
- Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training
- Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
- Jiang S, Fu S, Lin N, Fu Y (2021) Pretrained models and evaluation data for the Khmer language. Tsinghua Sci Technol 27(4):709–718
- Alghamdi A, Duan X, Jiang W, Wang Z, Wu Y, Xia Q, Ghaddar A (2023) Aramus: pushing the limits of data and model scale for Arabic natural language processing. arXiv preprint. arXiv: 2306.06800.
- Li L (2020) Tibetan pre-trained model based on ALBERT and its application. Lanzhou University
- Liu S, Deng J, Sun Y, Zhao X (2022) Tibert: Tibetan pre-trained language model. In: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp 2956–2961
- Yang Z, Xu Z, Cui Y, Wang B, Lin M, Wu D, Chen Z (2022) CINO: A Chinese minority pre-trained language model. arXiv preprint. arXiv: 2202.13558.
- An B (2023) Prompt-based for low-resource Tibetan text classification. ACM Trans Asian Low-Resour Lang Inform Process 22(8):1–13
- Zhou M, Daiqing Z, Qun N, Nyima T (2023) PEFTT: Parameter-Efficient Fine-Tuning for low-resource Tibetan pre-trained language models. arXiv preprint. arXiv: 2309.12109.

34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I (2017) Attention is all you need. *Advances in neural information processing systems*
35. Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, Sifre L (2022) Training compute-optimal large language models. *arXiv preprint. arXiv: 2203.15556*.
36. Cooper K (2023) OpenAI GPT-3: Everything you need to know. [Online]. Available: <https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/>. Accessed 17 May 2023
37. Yi Z, Wu Q, Yu J, Tang Y, Liu X, Peng L, Ma J (2022) Tibetan Syllable Prediction with Pre-trained Cross-lingual Language Model. In: 2022 IEEE 5th International Conference on computer and communication engineering technology (CCET), pp 162–166
38. Qun N, Li X, Qiu X, Huang X (2017) End-to-end neural text classification for Tibetan. In: *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 16th China National Conference, CCL 2017, and 5th International Symposium*, pp 472–480.
39. Gerganov, Georgi (2022) llama.cpp. GitHub. <https://github.com/ggerganov/llama.cpp>.
40. Lin CY, Hovy E (2003) Automatic evaluation of summaries using n-gram co-occurrence statistics. In: *Proceedings of the 2003 Human Language Technology Conference of the North American chapter of the association for computational linguistics*, pp 150–157.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.