TiLamb：基于增量预训练的藏文大语言模型

庄文浩[1,2]　孙媛[1,2,3,*]　赵小兵[1,2,3]

[1]中央民族大学　信息工程学院，北京　100081

[2]国家语言资源监测与研究少数民族语言中心

[3]民族语言智能分析与安全治理教育部重点实验室

*通讯作者：孙媛

tracy.yuan.sun@gmail.com

摘要

基于“预训练+微调”范式的语言模型展现了卓越的性能，随着模型规模和训练数据量的扩增，其解决多种自然语言处理任务的能力得到了显著的提高。当前的大语言模型主要支持英汉等主流语言，这限制了藏语等低资源语言在该领域的研究。针对藏语数据稀缺、现有藏语预训练模型效果不够好、下游任务可扩展性差等问题，本文汇总清洗得到26.43GB藏文数据，以开源的LLaMA2-7B作为基座模型，扩充LLaMA2现有词表，增加了约30,000个藏文tokens，提高其藏文编码效率和对藏文的语义理解能力，通过增量预训练得到藏文大语言模型基座TiLamb。根据多种藏文下游任务分别制作数千到几万条不等的微调数据集，微调后的TiLamb在藏文新闻分类、藏文实体关系分类、藏文机器阅读理解、藏文分词、藏文摘要、藏文问题回答、藏文问题生成共七个下游任务中进行验证，多项指标结果相较传统方法和其他藏文预训练模型有大幅提升。本文将TiLamb和部分资源开放供研究使用，https://github.com/NLP-Learning/TiLamb。

关键词： 增量预训练；藏文大语言模型；指令微调；下游任务

TiLamb: A Tibetan Large Language Model Based on Incremental Pre-training

Wenhao Zhuang[1,2] Yuan Sun[1,2,3,*] Xiaobing Zhao[1,2,3]

[1] School of Information Engineering, Minzu University of China, Beijing　100081

[2] National Language Resources Monitoring and Research Center for Minority Languages

[3] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE

*Corresponding author: Yuan Sun

tracy.yuan.sun@gmail.com

Abstract

Based on the ”pre-training + fine-tuning” paradigm, language models have demonstrated exceptional performance. With the expansion of model size and training data volume, their ability to solve a variety of natural language processing tasks has seen significant improvements. Current large language models primarily support mainstream languages such as English and Chinese, limiting research in low-resource languages like Tibetan. Addressing issues such as the scarcity of Tibetan data, the inadequate performance of existing Tibetan pre-trained models, and poor expandability in downstream tasks, this paper consolidates and cleans 26.43GB of Tibetan data. Utilizing the open-source LLaMA2-7B as the foundational model, it expands the existing vocabulary of LLaMA2 by adding approximately 30,000 Tibetan tokens to enhance Tibetan text encoding efficiency and semantic understanding capabilities. Through incremental pre-training, the Tibetan large language model foundation, TiLamb, was developed.

第二十三届中国计算语言学大会论文集，第254页-第267页，太原，中国，2024年7月25日至28日。
卷1：主会论文

For various Tibetan downstream tasks, fine-tuning datasets ranging from thousands to tens of thousands of entries were prepared. The fine-tuned TiLamb was validated across seven Tibetan downstream tasks, including Tibetan news categorization, Tibetan entity relationship classification, Tibetan machine reading comprehension, Tibetan word segmentation, Tibetan summarization, Tibetan question answering, and Tibetan question generation, achieving significant improvements in multiple metrics over traditional methods and other Tibetan pre-trained models. This paper makes TiLamb and some resources available for research use at https://github.com/NLP-Learning/TiLamb.

Keywords: Incremental Pre-training , Tibetan Large Language Model , Prompt-based Fine-tuning , Downstream Tasks

1 引言

作为预训练语言模型（Pre-training Language Model, PLM）初期的创新性研究之一，ELMo(Sarzynska-Wawer et al., 2021)采用预训练的双向LSTM（BiLSTM）网络来实现对词汇在不同上下文中的动态表示，通过对特定的下游任务进行微调，进一步优化了BiLSTM网络的性能。近年来，通过在大规模语料库上对基于自注意力机制的高度并行化Transformer(Vaswani et al., 2017)架构进行预训练，人们发现BERT(Devlin et al., 2018)、GPT-2(Radford et al., 2019)、BART(Lewis et al., 2020)等基于"预训练+微调"范式的PLM在解决多种自然语言处理（Natural Language Processing, NLP）任务方面表现出强大的能力。进一步研究发现随着模型规模的扩大，这些语言模型的性能不仅得到了巨大提升，而且还表现出上述小规模语言模型所不具备的特殊能力（如上下文学习）。为区分不同参数规模下的语言模型，研究界创造了术语——大语言模型（Large Language Model, LLM）代指大型的PLM（包含数十亿乃至数千亿参数）。

有关藏语的预训练模型已经有了一些进展。Facebook AI的XLM-R模型(Conneau et al., 2020)使用了超过2TB的跨语言语料库，提出了一种高效、可扩展的跨语言表示学习方法，在近百种语言上进行了实验，取得了句子分类、命名实体识别等多个任务上的良好表现。哈工大讯飞联合实验室Yang等人(Yang et al., 2022)推出的少数民族语言的多语言预训练模型CINO是基于多语言预训练模型XLM-R开发的，涵盖了藏语等8种语言，在藏语新闻分类任务数据集TNCC(Qun et al., 2017)、朝鲜语新闻分类YNAT(Park et al., 2021)上获得了良好的效果。Liu等人(Liu et al., 2022)提出藏文预训练语言模型TiBERT，在问题生成等任务上取得突破。安波等人(安波and 龙从军, 2022)抓取了一个较大规模的藏文文本数据集，基于这些数据训练了BERT-base-Tibetan预训练模型，能够显著提升藏文文本分类的性能。Deng等人(Deng et al., 2023b)构建了包含蒙藏维哈朝五种少数民族语言的预训练模型MiLMo，为少数民族语言的各种下游任务提供支持。湖南大学Li等人(Li et al., 2022)使用超过10GB的汉英维藏蒙语料，在BART的基础上，加入DeepNorm训练得到少数民族预训练模型CMPT。Deng等人(Deng et al., 2023a)为了解决藏文预训练语言模型在知识记忆和推理能力上的缺陷，提出基于知识增强的藏文预训练语言模型TiKEM。

以上有关藏文的预训练模型虽然在多个下游任务上都取得了不错的表现，但均基于BERT设计，缺少解码器结构，并不属于真正的生成式藏文语言模型，且局限于分类等任务，模型参数量远低于LLM的参数量，对于藏文这种低资源语言，使用的训练数据还不够多。针对以上问题，本文提出基于LLaMA2-7B进行增量预训练的藏文大语言模型TiLamb（Tibetan Large Language Model Base），并在多个下游任务上验证了模型性能。本文的主要贡献如下：

（1）汇总清洗了26.43GB藏文预训练语料，使用SentencePiece(Kudo and Richardson, 2018)的BPE算法在近10GB藏文文本上训练得到词表大小为32,000的藏文分词模型。在原始LLaMA2的词表中增加额外约30,000个藏文tokens，增强了藏文的编码和解码效率，并提高了LLaMA2对藏文的理解能力。

（2）采用低秩适配（Low-Rank Adaptation, LoRA）方法，高效地对LLaMA2-7B进行增量预训练，并微调TiLamb适配藏文下游任务，在藏文新闻分类、藏文实体关系分类、藏文分

词、藏文摘要、藏文问题回答、藏文问题生成共六个下游任务中性能得到较大提升，在藏文机器阅读理解任务中取得较好结果。

（3）本文将藏文大语言模型TiLamb和部分资源公开，过程方法对藏语等低资源语言的LLM研究具有一定借鉴意义，促进低资源语言NLP的合作和发展。

2 相关工作

LLM的预训练对大规模文本数据的自动学习以捕获语言的复杂性和细粒度信息，能够在无需明确指示的情况下，捕获和理解文本中的隐含意义，从而增强其对新任务和未知数据的适应能力。LLM从零到一的预训练过程需要大规模的数据、计算资源以及时间，成本非个人和一般研究机构所能承受，因此，许多研究者利用现有的开源预训练模型作为基座，根据实际需求进行增量预训练和微调，往往能将大模型用于垂直领域，比如金融(Li et al., 2023)、法律(Blair-Stanek et al., 2023)等领域，并取得较好的效果。部分LLM训练统计数据如表1所示。

模型	参数大小	数据规模	计算资源	训练时间
T5(Raffel et al., 2020)	11B	1万亿tokens	1024 TPU v3	-
LLaMA(Touvron et al., 2023a)	65B	1.4万亿tokens	2048 80G A100	21天
LLaMA2(Touvron et al., 2023b)	70B	2万亿tokens	2048 80G A100	36天
BLOOM(Le Scao et al., 2022)	176B	3600亿tokens	384 80G A100	105天
CodeGeeX(Zheng et al., 2023)	13B	8500亿tokens	1536 Ascend 910	60天
ERNIE3.0(Zhang et al., 2019)	10B	3750亿tokens	384 V100	-
FLAN(Wei et al., 2021)	137B	-	128 TPU v3	60小时
Gopher(Rae et al., 2021)	280B	3000亿tokens	4096 TPU v3	920小时

表 1. 部分LLM预训练统计数据

微调旨在通过在特定任务相关数据上进行有监督学习来调整预训练模型的参数(Wei et al., 2021)，通常涉及对模型的最后几层或所有层进行调整，以便模型能够更好地适应特定的数据集和任务。由于LLM包含大量的参数，参数高效微调是一个重要的课题，旨在减少可训练参数的数量，同时尽可能保持良好的性能。适配器微调在Transformer模型中引入了小型神经网络模块（称为适配器）(Houlsby et al., 2019)，适配器模块将被集成到每个Transformer层中，通常使用串行插入的方式，分别在Transformer层的两个核心部分（即注意力层和前馈层）之后。前缀微调(Li and Liang, 2021)在语言模型的每个Transformer层前添加了一系列前缀，这些前缀是一组可训练的连续向量。与前缀微调不同，提示微调(Lester et al., 2021)主要是在输入层中加入可训练的提示向量。低秩适配LoRA(Hu et al., 2021)通过添加低秩约束来近似每层的更新矩阵，以减少适配下游任务的可训练参数。以上多种微调方式可以在不重复昂贵预训练过程的同时，实现对模型的快速适应和优化。

预训练不同语种语料的占比影响着LLM对不同语言的学习能力。LLaMA预训练使用的语料中近90%是英文语料，中文语料占比不足1%，对中文理解和生成能力有所欠缺，为了提高其中文对话能力，需要加入中文数据进行增量预训练，之后再微调使其适应下游任务，许多开源的Github项目基于这种思想开发，如Llama-Chinese、Chinese-LLaMA-Alpaca1/2(Cui et al., 2023)、Chinese-Vicuna(Chenghao Fan and Tian, 2023)等。基于这种迁移学习的思想，使用藏文语料在开源模型上进行增量预训练和微调，有望让LLM理解和生成藏文。

在藏语方面，哈工大讯飞联合实验室(Yang et al., 2022)提出了少数民族多语言预训练模型CINO，其中包含0.13B tokens藏文，CINO-Large版本的参数量为0.585B，参与训练的多种少数民族语言的tokens总数为3.37B。Liu等人(Liu et al., 2022)构建了覆盖藏语语料库99.95%的词汇表，提出了藏文预训练语言模型TiBERT，参数量为0.11B，藏文语料约3.56GB。Deng等人(Deng et al., 2023a)针对藏文预训练语言模型缺少外部知识指导，知识记忆能力和知识推理能力受限的问题，使用含有50万个三元组知识的藏文知识增强预训练数据集，训练了基于知识增强的藏文预训练语言模型TiKEM，参数量为0.115B，参与训练tokens总数为0.245B。安波等

人(安波and 龙从军, 2022)抓取了较大规模的藏文文本数据集，基于这些数据训练了BERT-base-Tibetan预训练模型。现有的藏文预训练模型在藏文下游任务上虽然表现不错，但其可拓展性有限，结果仍有很大的提升空间。主要原因是模型规模小，训练数据不足。解决这些问题，提高藏文预训练模型的能力，是本文研究的重点和难点。

3 TiLamb藏文大语言模型

3.1 LLaMA2模型基础

LLaMA2系列模型(Touvron et al., 2023b)是基于transformer架构(Vaswani et al., 2017)的纯解码器网络，集成了预归一化(Zhang and Sennrich, 2019)、SwiGLU激活函数(Shazeer, 2020)和旋转位置编码(Su et al., 2024)，实现了模型性能与训练效率的优化。覆盖不同参数规模（7B至65B），LLaMA2通过混合数据源预训练，在语言建模任务上展现了卓越性能。该模型结构包括32层，每层采用自注意力机制，其中线性投影实现了查询、键、值和输出的转换，而旋转嵌入强化了序列位置的表征。模型的MLP部分采用门控单元，增加了处理能力的非线性，而归一化层保证了训练的稳定性。最终，语言模型头部（lm_head）将隐藏状态映射到与词汇表大小相当的输出空间。

3.2 构建藏文分词模型

LLaMA2原始的分词模型是基于SentencePiece库(Kudo and Richardson, 2018)训练的。SentencePiece是一个用于训练分词器的开源库，它支持多种训练算法，包括unigram、bpe、word和char。在训练过程中，LLaMA2使用了BPE（Byte Pair Encoding）算法，它通过合并常见的字节对来构建词汇表，从而能够有效地处理各种语言。

本文收集近10GB的原始藏文语料，使用SentencePiece训练得到词表大小为32,000、覆盖率为99.95%的藏文分词模型，重要参数选择如表2所示，其余参数均为默认。

为了优化和适配LLaMA2的词表设计，本文在模型训练时进行了多项策略优化。首先，将所有数字字符拆分为单独的单元，减少了词表的数字量。其次，开启字节回退，在遇到未知或少见字符时将其分解为UTF-8字节来表示，避免OOV问题。此外，除了标准的藏文语料外，还加入了藏文的方言和变体（如古藏语），以确保模型能够处理各种地域性语言特征，提高分词模型的鲁棒性。

参数	值	参数	值
词表大小(`vocab_size`)	32000	分词算法(`model_type`)	BPE
数字拆分(`split_digits`)	True	字节回退(`byte_fallback`)	True
最大句子长度(`max_sentence_length`)	5000	字符覆盖率(`character_coverage`)	0.9995

表 2. 藏文分词模型训练参数

3.3 LLaMA2词表藏文扩充

LLaMA2-7B的预训练语料包含约2万亿个tokens，其中将近90%是英语，少部分是使用拉丁或西里尔字母的其他欧洲语言，因此LLaMA2具备多语言和跨语言理解能力。

本文在对LLaMA2进行初步研究时发现两个问题：（1）其对藏文的理解能力很差，无法理解藏文指令，这个问题的原因是LLaMA2的预训练语料缺少藏文，或者藏文语料占比太小导致预训练过程无法有效关注到藏文的向量表示；（2）LLaMA2在对藏文指令回复时，大部分是用英文回复无法理解指令，偶有的藏文回复也是对问题的重复或者错乱的藏文字符组合，且推理速度相比英文问答极其缓慢，导致速度缓慢的原因是LLaMA2词表中几乎没有藏文字符，尽管LLaMA2的分词器通过将未在词表中的内容分成多个Unicode字符来规避问题，但这样显著增加了序列的长度，而且降低了藏文文本的编码和解码效率。此外，每个藏文字符被分割为若干个Unicode字符，这些Unicode字符也可以表示其他语言，这会破坏藏文语义和语法结构的完整性，从而使得LLaMA2很难有效地学习捕捉藏文字符的语义表示。

本文将**3.2**节训练得到的藏文词表合并到LLaMA2原始词表中，合并后的词表大小为61,221，称为TiLamb分词器。对于同样一句藏文，使用LLaMA2原始分词器和TiLamb分词器的对比如表3所示。

	长度	内容
藏文语句	30	ཅ་དངོས་ཀྱགས་སྒམ་ཀ་ཕྱེ་ནས་ཞིབ་བཤེར་བྱས་རྗེས། འཕགས་སྐོ་ལས་ཁྲལ་ཀྱི་འཇེལ་ཡོད་ཤེ་སྣལ་ཤུམ་སྒམ་ནང་ངའི་སྒམ་ཆུང་ཞིག་རྙེད་ཕེབ།
		（打开行李箱检查后，海关人员在包装箱内发现了一个茶盒。）
LLaMA2 分词器	174	'▁', '<0xE0>', '<0xBD>', '<0x85>', '་', 'ད', 'ང', 'ོ', 'ས', '་', 'ག', '<0xE0>', '<0xBE>', '<0xB3>', 'ྱ', 'ག', 'ས', '་', 'ས', '<0xE0>', '<0xBE>', '<0x92>', 'མ', '་', ••• 省略剩余的 150 个 tokens
TiLamb 分词器	19	'▁ཅ', 'དངོས', 'ཀྱགས', 'སྒམ', 'ཀ', 'ཕྱེ་ནས', 'ཞིབ་བཤེར', 'བྱས་རྗེས', '▁', 'འཕགས་སྐོ', 'ལས་ཀྱི', 'འཇེལ་ཡོད', 'ཤེ་སྣལ', 'ཤུམ', 'སྒམ', 'ནང', 'ངའི', 'སྒམ་ཆུང', 'ཞིག་རྙེད', 'ཕེབ'

表 3. LLaMA2原始分词器和TiLamb分词器的分词结果对比

实验对比表明，对于相同一段藏文文本，换用TiLamb分词器后token数量从174缩减至19，减少约8倍，大幅提高了模型对藏文的编码效率和推理速度。在固定的上下文长度下，模型的最大输入/输出文本长度、模型可以容纳的信息量均提升至原来的8倍以上。

3.4 扩展词嵌入和调整lm_head维度

为了适配TiLamb分词器，本文将词嵌入和语言模型头部从原始尺寸$V \times H$调整为新尺寸$V' \times H$，其中$V = 32,000$代表原始词汇表大小，$V' = 61,221$是TiLamb分词器的词表大小。

LLaMA2-7B原始的Embedding层的大小为$(32000, 4096)$，即词表里的每一个token对应一个1×4096的Embedding向量。将新增的token在模型的Embedding层和lm_head层进行初始化，常用的有均值扩充和随机扩充两种方法，本文选用业界常用的均值扩充方法，即新增token的Embedding用原来token的Embedding的均值来表示。新的藏文词表被添加到原始嵌入矩阵的末尾，确保原始词汇表中的词嵌入不受影响。初始化后可得新增训练参数量为$(61221 - 32000) \times 4096 \times 2 = 239,378,432$。

3.5 预训练数据收集与处理

本文预训练数据主要包括：藏文新闻分类数据集TNCC(Qun et al., 2017)、UTibetNLP分类的藏文新闻数据(Zhang et al., 2022)、早期积累的藏文语料约15.5GB。爬取人民网藏文版、云藏网、西藏新闻网等二十余个藏文网站共计约7.5GB网页数据，包括党政文件、通知公告等多个领域。爬取云藏百科、维基百科等藏文百科数据3.1GB。爬取小红书和微信公众号文章约0.3GB。

以上数据均经过去重、隐私去除、质量过滤三种方式处理：

（1）去重：语料中的重复数据会降低语言模型的多样性，可能导致训练过程不稳定，从而影响模型性能(Hernandez et al., 2022),因此本文直接对重复句子进行过滤，同时删除包含重复单词或短语的低质量语句。

（2）隐私去除：本文的预训练文本数据大都来源于网络，包含涉及敏感或个人信息的内容，如联系方式、地址、邮箱等，这会增加隐私泄露的风险(Carlini et al., 2021)。本文使用直接有效的基于规则的方法，包括关键字识别、正则匹配等方法来对隐私内容进行去除。

（3）质量过滤：本文结合特定关键词的集合以及正则匹配，识别并删除文本中的噪声和无用元素，包括一些HTML标签、超链接、模板和攻击性词语。

处理后的藏文无标签预训练数据大小为26.43GB，藏文tokens总数约3B。

3.6 使用LoRA进行参数高效微调

传统的全参数更新训练范式对于大型语言模型来说计算时间和成本过高，低秩适配LoRA(Hu et al., 2021)是一种参数高效的训练方法，它保持预训练模型权重不变，同时引入可训练的秩分解矩阵。LoRA冻结了预训练模型的权重，并在每一层注入可训练的低秩矩

阵。具体来说，对于一个线性层，其权重矩阵为$W_0 \in \mathbb{R}^{d\times k}$，其中$k$是输入维度，$d$是输出维度，LoRA添加了两个低秩分解的可训练矩阵$B \in \mathbb{R}^{d\times r}$和$A \in \mathbb{R}^{r\times k}$，其中$r$是预定的秩。带有输入$x$的前向传播由公式(1)给出。

$$h = W_0x + \Delta Wx = W_0x + BAx, \quad B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times d} \tag{1}$$

在训练过程中，W_0 是固定的，不接收梯度更新，而B 和A 是可更新的。通过选择$r \ll \min(d, k)$，能够减少显存消耗。

为在成本和性能之间选择平衡点，实现参数的高效训练，本文将LoRA训练方法应用于增量预训练和微调两个阶段，主要将LoRA适配器集成到注意力模块和MLP层的权重中。

3.7 预训练目标

本文使用26.43GB藏文预训练数据在标准的因果语言模型（Causal Language Modeling, CLM）任务上对LLaMA2模型进行增量预训练。给定一个输入token序列$x = (x_0, x_1, x_2, \ldots)$，模型被训练以自回归的方式预测下一个token x_i，目标是最小化负对数似然，如公式(2)所示。

$$\mathcal{L}_{\text{CLM}}(\Theta) = \mathbb{E}_{x\sim D_{PT}}\left[-\sum_i \log p(x_i|x_0, x_1, \ldots, x_{i-1}; \Theta)\right] \tag{2}$$

其中，Θ 表示模型参数，D_{PT} 是预训练数据集，x_i 是待预测的token，$x_0, x_1, \ldots, x_{i-1}$ 构成上下文。

本文使用LLaMA-Factory框架(Zheng et al., 2024)进行LLaMA2-7B的增量预训练，部分参数的选择如表4所示。增量预训练之后得到藏文大语言模型TiLamb，过程中损失与学习率的记录如图1所示。

参数	值	参数	值
cutoff_len	1024	learning_rate	2×10^{-4}
finetuning_type	lora	num_train_epochs	1.0
per_device_train_batch_size	4	gradient_accumulation_steps	2
lr_scheduler_type	cosine	max_grad_norm	1.0
lora_rank	8	lora_dropout	0.1
lora_target	q_proj, v_proj	warmup_steps	0
additional_target	embed_tokens, lm_head, norm	fp16	True

表 4. 增量预训练过程参数选择

3.8 TiLamb监督微调

预训练语言模型难以遵循用户指令，经常会生成非预期内容。这是因为公式(2)中的建模目标是预测下一个token，而不是根据指令回答问题(Ouyang et al., 2022)。为了使语言模型的行为与用户的意图对齐，可以通过微调模型明确训练它遵循指令。本文采用LLaMA-Factory框架中的微调模板，其中instruction为用于描述任务的指令，input可选，为任务指令的补充输入，output为期望模型产生的回复，示例如图2所示。

在监督微调过程中，损失仅在输入序列的“output”部分计算，如公式(3)。

$$\mathcal{L}_{\text{SFT}}(\Theta) = \mathbb{E}_{x\sim D_{\text{SFT}}}\left[-\sum_{i\in output} \log p(x_i|x_0, x_1, \ldots, x_{i-1}; \Theta)\right] \tag{3}$$

其中，Θ 表示模型参数，D_{SFT} 是微调数据集，$x = (x_0, x_1, \ldots)$ 代表分词后的输入序列。

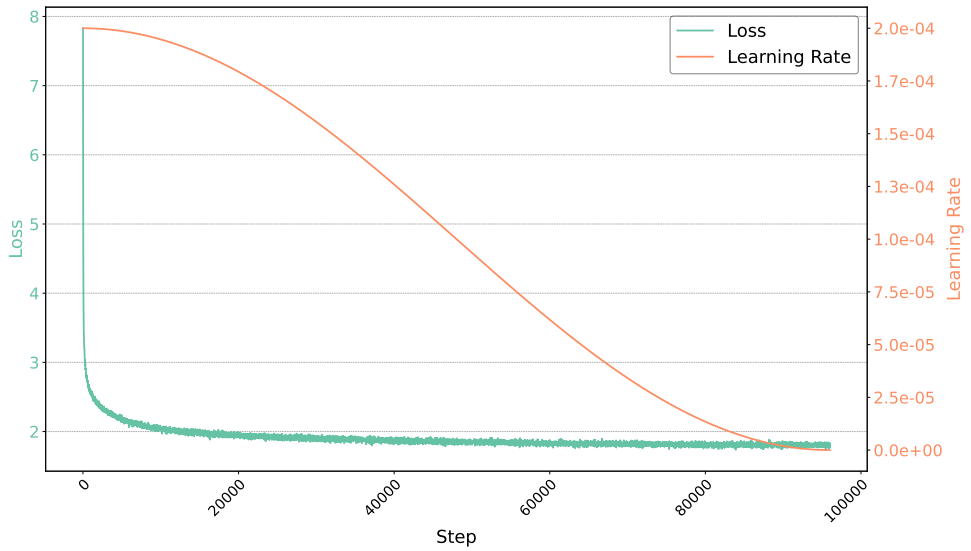

图 1. 增量预训练过程Loss与Learning Rate的变化

"instruction": "AIལས་རོགས་པ་ལ་མི་ཞེག་གིས་འཚམས་འདྲི་ཕྱིད་དུས་ལན་འདེབས་ཚོག་པའི་འཚམས་འདྲི་ཞུ་ཚོག་ཅིག་གསར་འཕྲུག་ཐུབ།"（创建一个 AI 助手在被人问候时可以说的问候语。）",

"input": "",

"output": "སྐུ་འཚས་བཟང་། དེ་རིང་ངས་ཁྱོད་ལ་ཅི་ཞེག་བསྒྲུབ་ཐུབ།"（“你好！今天我能帮你做些什么？”）"

图 2. 微调模板示例

4 TiLamb下游任务实验评估

在藏文大语言模型基座TiLamb上针对各项下游任务进行LoRA微调，合并适配器后调用API进行自动测试，收集实验结果并进行科学分析。TiLamb在藏文下游任务的监督微调中部分参数的选择如表5所示。

本文采用“一种下游任务对应一个微调模型”的策略，主要考虑到各类下游任务的数据规模和难度差异较大，使大模型难以同时掌握多种任务。此外，初步实验表明，设置不同提示词（prompt）虽能帮助模型区分任务，但会降低单个任务的性能。该策略无需在微调时设置提示词，提升了训练效率，并简化了API调用和用户使用过程。

参数	值	参数	值
cutoff_len	2048	learning_rate	2×10^{-4}
finetuning_type	lora	num_train_epochs	3.0
per_device_train_batch_size	4	gradient_accumulation_steps	2
lr_scheduler_type	cosine	max_grad_norm	1.0
lora_rank	8	lora_dropout	0.05
lora_target	q_proj, v_proj	fp16	True

表 5. 监督微调过程参数选择

4.1 藏文新闻分类

该任务选用由复旦大学自然语言处理实验室发布的藏语新闻数据集Tibetan News Classifi-

cation Corpus (TNCC)(Qun et al., 2017)。数据集包含9,204条样本，涉及政治、经济、教育、旅游、环境、艺术、文学、宗教等12个类别。本文按9:1的比例将其划分为训练集和测试集，训练集的数据用于制作微调数据集。评价指标为Accuracy(%)和Macro-F1(%)，实验结果如表6所示。

模型	Accuracy(%)	Macro-F1(%)
Transformer(Vaswani et al., 2017)	28.63	28.79
CNN(syllable)	61.51	57.34
TextCNN(Guo et al., 2019)	61.71	61.53
DPCNN(Johnson and Zhang, 2017)	62.91	61.17
TextRCNN(Lai et al., 2015)	63.67	62.81
Bert-base-Tibetan(安波and 龙从军, 2022)	-	51.00
TiBERT(Liu et al., 2022)	71.04	70.94
CINO-base(Yang et al., 2022)	73.10	70.00
TiKEM(Deng et al., 2023a)	74.46	72.61
TiLamb+LoRA	**78.85**	**77.45**

表 6. 藏文新闻文本分类结果

TiLamb使用训练集经LoRA微调后，在藏语新闻文本分类上取得了最好的分类效果，准确率比CINO-base和TiKEM分别提高了5.75%和4.39%，F1值比CINO-base和TiKEM分别提高了7.45%和4.84%。

4.2 藏文实体关系分类

为了验证TiLamb模型对知识的记忆及融合运用能力，本文使用6,433条三元组-文本对齐数据集，三元组中共有11种关系。该任务要求在给定两个实体和包含该实体的对应文本后，给出两个实体之间的关系类别。本文按9:1的比例将其划分为训练集和测试集，训练集的数据用于制作该任务的微调数据集。评价指标为Accuracy(%)、Macro-P(%)、Macro-R(%)和Macro-F1(%)，实验结果如表7所示。

模型	Accuracy(%)	Macro-P(%)	Macro-R(%)	Macro-F1(%)
FastText(Joulin et al., 2016)	55.80	34.05	32.98	31.61
DPCNN(Johnson and Zhang, 2017)	70.94	54.21	49.23	48.65
TextCNN(Guo et al., 2019)	72.38	71.03	59.11	56.76
TiBERT(Liu et al., 2022)	84.70	76.66	68.82	67.94
CINO-base(Yang et al., 2022)	85.31	75.48	69.12	66.73
MiLMO(Deng et al., 2023b)	85.76	77.13	68.97	68.57
TiKEM(Deng et al., 2023a)	90.12	91.73	75.61	76.34
TiLamb+LoRA	**95.98**	**97.14**	**88.98**	**91.60**

表 7. 藏文实体关系分类结果

TiLamb微调后在该任务上的分类准确率（Accuracy）上达到了95.98%，相比于其他预训练模型，TiLamb在所有评价指标上都有明显提升。

4.3 藏文机器阅读理解

机器阅读理解任务是给定一段文本和一个问题，让模型回答对应问题。这需要模型理解问题和上下文语义，然后进行推理、判断等，给出具体答案。本文使用藏文机器阅读理解数据集TibetanQA(Sun et al., 2021a)对模型的阅读理解能力进行评估，该数据集包含了1,513篇文章和20,000个问答对。为了评估模型性能，本文使用EM值（精确匹配）和F1值作为评价指标。

本文以8:2的比例将数据划分为训练集和测试集，训练集用于制作该任务的微调数据集，实验结果如表8所示。

模型	EM(%)	F1(%)
R-Net(Wang et al., 2017)	55.8	63.4
BiDAF(Seo et al., 2016)	58.6	67.8
QANet(Yu et al., 2018)	57.1	66.9
TiBERT(Liu et al., 2022)	53.2	73.4
TiLamb+LoRA	46.6	77.4
Ti-Reader(Sun et al., 2021c)	67.9	77.4
TiKEM(Deng et al., 2023a)	**69.4**	**80.1**

表 8. 在TibetanQA上的藏文阅读理解评估结果

TiLamb作为生成式模型，在做抽取式阅读理解任务时相比其他模型难度大，因为其他模型只需根据文本和问题，在文本中找出答案的起止位置即可，但生成式模型本就基于概率生成。即便如此，微调后的TiLamb的F1值依然达到了77.4%，与藏文抽取式机器阅读理解模型Ti-Reader的指标相同，只比TiKEM的F1值低2.7%。

4.4 藏文分词

对于藏文这种结构复杂、资源相对较少的语言而言，正确的分词对于进一步的语言处理任务，如语义分析、机器翻译和信息检索等，都具有至关重要的作用。该任务使用中文信息学会举办的第一届藏文分词评测所用的数据集，数据集中藏语短句去重后有21,427条，本文保留了1,000条作为测试集，剩余的20,427条用于制作微调数据集，测试结果与第一届藏文分词评测第一名TIP-LAS(李亚超et al., 2015)对比如表9所示。

模型	Precision(%)	Recall(%)	F1(%)
TIP-LAS(李亚超et al., 2015)	93.14	92.17	92.66
TiLamb+LoRA	**93.58**	**93.71**	**93.64**

表 9. 藏文短句分词评估结果

实验结果表明，TiLamb微调后在藏文短句分词任务中表现出色，其Precision、Recall 和F1值均超过了TIP-LAS 模型。具体而言，微调后的TiLamb的F1 值达到93.64%，比TIP-LAS提高了0.98个百分点。

4.5 藏文摘要

文本摘要是自然语言处理领域中的一项关键技术，它通过对大量文本进行提炼和压缩，去除冗余信息，生成简洁而全面的概要，从而帮助用户迅速把握文本的核心内容，大幅提升阅读效率。本文使用47,088条新闻与对应摘要做微调，测试集使用5,232条新闻与对应摘要。用于本文实验的数据均以新闻标题作为摘要，这种方法能够很好地捕捉新闻的主要内容，确保生成的摘要简明扼要且信息丰富。实验结果如表10所示。

模型	ROUGE-1(%)	ROUGE-2(%)	ROUGE-L(%)
统一模型(Huang et al., 2023)	19.81	13.27	16.90
CMPT模型（Ti-SUM）(Huang et al., 2023)	39.53	26.42	38.02
CMPT（50000条）(Huang et al., 2023)	49.16	33.43	48.66
TiLamb+LoRA	**53.99**	**37.22**	**52.89**

表 10. 藏文摘要生成评估结果

4.6 藏文问题回答

该任务使用本组人工制作的TiconvQA藏文多轮对话数据集，其中包含2,120条藏文文章段落及20,420对多轮问答对。为了进行实验评估，本文按照8:2的比例将数据集划分为训练集和测试集。在评估实验中，使用EM值和F1值作为评估指标，实验结果如表11所示。

模型	EM(%)	F1(%)
DrQA(Chen et al., 2017)	41.49	61.51
TiBERT(Liu et al., 2022)	45.12	65.71
TiLamb+LoRA	**45.28**	**72.84**

表 11. 单轮藏文问题问答评估结果

4.7 藏文问题生成

问题生成是自然语言生成的一项任务，它以文本和目标答案为输入，自动从答案中生成问题。本文使用用于机器阅读理解的藏文问答数据集TibetanQA(Sun et al., 2021a)，按照约9:1的比例划分训练集和测试集，其中用于微调的数据共17,762条，用于测试的数据为1,976条，实验指标数值均为百分比，实验结果如表12所示。

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
S2S+ATT+CP(Sun et al., 2021b)	29.99	20.14	13.90	9.59	31.45
TiBERT(Liu et al., 2022)	35.48	28.60	24.51	21.30	40.04
TiBERT+wh(Sun et al., 2022)	42.45	35.07	**29.64**	**25.58**	43.28
TiLamb+LoRA	**44.60**	**35.24**	28.88	24.47	**50.42**

表 12. 藏文问题生成评估结果

微调后的TiLamb在问题生成任务中有不错的表现，尤其在ROUGE-L上达到了50.42%，相较其他模型有较大提升。

5 总结与展望

本文使用近10GB藏文数据训练得到词表大小为32,000的藏文分词模型，并用约30,000个藏文tokens扩充了LLaMA2现有词表，显著提高了其对藏文的编码效率，收集整理了26.43GB藏文预训练语料，并对LLaMA2-7B进行增量预训练得到藏文大语言模型TiLamb。针对七个藏文NLP下游任务分别微调后的TiLamb，经实验验证相较其他模型均有不同程度的提升。本文的过程和方法对其他低资源语言在大型语言模型方向的研究提供了一定的参考价值。当前的TiLamb仍存在一定局限性，由于预训练语料大部分为藏文新闻，对TiLamb进行百万数量级的指令微调后，虽能显著提升LLaMA2的藏文对话能力和指令遵循能力，但TiLamb的藏文理解

能力还有一定进步空间。在未来将使用更多有关藏族文化、历史和宗教的相关数据在参数量更大的LLM上进行预训练和通用对话微调，并加入人类反馈的强化学习（RLHF）过程，使模型的输出与人类偏好对齐，同时探索DoRA(Liu et al., 2024)、GaLore(Zhao et al., 2024)等更先进的微调算法对藏文大语言模型的适用性。

致谢

本论文得到了国家社科基金(22&ZD035), 国家自然科学基金(61972436), 中央民族大学项目(GRSCP202316, 2023QNYL22, 2024GJYY43)的资助。

参考文献

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 22–31.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Zhenyi Lu Chenghao Fan and Jie Tian. 2023. Chinese-vicuna: A chinese instruction-following llama-based model.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Junjie Deng, Long Chen, Yan Zhang, Yuan Sun, and Xiaobin Zhao. 2023a. Tikem: 基于知识增强的藏文预训练语言模型(tikem: Knowledge enhanced tibetan pre-trained language model). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 135–144.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023b. Milmo: minority multilingual pre-trained language model. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 329–334. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi Ma. 2019. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*, 363:366–374.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. 2022. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shuo Huang, Xiaodong Yan, Xinpeng OuYang, and Jinpeng Yang. 2023. 基于端到端预训练模型的藏文生成式文本摘要(abstractive summarization of tibetan based on end-to-end pre-trained model). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 113–123.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. 2022. A multi-tasking and multi-stage chinese minority pre-trained language model. In *China Conference on Machine Translation*, pages 93–105. Springer.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. Cfgpt: Chinese financial assistant with large language model. *arXiv preprint arXiv:2309.10654*.

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

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Sungjoon Park, Sungdong Kim, Jihyung Moon, Won Ik Cho, Kyunghyun Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, et al. 2021. Klue: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*. Advances in Neural Information Processing Systems.

Nuo Qun, Xing Li, Xipeng Qiu, and Xuanjing Huang. 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, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings 5*, pages 472–480. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Y Sun, S Liu, C Chen, Z Dan, and X Zhao. 2021a. Construction of high-quality tibetan dataset for machine reading comprehension. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 208–218.

Yuan Sun, Chaofan Chen, Andong Chen, and Xiaobing Zhao. 2021b. Tibetan question generation based on sequence to sequence model. *Computers, Materials & Continua*, 68(3).

Yuan Sun, Chaofan Chen, Sisi Liu, and Xiaobing Zhao. 2021c. Ti-reader: 基于注意力机制的藏文机器阅读理解端到端网络模型(ti-reader: An end-to-end network model based on attention mechanisms for tibetan machine reading comprehension). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 219–228.

Yuan Sun, Sisi Liu, Zhengcuo Dan, and Xiaobing Zhao. 2022. Question generation based on grammar knowledge and fine-grained classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6457–6467.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, Vancouver, Canada, July. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations*.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 12381–12392.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Jiangyan Zhang, Deji Kazhuo, Luosang Gadeng, Nyima Trashi, and Nuo Qun. 2022. Research and application of tibetan pre-training language model based on bert. In *Proceedings of the 2022 2nd International Conference on Control and Intelligent Robotics*, ICCIR '22, page 519–524, New York, NY, USA. Association for Computing Machinery.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. 2023. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

安波and 龙从军. 2022. 基于预训练语言模型的藏文文本分类. 中文信息学报, 36(12):85–93.

李亚超, 江静, 加羊吉, and 于洪志. 2015. Tip-las: 一个开源的藏文分词词性标注系统. 中文信息学报, 29(6):203–207.