Modern Linguistics 现代语言学, 2026, 14(5), 328-334
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.145407

Hans汉斯

大语言模型时代壮语情感词典构建的方法路径

高　艳

广西民族大学外国语学院，广西　南宁

收稿日期：2026年4月9日；录用日期：2026年5月6日；发布日期：2026年5月18日

摘　要

壮语作为我国使用人口最多的少数民族语言，在自然语言处理领域长期处于低资源状态，情感词典等细粒度语义资源至今缺失。文章梳理情感词典构建方法的演进脉络，重点考察大语言模型时代涌现的三种新兴技术路径——人机协同主动学习、少样本上下文学习和参数高效微调，并结合壮语的资源现状逐一分析各路径的方法学要求与落地条件。研究表明，预训练语料覆盖薄弱、双语资源结构有限和专业人才稀缺三方面因素相互交织，使壮语情感词典建设难以依靠单一技术路径解决，需要根据资源条件灵活组合传统方法与新兴方法。

关键词

壮语，情感词典，大语言模型，低资源语言，跨语言映射

Methodological Pathways for Constructing a Zhuang Sentiment Lexicon in the Era of Large Language Models

Yan Gao

School of Foreign Languages, Guangxi Minzu University, Nanning Guangxi

Received: April 9, 2026; accepted: May 6, 2026; published: May 18, 2026

Abstract

Zhuang, the most widely spoken ethnic minority language in China, has long remained a low-resource language in natural language processing, with sentiment lexicons and other fine-grained semantic resources still absent. This paper traces the methodological evolution of sentiment lexicon construction and focuses on three emerging technical pathways enabled by large language models: human-in-the-loop active learning, few-shot in-context learning, and parameter-efficient fine-tuning.

文章引用: 高艳. 大语言模型时代壮语情感词典构建的方法路径[J]. 现代语言学, 2026, 14(5): 328-334.
DOI: 10.12677/ml.2026.145407

Each pathway is examined in light of the current resource conditions of Zhuang. The analysis reveals that three intertwined challenges, namely insufficient pretraining coverage, the limited structure of bilingual resources, and the scarcity of specialized expertise, prevent any single technical pathway from independently meeting the demands of Zhuang sentiment lexicon construction. A flexible combination of traditional methods and emerging approaches, calibrated to actual resource conditions, is therefore necessary. Among the three new pathways, few-shot in-context learning offers the strongest fit for current Zhuang research, owing to its lower technical threshold and reduced reliance on pretraining coverage.

Keywords

Zhuang Language, Sentiment Lexicon, Large Language Models, Low-Resource Languages, Cross-Lingual Mapping

1. 引言

情感分析作为自然语言处理(Natural Language Processing, NLP)的核心任务之一，已在社交媒体监测、商业智能、舆情研判等领域得到广泛应用。情感词典作为情感分析的基础资源，系统记录词汇的情感极性与强度等属性，通过词典匹配实现文本情感倾向的计算[1]。过去二十年间，研究者为英语等高资源语言构建了丰富的情感词典资源，推动了情感分析技术的快速发展。然而，全球约 7000 种语言中，仅少数高资源语言拥有充足的 NLP 资源，超过 90%的语言在数据集和工具建设方面严重滞后[2]。低资源语言(low-resource languages)通常被定义为因任务特定标注数据或辅助资源稀缺而难以有效应用计算技术的语言，其范围从拥有一定规模文本但缺乏标注的少数民族语言，到仅有数千条句子的濒危语言不等[3]。资源分布的不均衡不仅制约了这些语言情感分析技术的发展，也使其使用者在数字社会中面临信息获取与表达的双重障碍。

壮语是我国使用人口最多的少数民族语言，使用者超过 2000 万，主要分布在广西、云南、广东等地。作为壮侗语族的代表性语言，壮语承载着壮族独特的文化认知与情感表达方式，但在 NLP 领域却长期处于典型的低资源状态。情感词典等细粒度语义资源至今缺失，直接制约了壮语在情感分析、智能检索、机器翻译等下游任务中的发展。近年来，大语言模型(Large Language Models, LLMs)的快速发展为低资源语言情感词典建设带来了新的方法学契机[4]。本文以壮语为考察对象，梳理情感词典构建方法的演进脉络，重点考察大语言模型时代涌现的新兴技术路径，结合壮语的资源现状分析各路径的可行性，以期为壮语情感词典建设的方法选择提供参考。

2. 情感词典构建方法概述

情感词典的构建方法依据策略差异可归纳为四类：专家编纂与人工标注、语料驱动的统计诱导、基于词汇知识库的扩展，以及跨语言迁移与投射。专家编纂是最早出现也最为直观的构建方式，研究者通过人工甄选词汇并标注情感倾向，形成高一致性的情感词表，General Inquirer 与 MPQA 等英语情感词典正是经由这一路径建成[5]。该方式的优势在于标注质量稳定、可控，劣势则在于人力成本高昂，规模扩展受限。语料驱动的统计方法则借助大规模文本，通过词共现、模式挖掘或分布式度量自动诱导情感词汇，规模扩展能力较强，但前提是目标语言具备规模足够的文本语料[6]。对于壮语这类文本资源稀缺的

语言而言，语料统计路径短期内难以独立支撑情感词典建设。

基于词汇知识库的扩展方法利用已有的语义资源，通过同义词集关系将词汇的语义结构映射为情感结构，SentiWordNet 是这一路径最具代表性的资源。跨语言迁移与投射方法则借助翻译、并行语料或双语词典，将高资源语言的情感词典迁移到低资源语言，显著降低了从零构建的成本[7]。在四类方法中，前两类对壮语的适用性较弱，后两类相对具备可操作空间——壮语已有《壮汉英词典》等多语词典资源积累，为基于知识库的情感映射和跨语言迁移提供了基础接口。

3. SentiWordNet 跨语言映射路径与壮语的契合点

SentiWordNet 依托 WordNet 词汇数据库构建，为每个同义词集标注客观性、正面性和负面性三个 0 至 1 区间的分数，三者之和为 1 [8]。这一基于同义词集的情感标注体系为低资源语言通过英语中介实现情感迁移提供了可能。Das 和 Bandyopadhyay 最早将 SentiWordNet 迁移至孟加拉语和印地语，采用字典翻译与同义词集映射相结合的混合策略[9]；Dehkharghani 等[10]将这一思路应用于土耳其语情感词典 SentiTurkNet 的构建[10]；Shelke 等则探索了印地语–马拉地语词网映射的情感极性转移路径[11]。东南亚及东亚区域的低资源语言情感资源建设同样引发了学界关注。越南语方面，Nguyen 等构建了包含情感标注的学生反馈语料库 UIT-VSFC，这是越南语情感分析领域较早的系统性数据工作[12]；印度尼西亚语方面，Koto 等推出的 IndoLEM 基准数据集对印尼语 NLP 的多项基础任务进行了系统评测，呈现了该语言在词汇语义标注上的整体资源现状[13]。这些工作不仅验证了 SentiWordNet 跨语言迁移的可行性，也为后续低资源语言情感词典建设提供了可复用的方法论框架。

对壮语而言，SentiWordNet 路径的契合点主要体现在两个层面。一是资源接口的现成性：《壮汉英词典》作为壮语研究的基础多语词典，为壮语词汇与英文同义词集之间建立了天然的映射通道，避开了对成熟壮语 WordNet 的依赖。二是标注粒度的稳健性：SentiWordNet 的三分类极性体系粒度较粗，在跨语言映射中受文化差异的影响小于细粒度的情感强度标注。这两个特点共同决定了 SentiWordNet 路径在壮语情境下的方法学吸引力。然而，跨语言迁移方法也存在难以回避的局限——文化特异性词汇难以精确映射，语境依赖的极性转换难以处理，迁移结果的质量高度依赖双语词典本身的编纂深度。

4. 大语言模型时代的新兴技术路径

4.1. 大语言模型的技术基础与方法学优势

Transformer 架构的提出开启了预训练语言模型时代[14]，BERT、GPT 等通用模型以及 XLM、mBERT、mT5 等多语言模型通过在超大规模语料上预训练，习得了跨语言的语义与语法知识。相比传统方法，大语言模型在低资源语言情感词典建设中表现出几方面新特性。其一是语义理解能力的跨越，模型能够通过注意力机制捕捉上下文关联，处理一定程度的文化特异性表达，不再依赖逐词的字面映射。其二是对标注数据依赖度的下降，少样本学习与零样本推理使资源极度匮乏的语言也能借助提示完成任务。其三是任务适配的灵活性，同一基座模型可经由提示工程或轻量微调快速服务于不同的下游任务。这些特性恰好回应了壮语等低资源语言的资源约束条件，为壮语情感词典构建提供了过去不具备的方法学选项。

在上述三条路径之外，还有两个技术方向值得纳入视野：无监督跨语言表示学习与持续预训练(continual pre-training)。前者以 XLM-R 为代表，通过对百余种语言的无标注文本进行联合预训练，在不依赖对齐语料的条件下实现了跨语言语义空间的隐式对齐，为低资源语言借力高资源语言的表示知识提供了可行路径[15]。后者的思路则是在已有多语言模型的基础上，利用目标语言的少量文本继续训练，使模型对该语言的词汇和语义形成更稳定的表征，同时保留原有的跨语言迁移能力。Gururangan 等的研究表明，用规模有限的领域语料对预训练模型进行持续自适应训练，即便数据量不大，也能为下游任务带来明显的性

能改善[16]。就壮语而言，现有的壮文政府文件、学术论著及双语平行语料虽然规模有限，却可以作为持续预训练的语料来源，用于在 XLM-R 等通用多语言模型基础上训练一个对壮语语义更为敏感的适配版本。如此，后续的少样本学习或微调便有了更好的模型起点，从而形成“语言适应–知识注入–任务适配”的递进逻辑。当然，持续预训练的实际收益在壮语上究竟有多大，还需要实证研究的进一步检验。

4.2. 人机协同的主动学习

人机协同主动学习将大语言模型嵌入标注循环，由模型完成初步自动标注，系统再通过不确定性采样策略筛选信息量最大的样本交由人工复核，形成机器与人工相互补充的反馈回路。Kholodna 等将 GPT-4 集成至主动学习循环，在多种非洲语言上验证了该模式的有效性，研究表明大语言模型的引入可显著降低标注的人工工作量，同时维持接近全量标注的性能水平[17]。这一模式适合具备一定标注预算和专业人力但希望降低成本的语言场景，能够在标注质量与建设成本之间取得平衡。

对壮语而言，主动学习路径的潜在价值在于可借助大语言模型分担初步标注的工作量，使有限的壮语专业人力集中处理需要语言学判断的疑难词汇。但这一路径在壮语上的应用面临一个前提性难题：主动学习的有效性依赖大语言模型对目标语言的基础理解能力，而主流大模型对壮语的预训练覆盖严重不足，初步标注质量可能难以达到可用水平。要使主动学习在壮语场景下有效运行，需要先通过提示工程或预备性的语言注入手段，提升大模型对壮语词汇语义的基础处理能力，否则反馈回路本身将失去意义。

4.3. 少样本上下文学习

少样本上下文学习(In-Context Learning, ICL)通过在提示中提供若干标注示例，引导大语言模型完成未见过的任务，整个过程无需更新模型参数。Cahyawijaya 等系统考察了少样本 ICL 在多种低资源语言上的表现，发现精心设计的示例对齐策略显著优于简单的标签匹配[18]。Li 等进一步指出，对于资源极其匮乏的语言，零样本或少样本 ICL 配合合适的上下文对齐甚至能够取得超过传统参数微调的效果，证明了上下文学习在极低资源情境下的独特优势[19]。

ICL 路径在壮语上的可行性已经获得直接证据。Zhang 等针对完全未被大语言模型支持的壮语提出了 DiPMT++框架，该研究仅使用壮汉词典和少量平行语料，通过动态检索示例为模型提供语法和词汇信息，无需参数更新即在壮语翻译任务上取得显著提升[20]。这一研究表明即便壮语在大模型预训练阶段几乎缺席，通过设计的示例提示，模型仍能完成壮语相关的语义任务。这一结论对壮语情感词典构建具有直接的启示价值：借助壮语多语词典中既有的语义对应关系构建高质量的示例对，能够在不依赖大规模标注数据的前提下引导大模型完成情感标注任务。少样本 ICL 的另一优势是技术门槛和计算资源需求相对较低，这对研究力量薄弱的壮语 NLP 研究而言尤其友好。综合来看，少样本上下文学习是当前壮语情感词典构建中最具落地条件的大模型路径。

4.4. 参数高效微调

参数高效微调(Parameter-Efficient Fine-Tuning, PEFT)的核心思路是冻结大模型的主体参数，仅训练少量额外参数实现语言适配，在定制性与资源效率之间取得平衡。LoRA(Low-Rank Adaptation)是其中最具代表性的技术。余杰等以蒙古语为例，结合多语言编码、投影层映射与 LoRA 微调，在输入输出对齐和句子级语义增强方面取得了实质性进展，展示了本土语料融合对提升语言特异性语义捕捉的作用[21]。然而，参数高效微调在壮语情感词典建设中的应用条件相对苛刻。该路径的前提是具备一定规模的壮语标注语料，而壮语当前的标注资源规模远未达到这一门槛，短期内独立实施难度较大。

5. 壮语情感词典建设的方法论思考

5.1. 三方面挑战的相互交织

预训练语料覆盖不足是壮语情感词典建设面临的首要挑战。主流大语言模型的训练语料以英语和其他高资源语言为主，壮语所占比重极低，甚至几近于无。这一现状使得通用大模型在面对壮语任务时，往往缺乏基本的语义理解能力，需要借助额外的提示设计加以弥补。换言之，大模型路径在壮语上的应用并非“开箱即用”，而是需要先解决模型对壮语的“陌生感”问题。

第二个挑战在于双语资源结构的有限性。《壮汉英词典》为壮语和英语之间的跨语言映射提供了基础接口，但任何一部双语词典在收词范围和释义深度上都有边界。壮族文化中的一些特有情感表达、地方性的语义色彩，往往难以在英文中找到精确对应的词汇。这类文化特异性词汇正是情感词典建设中最关键、最敏感的部分，恰恰也是跨语言映射最容易失真的部分。

第三个挑战来自壮语 NLP 专业人才的稀缺。国内从事壮语计算语言学研究的团队主要集中在广西区内少数高校，整体规模有限，持续性投入也不足。无论选择哪一条技术路径，标注规范的制定、疑难词汇的判断、文化特异性表达的处理，都离不开既精通壮语又熟悉计算语言学的复合型人才。在某种意义上，人才匮乏比技术路径的选择更具决定性，它直接决定了壮语情感词典建设能走多远、走多稳。

5.2. 技术路径的适配性比较与组合策略

回到本文所讨论的几条技术路径，可以看到它们在壮语情境下的适配度并不一致。专家编纂在标注质量上无可替代，但壮语专业研究力量薄弱，难以支撑大规模的人工建设。语料统计依赖大规模文本，而壮文语料的稀缺直接堵住了这条路。SentiWordNet 跨语言映射借助《壮汉英词典》的现成接口，可操作性相对较强，适合作为壮语情感词典的起步框架。在大语言模型一侧，参数高效微调对标注数据的规模要求较高，短期内难以独立支撑壮语建设；主动学习的有效运行又依赖大模型对壮语具备基本的理解能力，而这一前提目前并不成立。相比之下，少样本上下文学习对预训练覆盖的依赖较小，技术门槛和算力需求也较低，是当前最贴近壮语实际的大模型路径。

由此可见，壮语情感词典建设不宜寄望于某一种方法的单兵突进，而应让不同方法各司其职、相互补足。一种较为务实的组合思路是先以 SentiWordNet 跨语言映射搭建起基础框架，再借助少样本上下文学习对基础框架进行扩展与精修，最后通过人机协同的方式对文化特异性词汇和争议性词条进行复核。这样的组合在起步阶段成本较低，在扩展阶段能借力大模型的语义理解能力，在质量把关阶段又保留了人工的判断空间。等壮语标注资源积累到一定规模之后，参数高效微调便可顺势接入，沿着已有的资源基础构建更具壮语适配性的情感标注模型，使整个建设过程形成由浅入深的递进结构。为使上述组合策略具备实际的可操作性，有必要对各阶段的技术流程加以细化。

具体而言，第一步是以 SentiWordNet 跨语言映射搭建种子词典。从《壮汉英词典》中提取情感相关词条，借助汉英对应关系建立壮语词汇到英文 WordNet 同义词集的映射链，再通过 SentiWordNet 查询各同义词集的正面性、负面性和客观性得分。对情感极性较为明确的词条(如正面或负面得分超过一定阈值)直接纳入初始框架，歧义词条和低置信度词条则暂时标注待复核，不急于入典。这一步的主要价值在于低成本地建立词典雏形，提供后续工作的操作基础。

第二步是以少样本 ICL 对种子词典进行扩展和精修。从已有的高置信度词条中抽取典型示例，构造包含壮语词汇、汉语释义和情感极性标注的提示模板，用于引导大语言模型对第一步遗留的低置信度词条及文化特异性词汇进行再标注。壮汉双语对照可以作为辅助信号：由于现有大模型对汉语语义的掌握远优于壮语，借助汉语释义传递语义信息，能在一定程度上弥补模型对壮语直接理解能力的不足。每轮

标注完成后，置信度较高的词条可补充进入词典，仍有争议的词条继续挂起，待人工介入处理。

第三步是人机协同复核，重点处理文化负载词、方言差异词和语境极性依赖词这三类机器标注最容易出错的词汇。建议采用双人独立标注加裁决的方式，以 Cohen’s Kappa 等一致性指标作为质量把关依据。人工复核的成果不应止步于修正个别词条，还应及时反馈至示例库——经专家审定的高质量标注对可以用于优化后续 ICL 的示例选取，使每轮人工投入都能转化为提示质量的持续改善。待词典规模积累到一定量级，参数高效微调便具备了启动条件，可以在已有资源基础上训练更贴合壮语语义特点的情感标注模型，整个流程由此形成由浅入深的递进结构。

6. 结语

本文梳理了情感词典构建方法的演进脉络，重点考察了大语言模型时代的三种新兴技术路径，并结合壮语的资源现状分析了各路径的方法学要求与适配条件。需要进一步思考的是，壮语的“低资源”并非天然属性，而是长期以来研究投入和学术关注分配不均的结果。在大语言模型出现之前，低资源语言 NLP 长期被困在“先有资源–再做研究”的循环中，资源匮乏制约研究开展，研究薄弱又反过来加剧资源匮乏。大语言模型的出现首次为打破这一循环提供了可能，少样本上下文学习、人机协同标注等新路径不再以大规模标注数据为前提，使壮语这类资源极度匮乏的语言也能在 NLP 研究中获得一席之地。从这个意义上说，壮语情感词典建设不仅是一项具体的资源工程，更是低资源语言借助大模型时代技术红利走出资源困境的一次方法学探索。

基金项目

本研究为广西壮族自治区研究生教育创新计划基金项目“壮语形容词极性类别标注研究”(项目号：YCSW2025303)的阶段性成果。

参考文献

[1] Mohammad, S.M. (2016) Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In: Meiselman, H.L., Ed., *Emotion Measurement*, Elsevier, 201-237. https://doi.org/10.1016/b978-0-08-100508-8.00009-6

[2] Joshi, P., Santy, S., Budhiraja, A., Bali, K. and Choudhury, M. (2020) The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 6282-6293. https://doi.org/10.18653/v1/2020.acl-main.560

[3] Hedderich, M.A., Lange, L., Adel, H., Strötgen, J. and Klakow, D. (2021) A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. *Proceedings of the* 2021 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Online, 6-11 June 2021, 2545-2568. https://doi.org/10.18653/v1/2021.naacl-main.201

[4] 陆小飞, 金檀. 大语言模型微调技术在语言分析与测试中的应用与展望[J]. 现代外语, 2025, 48(3): 413-421.

[5] Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011) Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, **37**, 267-307. https://doi.org/10.1162/coli_a_00049

[6] Darwich, M., Mohd Noah, S.A., Omar, N. and Osman, N.A. (2019) Corpus-Based Techniques for Sentiment Lexicon Generation: A Review. *Journal of Digital Information Management*, **17**, Article 296. https://doi.org/10.6025/jdim/2019/17/5/296-305

[7] Xu, Y., Cao, H., Du, W. and Wang, W. (2022) A Survey of Cross-Lingual Sentiment Analysis: Methodologies, Models and Evaluations. *Data Science and Engineering*, **7**, 279-299. https://doi.org/10.1007/s41019-022-00187-3

[8] Esuli, A. and Sebastiani, F. (2006) SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *LREC*, **6**, 417-422.

[9] Das, A. and Bandyopadhyay, S. (2010) SentiWordNet for Indian Languages. *Proceedings of the Eighth Workshop on Asian Language Resources*, Beijing, 21-22 August 2010, 56-63.

[10] Dehkharghani, R., Saygin, Y., Yanikoglu, B. and Oflazer, K. (2016) Sentiturknet: A Turkish Polarity Lexicon for Sentiment Analysis. *Language Resources and Evaluation*, **50**, 667-685. https://doi.org/10.1007/s10579-015-9307-6

[11] B. Shelke, M., Sawant, D.D., Kadam, C.B., Ambhure, K. and Deshmukh, S.N. (2023) Marathi sentiwordnet: A Lexical Resource for Sentiment Analysis of Marathi. *Concurrency and Computation*: *Practice and Experience*, **35**, e7497. https://doi.org/10.1002/cpe.7497

[12] Nguyen, K.V., Nguyen, V.D., Nguyen, P.X.V., Truong, T.T.H. and Nguyen, N.L. (2018) UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis. 2018 10*th International Conference on Knowledge and Systems Engineering* (*KSE*), Ho Chi Minh City, 1-3 November 2018, 19-24. https://doi.org/10.1109/kse.2018.8573337

[13] Koto, F., Rahimi, A., Lau, J.H. and Baldwin, T. (2020) Indolem and Indobert: A Benchmark Dataset and Pre-Trained Language Model for Indonesian NLP. *Proceedings of the* 28*th International Conference on Computational Linguistics*, Barcelona, 8-13 December 2020, 757-770. https://doi.org/10.18653/v1/2020.coling-main.66

[14] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**, 5998-6008.

[15] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., *et al*. (2020) Unsupervised Cross-Lingual Representation Learning at Scale. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 8440-8451. https://doi.org/10.18653/v1/2020.acl-main.747

[16] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., *et al*. (2020) Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 8342-8360. https://doi.org/10.18653/v1/2020.acl-main.740

[17] Kholodna, N., Julka, S., Khodadadi, M., Gumus, M.N. and Granitzer, M. (2024) LLMs in the Loop: Leveraging Large Language Model Annotations for Active Learning in Low-Resource Languages. In: Bifet, A., Krilavičius, T., Miliou, I. and Nowaczyk, S., Eds., *Lecture Notes in Computer Science*, Springer, 397-412. https://doi.org/10.1007/978-3-031-70381-2_25

[18] Cahyawijaya, S., Lovenia, H. and Fung, P. (2024) LLMs Are Few-Shot In-Context Low-Resource Language Learners. *Proceedings of the* 2024 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (Volume 1: Long Papers), Mexico City, 16-21 June 2024, 405-433. https://doi.org/10.18653/v1/2024.naacl-long.24

[19] Li, Y., Zhao, Z. and Scarton, C. (2025) It's All about In-Context Learning! Teaching Extremely Low-Resource Languages to LLMs. *Proceedings of the* 2025 *Conference on Empirical Methods in Natural Language Processing*, Suzhou, 4-9 November 2025, 29532-29547. https://doi.org/10.18653/v1/2025.emnlp-main.1502

[20] Zhang, C., Liu, X., Lin, J. and Feng, Y. (2024) Teaching Large Language Models an Unseen Language on the Fly. *Findings of the Association for Computational Linguistics ACL* 2024, Bangkok, 11-16 August 2024, 8783-8800. https://doi.org/10.18653/v1/2024.findings-acl.519

[21] 余杰, 飞龙, 郭陆祥, 等. 基于通用大模型的民族语言大模型构建技术[J]. 中文信息学报, 2025, 39(8): 75-81.