

侗语：基于大语言模型的侗族知识获取系统构建

黄蕊¹ 秦臻¹ 崔嘉^{1*}
HUANG Rui QIN Zhen CUI Jia

摘要

针对侗族文化知识资料零散、专业性强、获取难度大等问题，文章基于大语言模型构建侗族文化知识库，以提升民族文化的数字化保护与传播效率。首先，以侗族建筑、服饰、音乐等多领域内容为基础，构建多维度语料库；其次，采用 LangChain 检索增强生成（RAG）对模型进行构建与对比实验。依托该方法，实现了具备知识动态更新与问答服务功能的侗族文化知识系统“侗语”。研究结果表明，RAG 技术能够有效降低侗族文化知识的获取门槛，提高知识服务质量，为民族文化数字化保护与创新应用提供可行的技术方案。

关键词

侗族文化；民族文化保护；大语言模型；数字化传播；LangChain

doi: 10.3969/j.issn.1672-9528.2026.03.042

0 引言

民族文化作为人类文明的瑰宝，是群体精神归属的根基，也是文化对话的关键动力^[1]。侗族作为少数民族的重要分支，具有深厚的文化价值和艺术价值^[2]，但其知识体系尚未得到系统整合，公众获取难度大，制约了文化的深入传播与创新应用。

近年来，人工智能技术尤其是大语言模型的快速发展，为民族文化的数字化重现与智能交互提供了新的可能。已有研究在传统文化领域的应用取得一定成效，如中医药古籍^[3]、古诗词生成^[4]等，但在少数民族文化领域仍存在提升空间。同时通用大语言模型在生成民族文化内容时往往缺乏准确性与深度，难以满足专业需求。

为此，本文以侗族文化为研究对象，探索大语言模型在民族文化知识建模与数字化传播中的适用路径。本文通过对比实验筛选合适的基座模型，并收集多维度侗族文化资料，构建侗族语料库；同时分别实践了三种大模型构建路径：基于 LangChain 的检索增强生成、Ptuning v2 微调与 SFT 微调。在此基础上，选择表现更优的 LangChain 方法搭建侗族文化问答系统，并通过评估验证其有效性。本文主要贡献如下：

（1）提出了侗族文化知识建模与大模型训练的完整路径：系统化地建立从基座模型选择、语料构建到不同模型构建方法对比的流程，为民族文化的数字化保护与智能应用提

供了可复现的技术范式。

（2）构建了多维度、高质量的侗族文化语料库。涵盖侗族大歌、建筑、服饰、织锦、武术等多个主题，规模达 40 余万字，为侗族文化的大模型训练与后续研究奠定了数据基础。

（3）实现了基于 LangChain 的侗族文化问答系统：构建了一个具备实用价值和审美价值的文化辅助工具，支持知识库匹配，在文化细节还原与事实准确性方面优于原生模型，为民族文化的溯源式学习和创造性转化提供了便捷、可靠的服务。

本研究通过大模型技术与侗族文化研究的结合，降低了公众和相关设计人员获取侗族文化相关知识的门槛，为侗族文化的数字化重现提供方法论支持与实践工具，也为推动民族文化的深度挖掘与广泛传播提供方法论和实践案例。

1 相关理论及技术

1.1 大语言模型及其在传统文化中的应用

当前，大语言模型在传统文化数字化领域展现出巨大的应用潜力。张君冬等^[3]基于 Ziya-Llama-13B 构建了中医药古籍大模型“Huang-Di”，实现了古籍文本的智能理解与生成；刘江峰等^[4]以《四库全书》为基础开发了 SikuGPT，专注于古诗词的自动生成与语义分析；周正达等^[5]基于大语言模型实现了非遗陶瓷工艺知识图谱的高质量自动构建。这些研究成果充分验证了大语言模型在传统文化数字化中的有效性，但当前大语言模型技术在少数民族文化的研究应用中仍存在空白，恰好为这类文化的数字化重现与传播提供了新的技术路径。

1.2 大语言模型相关技术

检索增强生成（RAG）是一种结合了信息检索和生成模型的技术，旨在通过引入外部知识库的信息来辅助大语言模

1. 华南理工大学设计学院 广东广州 510006

[基金项目] 广州市哲学社会科学“十四五”规划课题（2024GZGJ17）；浙江大学计算机辅助设计与图形系统国家重点实验室开放课题（Grant No.A2416）；广州市基础研究计划基础与应用基础研究专题一般项目（SL2024A04J00955）

型（large language models, LLMs）生成更准确且符合上下文的答案。当大模型需要生成文本或回答问题时，会先从一个庞大的文档集合中检索出相关信息，然后利用这些检索到的信息来指导文本的生成。本研究采用 LangChain^[6] 这一集成 RAG 能力的先进框架进行侗族文化大模型的构建。该框架具备良好的模块化设计与强大的功能集成能力，能够以较低的实现成本快速构建基于大语言模型的端到端应用。

尽管大语言模型具备强大的计算与预测能力，但其在特定领域的适用精度往往受限于语料差异与概念脱节，模型微调是缓解这一问题，实现领域迁移的关键技术^[7]。在文本分类等任务中，监督微调（supervised fine-tuning, SFT^[8]）作为一种常用训练范式，在不同微调策略中具有高度的灵活性。它既可与全参数微调结合，对模型进行全面优化，也可嵌入高效参数微调框架中，仅调整少量参数，以适应计算资源受限的训练环境^[9]。本文基于 ChatGLM-6B 基座模型，探索适合构建侗族文化大模型的高效微调策略，为构建领域化大模型提供实验依据。

2 侗族文化知识建模与大模型训练

通过对比相同数据微调的效果选择合适的基座模型，同时构建侗族语料库，再分别从三个路径进行侗族文化大模型的构建实践：基于 LangChain 的检索增强生成、Ptuning v2 微调和 SFT 微调。最后对比三种路径的实践效果，选择更优者搭建侗族文化问答系统——“侗语”，本文研究框架如图 1 所示。

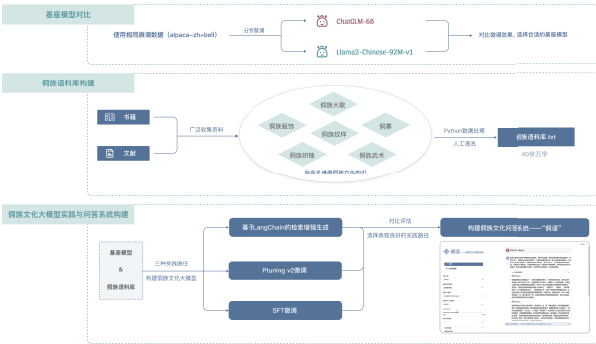


图 1 本文研究框架

2.1 基座模型

不同基座模型在微调后往往表现出差异化的性能特点。本研究选取当前广泛使用的两款开源轻量化模型——ChatGLM-6B^[10] 与 Llama2-Chinese-92M-v1 作为候选，在本地环境中使用相同规模数据集（alpaca-zh+bell，约 470×10⁶）进行微调，从知识吸收能力、知识保持能力和生成稳定性三个维度系统评估其微调表现，以确定更适合领域适应的基座模型。

对比结果如表 1 所示，ChatGLM-6B 在微调后表现出更

优的综合性能，能有效学习新语料中的细节知识，生成内容显著丰富具体，具有更强的知识吸收能力、知识保持能力和生成稳定性。而 Llama2-Chinese-92M-v1 微调后则表现不佳，其回答出现内容笼统、细节缺失等问题，甚至出现语义混乱或不合逻辑的生成结果。因此，本研究选择 ChatGLM-6B 作为侗族文化大模型构建的基座模型，并以此为基础开展后续工作。

表 1 基座模型微调效果对比

评估维度	测试问题	ChatGLM-6B	Llama2-Chinese-92M-v1
知识吸收能力	请简要介绍中国剪纸艺术的特点	能准确归纳剪纸艺术的“形态独特、手法细腻、寓意深刻、色彩丰富”等核心特征，体现对传统艺术的理解。	回答偏离主题，将剪纸艺术误述为“北京艺术”，内容失准。
知识保持能力	简要介绍中国的茶文化	完整保持对茶文化体系的理解，从茶叶种类、茶器茶席到饮茶方式均有准确说明。	出现知识遗忘和混淆，回答零散且逻辑混乱。
生成稳定性	请阐述中国传统节日清明节的文化内涵	回答结构完整，从祭祖、感恩、缅怀等多维度系统阐述，逻辑清晰连贯。	内容完全偏离主题，生成失常。

2.2 侗族语料库构建

侗族文化相关的语料数量有限且分布零散，为系统整合该领域的文本资源，研究人员广泛搜集了涉及侗族纹样、侗族大歌、侗族建筑等多方面的书籍与学术文献，覆盖了侗族文化的主要维度，如图 2 所示。

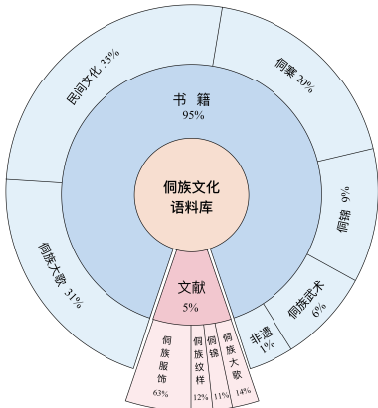


图 2 侗族语料库构成

通过对原始材料进行清洗与整理，最终构建了一个规模达 40 余万字的侗族文化语料库，为后续模型预训练奠定了数据基础。在语料预处理阶段，针对原始 PDF 格式文献的转换需求，选用 fitz 第三方库完成批量格式转换任务。转换为 TXT 格式后，语料经过人工检查与二次清洗，剔除其中存在

的乱码、文献编号、刊物说明等无关信息，为模型训练提供可靠的数据支持。

2.3 侗族文化大模型构建

为探索适合侗族文化大模型构建的技术路径，本研究系统对比了三种实践方法：基于 LangChain 的检索增强生成、Ptuning v2 微调以及 SFT 微调。本文从文化丰富度、文化准确性、可维护性和资源效率四个维度进行评估，以确定最适合侗族文化数字化传承的技术方案。

(1) 文化准确性。在案例 1“侗族蜘蛛纹有什么含义”中，基于 LangChain 的回答则明确关联至“萨神”这一侗文化核心图腾，并提出祖先崇拜、智慧祥瑞和生命繁衍的深层内涵，显示出对文化深层符号的准确把握。Ptuning v2 则错误地将蜘蛛纹含义解释为头饰相关内容，反映出模型对领域知识的理解存在偏差。回答详情如图 3 所示。

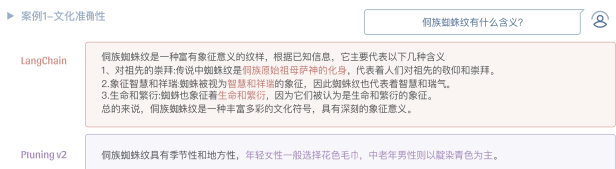


图 3 文化准确性对比

(2) 文化丰富度。在案例 2“侗锦根据色彩可以分为哪几类”中，基于 LangChain 的回答准确指出应分为“素锦”与“彩锦”两大类，并进一步阐明“素锦多使用白、灰、青等冷色系纱线编织”“彩锦采用互补色对比搭配”等细节，回应更具结构性与文化一致性。Ptuning v2 微调仅能回答出两个类别名称，缺乏具体细节说明，文化丰富度一般。回答详情如图 4 所示。



图 4 文化丰富度对比

(3) 可维护性。基于 LangChain 的检索增强生成方法支持动态更新，可通过知识库管理界面直接添加、编辑或删除知识条目，维护成本低。Ptuning v2 微调和 SFT 微调需要重新训练整个模型才能更新知识，维护成本较高，尤其是 SFT 微调需要调整大量参数，维护难度较大。

(4) 资源效率。基于 LangChain 的检索增强生成方法无需训练过程，仅需存储知识库和运行检索算法，资源效率高，Ptuning v2 微调和 SFT 微调均有较高的计算资源和时间成本要求，资源效率较低。

基于以上对比分析，三种方法在各维度上表现存在差异。LangChain 方法在文化准确性、文化丰富度、生成质量和可维护性方面均表现最优，适合文化细节要求高、需要频繁更

新的侗族文化应用场景；Ptuning v2 微调在资源效率方面具有一定优势，但文化准确性和生成质量不稳定，适合对实时性要求不高的场景；SFT 微调虽然训练充分，但文化适应性较差，且资源消耗大，不适合小样本文化语料场景。综合考虑侗族文化传播的实际需求——包括文化准确性要求高、知识需要持续更新、计算资源有限等特点，本研究选择基于 LangChain 的检索增强生成方法作为侗族文化大模型的构建方案。该方法不仅能够提供准确、丰富的文化内容，还支持知识库的动态维护，为侗族文化的数字化传承提供了可靠且可持续的技术路径。

3 侗族文化问答系统设计

基于前文实验对比，本研究基于 LangChain 检索增强生成，构建了名为“侗语”的侗族文化问答系统。系统前端采用 Streamlit 框架实现。系统部署于本地服务器，提供知识库管理功能，支持对文化语料进行动态更新与维护。视觉上深度融合侗族文化元素，旨在为用户提供准确、专业且具有文化认同感的交互体验。系统支持查看知识库匹配结果，支持用户开展溯源式学习，确保内容的真实性与可靠性。最终效果如图 5 所示。



图 5 侗族文化问答系统“侗语”

网页整体设计注重侗族文化符号的视觉化表达，将侗族文化融入网页设计，有助于激发用户的情感共鸣与文化认同。Logo 选用侗族最常见的凤鸟纹作为核心元素，承载着吉祥、美好的象征意义；Logo 内部以侗锦花纹填充，体现了侗族文化的独特性与深厚底蕴。网站整体配色提取自侗锦经典色彩，问答头像采用侗族“阿哥”形象，强调侗族文化氛围，也进一步突出了本研究在数字化平台中传承活化民族文化的目标。

4 侗族文化大模型问答评估实验

为系统评估“侗语”文化问答系统的实际效果，本文从文化准确性、术语规范性、信息完整度三个核心维度，与原生 ChatGLM-6B 模型进行了对比分析，并辅以案例说明。

案例 1 中，面对“详细介绍一下侗族原始祖母”问题，如图 6 所示，原生模型将萨神简单归类为“萨满教女神”，缺乏侗族特有的文化称谓和具体事迹描述。而“侗语”系统的回答则准确使用侗族文化特有的称谓“萨岁”“萨玛”，并提供具体的历史传说内容，体现更高的文化准确性。

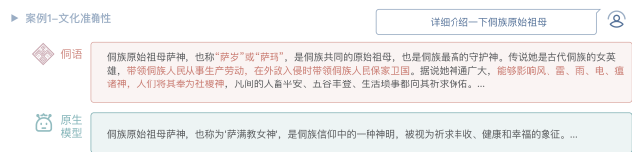


图6 文化准确性对比

案例2中,“侗语”系统回答严格限定在“口传类民间文艺”,列举的六大类全部符合这一范畴,准确使用了侗族文化研究中的专业术语,并指出了“珠郎娘美”这一特定叙事歌及其“汉字记侗音”的特殊传承方式,展现了深厚的专业知识。如图7所示。原生模型的回答则混淆了“口传文艺”与“广义文化”,信息量有限,甚至出现专业性错误:侗语属于汉藏语系-壮侗语族-侗水语支,而非原生模型所提到的苗语族瑶语族。



图7 术语规范性对比

案例3中,“侗语”系统将侗族妇女服饰科学划分为“紧身束型裙装、宽松型裙装和裤装”三大类,从色彩搭配、图案特征、地域差异、场合区分等多维度系统阐述,并援引专家研究作为权威佐证,体现了更高的信息完整度与结构层次性。而原生模型缺乏分类体系与细节支撑,信息零散模糊。如图8所示。



图8 信息完整度对比

实验结果表明,本研究构建的“侗语”系统在文化准确性、术语规范性、信息完整度三个核心评估维度上均显著优于原生 ChatGLM 模型。该系统能够准确使用文化特有称谓和具体事迹描述,避免笼统表述;且严格使用专业术语,避免学术性错误和范畴混淆;同时提供系统化的分类体系和详细的特征描述。经评估,本研究构建的侗族文化系统有效提升了民族文化数字化传承的质量和可靠性。

5 结语

本文系统探讨了侗族文化大语言模型的构建路径及其应用价值。研究从基座模型的选择、侗族文化语料库的构建,到多种技术方案下的大模型实践,再到侗族文化问答系统的实现,形成了一套完整的侗族文化大模型研究和实践框架。本研究不仅为侗族文化的传承与活化提供了新的方法与技术工具,也为民族文化的数字化传播提供了可借鉴的路径与实

践案例,对推动民族文化在数字时代的创新表达与传播具有积极意义。

参考文献:

[1] 罗国,蔡安宁.民族文化展示空间:多感官体验与空间结构优化的协同探索[J].新楚文化,2025(24):35-37.

[2] 曾斌.侗族文化的信息可视化设计研究[D].桂林:桂林电子科技大学,2024.

[3] 张君冬,杨松桦,刘江峰,等.AIGC 赋能中医古籍活化:Huang-Di 大模型的构建[J].图书馆论坛,2024,44(10):103-112.

[4] 刘江峰,刘维菲,齐月,等.AIGC 助力数字人文研究的实践探索:SikuGPT 驱动的古诗词生成研究[J].情报理论与实践,2023,46(5):23-31.

[5] 周正达,王昊,汪琳,等.ChatKG:一种基于大语言模型和提示工程的非遗知识图谱构建框架:以中国非遗陶瓷制作工艺为例[J/OL].图书馆杂志,1-30[2026-02-24].https://link.cnki.net/urlid/31.1108.g2.20250224.1136.008.

[6] TOPSAKAL O, CETIN AKINCI T. Creating large language model applications utilizing langchain: a primer on developing LLM Apps fast[C]//International Conference on Applied Engineering and Natural Sciences. Berlin: Springer, 2023: 1050-1056.

[7] 刘旭凯,刘洋,黄浩桢.EC-MIP:基于高效微调的知识产权小参数大语言模型[J/OL].计算机科学,1-15[2026-02-24].https://link.cnki.net/urlid/50.1075.TP.20250902.1423.017.

[8] PROTASHA N J, SAMI A A, KOWSHER M, et al. Transfer learning for sentiment analysis using BERT based supervised fine-tuning[J].Sensors, 2022, 22(11): 4157.

[9] 孙青琳,王小梅,陈挺,等.基于大语言模型微调的论文研究阶段分类方法改进:以科学结构图谱为例[J/OL].数据分析与知识发现,1-17[2026-02-24].https://link.cnki.net/urlid/10.1478.G2.20250903.1653.002.

[10] GLM T, ZENG A, XU B, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools[EB/OL].(2024-07-30)[2026-01-22].https://doi.org/10.48550/arXiv.2406.12793.

【作者简介】

黄蕊(2003—),女,安徽六安人,硕士研究生,研究方向:侗族文化、设计计算、风格分析。

秦臻(1988—),女,湖南益阳人,博士,讲师,研究方向:纺织服装设计、设计介入社会创新。

崔嘉(1982—),通信作者(email:cuijia1247@scut.edu.cn),男,山东济南人,博士,硕士生导师、副教授,研究方向:智能设计、衍生设计、民族文化推广。

(收稿日期:2025-12-05 修回日期:2026-03-11)