



文献 CSTR:

32001.14.11-6035.csd.2024.0169.zh



文献 DOI:

10.11922/11-6035.csd.2024.0169.zh

数据 DOI:

10.57760/sciencedb.j00001.01088

文献分类: 计算机科学技术

收稿日期: 2024-09-30

录用日期: 2025-02-19

发表日期: 2025-09-03

2024 年西北五省旅游资源 5 个关键要素的大模型

微调数据集

卢保青^{1,2}, 万福成^{1,2*}, 于洪志^{1,2}, 陈敏^{1,2}

1. 西北民族大学, 语言与文化计算教育部重点实验室, 兰州 730030

2. 西北民族大学, 甘肃省民族语言文化智能信息处理重点实验室, 兰州 730030

摘要: 随着大模型在各领域的广泛应用, 旅游行业对高质量数据集的需求日益增长, 以支持模型在旅游信息理解与生成方面的能力提升。本数据集专注于旅游领域文本数据的构建, 旨在为面向旅游大模型的微调任务提供高质量的数据支持, 以提升大模型在旅游相关信息的理解与生成能力。数据集的多样性和质量对模型性能至关重要, 因此本研究采用了爬虫技术与人工标注相结合的方式, 结合数据清洗、去噪与停用词处理, 确保数据的高质量与准确性。此外, 为保证数据一致性, 本研究还使用了自动化标注工具进行文本指令生成与一致性校验。本数据集 (LLM-Tourism) 主要依托携程网与百度百科两个平台, 涵盖甘肃、宁夏、青海、陕西和新疆五个西北省份, 包含 53280 对结构化数据, 格式为 json 文件。本数据集的构建不仅将提升旅游大模型的生成精度, 也为旅游领域的大模型数据集共享与应用提供了重要支持。

关键词: 文本数据; 旅游大模型; 微调; 生成

数据库(集)基本信息简介

数据库(集)名称	2024 年西北五省旅游资源 5 个关键要素的大模型微调数据集
数据通信作者	万福成 (wanfucheng@126.com)
数据作者	卢保青, 万福成, 于洪志, 陈敏
数据时间范围	2020 年 1 月–2024 年 9 月
地理区域	甘肃省、宁夏回族自治区、青海省、陕西省、新疆维吾尔自治区
数据量	46.6 MB
数据格式	*json
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01088
基金项目	甘肃省基础研究创新群体项目 (24JRRA154); 兰州市青年科技人才创新项目 (2024-QN-167); 甘肃省高校研究生“创新之星”2025CXZX-243)。
数据库(集)组成	数据集是一个文本文件构成, 文本的格式为 json 格式保存, 数据集的样本包括 53280 对结构化数据。

引言

近年来, 随着旅游业的蓬勃发展, 大量旅游相关数据的涌现为智能旅游推荐、景点评价与规划提供了丰富的研究基础。然而, 现有旅游大模型的性能仍受限于数据集的质量与多样性, 特别是在区域性文化深度挖掘和高精度语义生成方面的不足尤为显著。甘肃、宁夏、青海、陕西和新疆五省, 作为中国西北地区的重要

* 论文通信作者

万福成: wanfucheng@126.com

旅游目的地，尽管以其丰富的自然景观和文化遗产吸引了众多游客，但相关的数据资源却未得到系统的整合与充分利用。其中，甘肃省有 480 家 A 级旅游景区，宁夏有 137 家 A 级旅游景区，青海有 179 家 A 级旅游景区，陕西有 257 家 A 级旅游景区，新疆有 599 家 A 级旅游景区。因此，开发一个覆盖西北地区特色旅游资源的高质量数据集，既能为区域旅游文化的传播提供数据支撑，也能推动智能旅游大模型的性能提升，具有重要的理论和实际意义^[1]。

目前，国内外在旅游数据集的开发方面已取得一定成果，为区域旅游资源的研究和智能化应用奠定了基础。韩立钦等^[2]针对黄河流域的旅游资源进行了空间分布分析，为旅游资源的区域规划提供了参考；康建芳等^[3]聚焦于黄河流域的人文经济数据，探索了经济与文化资源之间的关联性，为跨领域研究奠定了基础。此外，万福成等^[4]基于北京旅游景点评论文本，构建了情感分析数据集，推动了自然语言处理在旅游情感研究中的应用。R Catelli 等^[5]使用 TripAdvisor 数据集进行建模，该数据集包含旅游评论和推荐，也常使用其评论数据进行情感分析、景点评价和推荐系统的构建。Nabiha Asgha 等^[6]使用 Yelp 公开数据集构建评论评分预测模型，该数据集提供了丰富的商户和用户评价数据，其中包括大量旅游相关的景点、餐饮和住宿评价。该数据集被广泛用于自然语言处理、推荐系统和用户行为分析等研究，该数据集通常包含用户评论、评分、景点信息等。尽管这些数据集在旅游分布、情感分析和评价分类等特定领域取得了显著进展，但在多任务适配性、区域文化深度挖掘和高质量语义生成方面仍显示出一定的局限性。

而本数据集（LLM-Tourism）通过多任务统一表示、实施递进式标注流程和文化特异性采集的方法，使得在这些方面具有显著优势。本数据集专注于西北五省的特色旅游资源，覆盖了从景点描述到文化背景的广泛内容，补充了黄河流域上游地区旅游资源数据。数据集结合了携程网的实用信息与百度百科的深度文化内容，兼顾了实际应用需求与理论研究的需求。在数据格式方面，LLM-Tourism 数据集采用了 LLaMA-Factory 微调工具所制定的通用微调的数据格式。LLaMA-Factory 整合了主流的多种高效训练微调技术，适配市场主流开源模型，形成一个功能丰富、适配性好的训练框架，提供了多个高层次抽象的调用接口，包含多阶段训练、推理测试、benchmark 评测和 API Server 等，大幅度降低了大模型部署和微调的难度。最后，通过严格的数据清洗与结构化处理，本数据集在多样性和质量控制方面更具优势，为复杂任务场景下的大模型微调提供了可靠支持。

本数据集涵盖的 53,280 对结构化文本数据，以 JSON 格式组织，主要针对西北五省甘肃、宁夏、青海、陕西和新疆的旅游资源构建多维度描述。数据包含以下核心类别：

基础信息：景点名称、所在省份与具体地理位置（如地址、经纬度）、开放时间及景区等级（如 AAAAA 级）。

内容描述：涵盖景点简介、特色亮点、自然与人文景观的详细说明。

历史文化：包括历史沿革、文物价值、保护现状及相关传说轶事。

服务信息：周边餐饮、住宿推荐及游览路线建议。

文化扩展：地方美食的制作工艺、民族节庆活动及非物质文化遗产相关内容。

数据字段通过携程网与百度百科双源互补，前者侧重实用信息（如地址、开放时间），后者提供文化深度（如历史背景、建筑特色）。

通过整合多源数据，涵盖景点、历史、美食等多维度内容。相比于现有数据集，本数据集的主要优势体现在以下方面：

- （1）区域性与文化深度：全面覆盖西北五省的旅游资源，注重历史文化与民族特色的挖掘。
- （2）数据质量高：通过严格的数据清洗、自动标注和一致性校验，确保数据准确性与适用性。
- （3）结构化与多样性：数据集存储格式规范、内容丰富，可支持多种任务场景（如文本分类、生成及推荐）。

本数据集的构建不仅丰富了西北旅游领域数据资源的收集，也为未来的旅游大模型研究提供了数据支持。

1 数据采集和处理方法

1.1 数据来源

实验数据集的构建是通过从携程网和百度百科中提取西北旅游相关的多种信息，包括景点、美食文化、酒店住宿以及历史文化等文本内容。为确保数据的丰富性和多样性，使用了网络爬虫技术，爬取时严格遵守相关法律法规与平台服务条款，并限制了爬取频率，以避免对平台服务器造成过载。同时，仅收集公开的非敏感信息，用于学术研究，不涉及个人隐私或商业用途，确保数据获取过程的合规性。爬虫自动抓取这些平台上发布的信息，能够高效地访问网页，提取结构化和非结构化数据，从而收集到关于甘肃的最新旅游资讯。这些文本数据不仅帮助构建一个全面且具有实用价值的数据集，还为后续的分析研究奠定了坚实的基础^[7]。通过这些精确而丰富的数据，研究者能够深入挖掘西北旅游的特色与潜力，提供更具参考价值的旅行建议。

1.2 数据采集与合法性说明

第一阶段：导入必要的库，用于发送网络请求、解析 HTML、处理 Excel 文件等任务^[8]。

第二阶段：定义全局变量。在代码开始处定义了一个 `users` 集合，用于记录已经保存过的用户名，避免重复写入。定义两个函数分别用于发送 HTTP 请求获取网页内容和解析单个网页的 HTML 内容，并提取相应文本信息。

第三阶段：运行主函数，创建一个 Excel 文件，并定义一个工作表。循环爬取指定页面范围的相应文本信息，每次循环发送一个 HTTP 请求获取网页内容，并调用函数解析页面内容，将提取的数据写入 Excel 表格中。并使用多线程进行爬取，创建一个线程池，通过 `map` 方法并发执行多个爬取任务，每个任务只爬取一部分页面的数据，提高爬取效率，最后保存 Excel 文件^[9]。网络爬虫流程如图 1 所示。

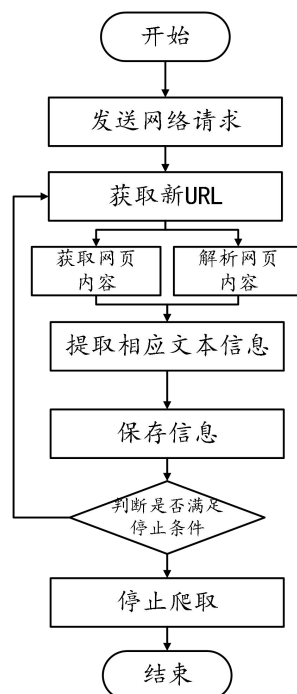


图 1 爬虫流程图

Figure 1 Web crawler flowchart

合法性说明：在数据采集过程中，严格遵循了各数据源网站的使用协议与隐私政策。为了确保合规性，所有的爬虫请求均遵守目标网站的 `robots.txt` 文件规定，并在合理的时间间隔内发送请求，

避免对网站服务器造成过大负担。同时，在爬取数据前，已审查并确保获取的数据不会侵犯网站的版权或知识产权。所有用于研究的数据仅为非商业性用途，且将严格保密处理，确保数据合法使用。

1.3 文本处理

数据集的文本处理是一个复杂而关键的过程，它涉及数据的收集和清洗等多个步骤。以下是对这一过程的详细描述：

数据收集是数据集构建的第一步，这一步骤通过运用网络爬虫工具实现。专注于从旅游评论网站、社交媒体平台、旅游攻略网站等网络资源中，广泛地收集与西北旅游相关的各类文本数据。这些数据涵盖了游客在旅游过程中的亲身体验和感受，包括他们的评论、游记、攻略以及景点介绍等内容。

数据清洗是一个重要的环节，它能够提高数据的质量和分析的准确性。收集到的文本数据中往往混杂着大量的噪声和无关信息，例如广告、链接、特殊字符等，这些噪声信息的存在可能会对分析结果的准确性产生影响。因此，需要进行数据清洗工作，以去除这些噪声和无关信息，从而保留有用的旅游文本数据。

在数据清洗的基础上，还需要进行去停用词处理。在文本中，存在一些高频的停用词，如“的”“和”“是”等。这些词在文本中出现的频率较高，但对于文本的意义贡献较小。因此，可以通过使用停用词表，将这些停用词从文本中去除，从而保留更有用的词语。去停用词处理有助于减少冗余信息，突出文本中的关键信息^[10]。

在处理数据集的过程中，采取了一系列系统化的步骤，以确保数据的质量和可用性。首先，提取了正文内容，以便集中分析所需的信息。接着，将连续的“##”符号替换为换行符，以改善文本的可读性，使数据在后续处理时更具结构性。为了进一步清洗数据，进行了多项处理，包括删除除了中文字符、字母、数字、逗号（“，”）和句号（“。”）之外的所有特殊符号，以确保文本的纯净度，避免影响后续分析。同时，还去除了字母“n”，这一步骤旨在消除不必要的字符干扰，确保文本的整洁^[11]。

在处理行内空格时，选择删除多余的空格，但保留了行的基本结构，这样做的合理性在于，它使得文本格式保持一致，有助于后续的机器学习模型更好地理解文本内容。此外，还跳过了空行，以避免在分析时引入无效数据，这样可以提高数据集的密度和信息量。之后，调用 API 生成每句的 instruction，为模型训练提供了明确的上下文信息。instruction 能够为模型的训练和生成提供明确的指引，instruction 的设计紧密结合数据集的实际内容，例如针对景点介绍数据，生成“请简要介绍某景点的主要特点”“列出该景点的开放时间 & 地址”等提示词；针对美食文化数据，则设计“描述该美食的制作过程及文化背景”等任务型指引。instruction 的主要作用在于引导模型高效学习目标任务，使得模型能够更准确地理解上下文并生成所需的输出。其优点包括：提高模型训练效率，通过明确的任务描述，减少模型在学习阶段的模糊性，加速模型收敛；增强任务适配性，不同 instruction 可以针对分类、生成、问答等多任务场景灵活调整，提高数据的适用范围，支持多样化输出。instruction 通过丰富的指引内容，扩展了数据集在模型训练中的潜在应用场景，提升生成结果的多样性和质量。最后，删除了过短的无用信息，并通过调用 API 完善了回答，如表 1 所示。这样不仅提升了数据的质量，也为后续的模型训练提供了丰富的背景知识^[12]。

表 1 数据处理流程

Table 1 Data processing workflow

序号	示例	处理方法	处理效果
1	那拉提景区 ::Document(url='https://baike.baidu.com/item/%E9%82%A3%E6%8B%89%E6%8F%90%E6%97%85%E6%B8%B8%E	提取正文	那拉提旅游风景区\n\n新疆维吾尔自治区境内AAAAA旅游风景区\n\n##简介\n\n那拉提旅游风景区，位于新疆维吾尔自治区伊犁哈萨克自治州新源县境内，地处天山

序号	示例	处理方法	处理效果
	9%A3%8E%E6%99%AF%E5%8C%BA/10952929', metadata={}, page_content='#那拉提旅游风景区\n\n新疆维吾尔自治区境内 AAAAA 旅游风景区\n\n## 简介\n\n那拉提旅游风景区，位于新疆维吾尔自治区伊犁哈萨克自治州新源县境内，地处天山腹地，伊犁河谷东端。1999 年，那拉提景区成立，总规划面积 1848 平方千米。风景区自南向北由高山草原观光区、哈萨克民俗风情区、旅游生活区组成。		腹地，伊犁河谷东端。1999 年，那拉提景区成立，总规划面积 1848 平方千米。风景区自南向北由高山草原观光区、哈萨克民俗风情区、旅游生活区组成。
2	那拉提旅游风景区\n\n新疆维吾尔自治区境内 AAAAA 旅游风景区\n\n## 简介\n\n那拉提旅游风景区，位于新疆维吾尔自治区伊犁哈萨克自治州新源县境内，地处天山腹地，伊犁河谷东端。1999 年，那拉提景区成立，总规划面积 1848 平方千米。风景区自南向北由高山草原观光区、哈萨克民俗风情区、旅游生活区组成。	将连续的 ## 替换为换行符	那拉提旅游风景区\n\n新疆维吾尔自治区境内 AAAAA 旅游风景区\n\n简介\n\n那拉提旅游风景区，位于新疆维吾尔自治区伊犁哈萨克自治州新源县境内，地处天山腹地，伊犁河谷东端。1999 年，那拉提景区成立，总规划面积 1848 平方千米。风景区自南向北由高山草原观光区、哈萨克民俗风情区、旅游生活区组成。
3	那拉提旅游风景区\n\n新疆维吾尔自治区境内 AAAAA 旅游风景区\n\n简介\n\n那拉提旅游风景区，位于新疆维吾尔自治区伊犁哈萨克自治州新源县境内，地处天山腹地，伊犁河谷东端。1999 年，那拉提景区成立，总规划面积 1848 平方千米。风景区自南向北由高山草原观光区、哈萨克民俗风情区、旅游生活区组成。	1.删除除了中文字符、字母、数字、逗号（，）和句号（。）之外的所有特殊符号 2.# 删除字母 'n' 3. 删除行内的空格，但保留行结构 4. 跳过空行	那拉提旅游风景区新疆维吾尔自治区境内 AAAAA 旅游风景区 简介那拉提旅游风景区，位于新疆维吾尔自治区伊犁哈萨克自治州新源县境内，地处天山腹地，伊犁河谷东端。1999 年，那拉提景区成立，总规划面积 1848 平方千米。风景区自南向北由高山草原观光区、哈萨克民俗风情区、旅游生活区组成。
4	那拉提旅游风景区新疆维吾尔自治区境内 AAAAA 旅游风景区 简介那拉提旅游风景区，位于新疆维吾尔自治区伊犁哈萨克自治州新源县境内，地处天山腹地，伊犁河谷东端。1999 年，那拉提景区成立，总规划面积 1848 平方千米。风景区自南向北由高山草原观光区、哈萨克民俗风情区、旅游生活区组成。	调用 api 生成本句的 instruction	{" instruction ": "那拉提旅游风景区 5A 景区", " output ":"那拉提旅游风景区新疆维吾尔自治区境内 AAAAA 旅游风景区"} {" instruction ": "那拉提旅游风景区", " output ":"简介那拉提旅游风景区，位于新疆维吾尔自治区伊犁哈萨克自治州新源县境内，地处天山腹地，伊犁河谷东端。1999 年，那拉提景区成立，总规划面积 1848 平方千米。风景区自南向北由高山草原观光区、哈萨克民俗风情区、旅游生活区组成。"} -
5	{" instruction ": "景点", " output ":"主要景点"}	删除过短无用信息	-
6	{" instruction ": "那拉提旅游风景区 5A 景区", " output ":"那拉提旅游风景区新疆维吾尔自治区境内 AAAAA 旅游风景区"}	调用 api 完善回答	" instruction ": "那拉提旅游风景区 5A 景区", " output ":"那拉提旅游风景区是新疆维吾尔自治区境内 AAAAA 旅游风景区"}

2 数据样本描述

2.1 数据组成

本数据集专注于西北地区旅游景点的文本信息，旨在为研究和分析提供全面的景点资料来源。数据主要来源于两个平台，通过携程网，获取了关于景点的基本信息，包括地区、简介、地址和开放时间等具体数据，这些信息详见表 2 中的景点四维信息。携程网的数据以实用性为主，便于用户快速了解各个景点的基本情况。

表 2 景点四维信息

Table 2 Four-dimensional information of scenic spots

instruction	output
甘肃兰州有什么景点？	五泉山公园
请介绍五泉山公园的主要特点。	五泉山公园位于兰州市区南侧，是兰州较出名的公园。五泉山的历史很是悠久，因传说霍去病将军曾在山上鞭打五下打出五眼泉水而得名。如今的五泉山上泉水已经很小，但是山间修建了众多的古代建筑和寺庙等，公园内绿树成荫，环境古朴优雅。来此可以登山锻炼、拜佛祈福，还能够登到山顶俯瞰兰州城区的美景。五泉山相对高度并不高，爬山大概 1-2 小时左右，山顶处可以俯瞰兰州城。很多游客来此并不会登顶，在山腰之间也能看到不错的城区景色。公园内大部分的建筑都分布在山腰之间，出名的有浚源寺、嘛呢寺、卧佛寺、地藏寺等，多为明清建筑，香火很旺，可以进入祈福。山下处有儿童游乐设施，可以带着小朋友前来娱乐，这里也是老年人晨练、唱歌、跳舞的地方。在公园旁边，还开辟了一座五泉山动物园，里面有上百种动物，适合带小朋友前来游玩参观。
五泉山公园的地址？	兰州市城关区五泉南路 103 号
五泉山公园的开放时间？	全年 08:00-18:00 开放，具体营业状态以当天开放情况为准。

另一方面，百度百科提供的数据则呈现出更为丰富的层次，涵盖了景点的介绍、历史沿革、特色亮点、文化背景、文物价值及其保护状况以及相关的旅游信息等，详见表 3 的景点详细信息。与携程网相比，百度百科对每个景点的描述更加详尽，能够为游客提供更深刻的背景知识和文化内涵，从而丰富了景点信息的维度。

这两个平台的数据相辅相成，携程网侧重于实用的旅游指导，而百度百科则注重于景点的历史文化脉络与价值观念。通过将这两种信息整合在一起，本数据集不仅能快速获取基本的旅游信息，还能深入理解每个景点的独特魅力与文化意义，从而提升整体的旅游体验。

2.2 样本数据描述

通过这四维信息可以快速的了解景点的基本信息，景点四维信息的优点在于其全面性、便捷性和易于比较，如表 2 所示。游客可以根据不同景点的四维信息进行直接比较，从而选择最符合个人兴趣和需求的景点。详细的景点信息还能增强期待感，在游览时更好地理解 and 欣赏景点的特色与文化。同时，通过结构化的数据呈现，四维信息减少了信息传递中的模糊性，提高了准确性。

景点详细信息的优点在于能够为游客提供全面而深入地理解，在旅行前对目的地有更清晰地认识，如表 3 所示。这些信息不仅涵盖景点的历史背景、文化意义和独特特色，还能提供相关的活动介绍和游览建议，帮助游客更好地规划行程。通过详细描述，能够让游客了解景点的独特之处，从而激发兴趣和期待感。此外，详细信息有助于消除不确定性，增强游览安全感。同时，这些信息的丰富性也能够满足不同游客的需求，提升整体的旅游体验。

表 3 景点详细信息

Table 3 Detailed information on scenic spots

Instruction	Output
兰州景点	兰州黄河铁桥
兰州黄河铁桥介绍	兰州黄河铁桥，又名中山桥，位于甘肃省兰州市白塔山下的滨河路中段。由美国桥梁公司设计、德国泰来洋行承建、中国工匠施工的合作模式建造。清光绪三十四年（1908 年）5 月 9 日，兰州黄河铁桥工程正式开工。清宣统元年（1909 年）8 月 19 日，兰州黄河铁桥竣工通行。民国十七年（1928 年），为纪念孙中山先生始称“中山桥”。 兰州黄河铁桥长 233.5 米，总宽 8.36 米。为四墩五孔，每孔跨径 45.9 米。上部结构为梯形穿式钢桁架，每一跨都是超静定结构组成桁架体系，由竖向钢架划分成五个长方形。 兰州黄河铁桥是中国近代史上整个西北地区第一座引进外国技术建造的桥梁，这一特殊的建设背景及建设年代使兰州黄河铁桥成为研究近代历史的钥匙，在中国的建筑历史上占有独特的地位。兰州黄河铁桥作为一个交通结构工程，体现了中国近代历史上洋务运动时期建筑艺术发展史的风格、流派、特征。同时，建造桥梁所需物资的运输过程也创造了近代运输史上的奇迹。2006 年 5 月 25 日，兰州黄河铁桥被中华人民共和国国务院公布为第六批全国重点文物保护单位。2019 年 4 月 12 日，入选由中国科学技术协会调查部主办，中国科学技术协会创新战略研究院、中国城市规划学会共同承办的“中国工业遗产保护名录（第二批）”。
历史沿革	清光绪三十二年（1906 年）农历五月初，甘肃洋务总局彭英甲等与德国泰来洋行喀佑斯就包修兰州黄河铁桥一事拟定了初步的合作合同，喀佑斯请工程师进行实际勘测，得到勘测结果后，德方认为“黄河水性，虽云湍急，若如所议章程架修铁桥，甘愿保固八十年”。 清光绪三十二年（1906 年）10 月 28 日，甘肃洋务总局与德国泰来洋行正式签订兰州黄河铁桥包修合同，兰州黄河铁桥的各项建设筹备工作全面展开。 清光绪三十四年（1908 年）3 月 23 日，清朝廷的朱批传达到甘肃：“该部知道。钦此。”至此，兰州黄河铁桥的建设工程正式得到国家认可。 清光绪三十四年（1908 年）4 月 2 日，庄浪县举人牛献珠致禀彭英甲，请求停修兰州黄河铁桥。他还列举了修建兰州黄河铁桥在发生战争时的六大弊端。彭英甲对牛献珠的观点予以回击，认为建桥正当时宜，刻不容缓。 清光绪三十四年（1908 年）5 月 9 日，兰州黄河铁桥工程正式开工。
建筑特色	结构：兰州黄河铁桥为四墩五孔，每孔跨径 45.9 米。上部结构为梯形穿式钢桁架，每一跨都是超静定结构组成桁架体系，有竖向钢架划分成五个长方形，再加上斜撑竿件，构成一个框架清晰又有韵律的立面。 特点：兰州黄河铁桥由美国桥梁公司设计、德国泰来洋行承建、中国工匠施工的合作模式建造。初建桥所用料件均由泰来洋行从德国购置。 兰州黄河铁桥设计简洁明了，没有多余冗杂的装饰，它不是中国的传统的建筑类型，而是工业化时代的新材料、新结构、新技术、新工艺以及先进的施工、管理水平在空间构筑物上的形象体现。
主要建筑	兰州黄河铁桥由上部结构（桥身）和下部结构（桥墩）两部分组成。兰州黄河铁桥长 233.5 米，总宽 8.36 米，其中车行道宽 6 米，两边人行道各宽 1 米。 桥墩：兰州黄河铁桥下部结构由 2 座桥台、4 个桥墩组，沉井基础开挖至岩层。桥墩、桥台的材料不一样，南桥台是条石砌的，北桥台是混凝土的。南面第一墩上部是浆砌条石，下部是混凝土。其余三个桥墩上部是混凝土，下部是浆砌疙瘩石。2010 年–2011 年将桥墩顶整体现浇筑混凝土接高 1.2 米。 桥身：兰州黄河铁桥桥身原为穿式钢桁架，共计 5 孔，每孔跨径 45.9 米，全长 233.5 米；每孔 2 片桁架，共计 10 片桁架；每片桁架有 9 个节间，每节间各为 5.1 米。桁架横梁为钢梁，栏杆由角铁及钢管焊接而成。桥身为五座弧形钢架拱梁，桥面为铁板。
历史文化	传说轶事 兰州历来是东西交通要冲，中原与西域往来的必经之途，穿城而过的黄河则是横亘在兰州这一交通枢纽之上的障碍，民间曾有“隔河如隔天，渡河如渡鬼门关”的歌谣。从明洪武五年（1372 年）起，宋国公冯胜、卫国公邓愈都曾因军事需要在兰州黄河段修建过浮桥，但都很

Instruction	Output
	快被拆或被毁了。直到洪武十八年（1385 年），兰州卫指挥佥事杨廉才在如今的兰州黄河铁桥的位置上兴建了镇远浮桥。此后 500 多年间，用以构筑浮桥的船数和用以固定的铁柱与木柱数虽有变化，但镇远浮桥却以其扼守要津的重要地位，被誉为“天下第一桥”。矗立在铁桥南岸的将军铁柱正是镇远浮桥 500 年兴衰史的见证。据《创建兰州黄河铁桥碑记》记载，兰州黄河铁桥从清光绪三十三年（1907 年）二月开始修建，宣统元年（1909 年）七月竣工，历时近两年半，花费白银 30.66 万两。
文物价值	兰州黄河铁桥作为一个交通结构工程，体现了中国近代历史上洋务运动时期建筑艺术发展史的风格、流派、特征。所表现出来的大型结构建筑的力与美的结合，具有较高的艺术表现力、感染力和审美价值。 由美国桥梁公司设计、德国泰来洋行承建、中国工匠施工的合作模式，再加上清末洋务运动的历史条件、政府和民众的积极参与的社会背景，使兰州黄河铁桥在中国的建筑历史上占有独特的地位。兰州黄河铁桥是中国近代史上整个西北地区第一座引进外国技术建造的桥梁，这一特殊的建设背景及建设年代使兰州黄河铁桥变成了研究近代历史的钥匙。同时，桥梁所需物资的运输过程也创造了近代运输史上的奇迹。
文物保护	1989 年，兰州黄河铁桥被列为兰州市市级文物保护单位。 2005 年，兰州黄河铁桥被列为甘肃省省级重点保护文物。 2006 年 5 月 25 日，兰州黄河铁桥被中华人民共和国国务院公布为第六批全国重点文物保护单位。 2019 年 4 月 12 日，兰州黄河铁桥入选由中国科学技术协会调宣部主办，中国科学技术协会创新战略研究院、中国城市规划学会共同承办的“中国工业遗产保护名录（第二批）”。

3 数据质量控制和评估

数据集的数据质量控制与评估是一个至关重要的环节，涵盖了数据爬取、数据清洗和数据标注等多个方面。

在数据爬取阶段，使用网络爬虫工具收集了大量的西北旅游相关文本数据。为确保数据的质量，选择了权威且可信的网站作为数据源，并采用了合适的爬虫策略，以避免对源网站造成不必要的负担，从而实现高效的数据采集。

在数据清洗阶段，对爬取到的文本数据进行了严格的预处理，主要包括去除广告、链接和特殊字符等噪声信息，以确保数据的纯净度和准确性。这一过程不仅提高了数据的可用性，也为后续的分析奠定了坚实的基础。

在将数据处理成标准的文本文件时，采用了自动化程序来对这些文本进行标注，利用 API 为文本添加了 instruction。在标注过程中，结合自动化程序和人工校对，确保 instruction 设计与内容的高度匹配，同时统一将数据调整为 JSON 格式，以便于后续模型加载和解析。标注的质量通过交叉验证法进行评估，即把部分数据分配给多个标注人员进行标注，并通过计算标注的一致性系数来评估标注结果的一致性和准确性。如果一致性系数低于预定标准，则进行重新标注或人工调整，以确保标注的质量。

在数据质量评估阶段，采用了多种可信的技术验证方法，确保数据集的高质量与可靠性。对数据进行一致性和完整性检查，确保所有的文本数据符合预定的格式和标注要求，且与源网站内容保持一致。对于清洗后的数据，利用自然语言处理（NLP）技术进行语义分析，识别潜在的语义不一致和文本中的噪声。使用人工抽样和自动化脚本的结合，对数据的准确性进行验证，确保每条数据与原始内容的匹配度超过 85%。在标注质量的评估中，采用交叉验证法，通过不同标注人员之间的一致性评分来衡量标注的准确性和一致性。对于低一致性数据，进行了额外的人工复核和调整。通过这些技术手段，数据集在准确性、完整性和一致性方面达到高质量标准。

通过对数据爬取、数据清洗、数据标注和数据质量评估等环节的精细控制，成功构建了一个高

质量、可靠的西北五省旅游资源大模型微调数据集。

4 数据价值

本数据集具有重要的数据价值，当前在旅游领域的相关大模型微调数据集相对匮乏，尤其是西北旅游等特定区域。这种匮乏导致了大模型在处理旅游相关任务时的能力受限，无法充分挖掘区域性文化和特色。面向旅游的微调数据集的构建，使模型能够在特定领域进行精准的语义理解和应用，从而提升其在旅游推荐、问答系统和智能客服等实际应用中的表现。

本数据集的科研价值不容忽视。通过对旅游文本的分析，可以揭示出区域旅游的独特特点与潜在价值，促进对旅游文化的深层次理解。同时，研究者可以利用这一数据集进行多种实验，包括文本分类、景点分析、景点推荐等，从而推动相关领域的技术进步。此外，这一数据集的构建将为未来的旅游大模型研究提供重要的基础数据，促进学术界与产业界之间的交流与合作，推动旅游信息技术的创新与发展。

面向旅游大模型的微调数据集不仅能够补充现有资源的不足，还将为科研和实际应用带来重要的推动力，具有广泛的应用前景和深远的社会价值。

5 数据使用方法和建议

本数据集在 Llama3-8b 和 Chatglm4-9b 等模型上取得了良好的微调效果，这表明其数据质量和结构适合进行高效的模型训练。为了进一步提升模型的性能，可以尝试在其他大型语言模型上进行微调，以探索不同模型架构对数据集的适应性和表现。

实验在 4 张 NVIDIA RTX A6000 上完成，深度学习框架为 Pytorch，Cuda 版本为 12.2。实验使用官方量化版 ChatGLM4-9B-Chat、Llama3-8B、Baichuan2-13B-Base 和 Qwen2-7B-Chat 模型，具体实验参数如表 4。

表 4 实验参数.

Table 4 Experimental parameters

参数	值
Python	3.10
微调方法	LoRA
cutoff_len	1024
learning_rate	5e-05
per_device_train_batch_size	2
gradient_accumulation_steps	8
lora_rank	8
lora_alpha	16

采用了 predict_bleu-4、predict_rouge-1、predict_rouge-2 和 predict_rouge-l 作为评价指标。BLEU (Bilingual Evaluation Understudy) 是一种常用的用于评估机器翻译质量的指标。

BLEU-4 表示四元语法 BLEU 分数，它衡量模型生成文本与参考文本之间的 n-gram 匹配程度，其中 n=4。值越高表示生成的文本与参考文本越相似，最大值为 100。ROUGE (Recall-Oriented Understudy for Gisting Evaluation) 是一种用于评估自动摘要和文本生成模型性能的指标。

ROUGE-1 表示一元 ROUGE 分数，ROUGE-2 表示二元 ROUGE 分数，分别衡量模型生成文本与参考文本之间的单个词和双词序列的匹配程度。值越高表示生成的文本与参考文本越相似，最大值为 100。

ROUGE-L 衡量模型生成文本与参考文本之间最长公共子序列 (Longest Common Subsequence) 的匹配程度。值越高表示生成的文本与参考文本越相似，最大值为 100。

从表 5 中可以明显看出, Qwen2 模型在微调前和微调后的性能表现有显著差异。在微调前, Qwen2 模型的各项指标表现中等, 但与其他未微调模型相比略有劣势。例如, 在 BLEU-4 指标上, Qwen2 微调前仅得分 9.5, 低于 GLM4 的 13.8 和百川的 18.7。在 ROUGE-1 和 ROUGE-2 指标上, Qwen2 微调前得分分别为 22.5 和 7.8, 虽然略高于 Llama3-8B 微调前的表现, 但远不及 GLM4 和百川。这说明 Qwen2 在未微调状态下, 其生成文本的质量和旅游文本的匹配度较低, 尤其在捕捉词汇和短语级特征方面表现不足。

表 5 模型对比

Table 5 Model comparison

评价指标	GLM4 微调前	GLM4 微调后	Llama3-8B 微调前	Llama3-8B 微调后	百川微调前	百川微调后	Qwen2 微调前	Qwen2 微调后
predict_bleu-4	13.8	35.4	1.2	32.5	4.1	18.7	9.5	38.2
"predict_rouge-1"	27.1	50.8	5.9	47.8	18.2	35.7	22.5	52.7
"predict_rouge-2"	12.1	35.5	1.5	31.4	4.5	20.4	7.8	36.0
"predict_rouge-l"	21.5	47.3	5.6	44.9	13.1	31.1	17.7	50.8

经过微调后, Qwen2 的性能得到了极大的提升, 成为表格中表现最优的模型。在 BLEU-4 指标上, Qwen2 微调后得分达到了 38.2, 是微调前的 4 倍多, 同时也高于其他所有模型, 表明其在生成与旅游文本一致性上的显著改进。ROUGE-1 和 ROUGE-2 指标分别提升至 52.7 和 36.0, 显示出模型在关键内容词汇覆盖和短语级匹配方面的显著增强。ROUGE-L 指标的提升则表明 Qwen2 在捕捉文本的全局结构和顺序上也有了极大改进。已对优秀突出指标进行加粗, 微调后的 Qwen2 在生成旅游文本的精确性、一致性和语义结构完整性上达到了领先水平, 远超其他模型。

建议在训练过程中进行 1 轮微调, 以便快速评估模型的初步效果。如果需要进行多轮训练, 则应考虑对数据集进行优化, 去除其中较短的数据样本。这是因为较短的数据可能会影响模型的泛化能力, 导致模型在处理长文本或复杂场景时的表现不佳。处理后能够确保模型在微调后的性能更加稳定和可靠, 从而提升其在实际应用中的实用性和准确性。

致 谢

在此特别感谢携程网和百度百科为研究提供的公开数据支持, 其高质量内容为数据集的构建奠定了坚实基础。

数据作者分工职责

卢保青(1998—), 男, 河南安阳人, 硕士研究生, 研究方向为自然语言处理。主要承担工作: 数据采集、数据质量评估、论文撰写。

万福成(1985—), 男, 博士, 教授, 博士生导师, 研究方向为自然语言处理与大数据技术。主要承担工作: 数据质量管理、数据综合控制。

于洪志(1947—), 女, 山东省龙口市人, 教授, 博士生导师, 研究方向为自然语言处理、人工智能、互联网+多语言信息技术。主要承担工作: 数据审核。

陈敏(2001—), 女, 陕西宝鸡人, 硕士研究生, 研究方向为自然语言处理。主要承担工作: 数据采集、数据预处理。

参考文献

- [1] 马宁, 吕文蓉, 郭泽晨. 面向机器阅读理解医学域数据集 MedicalQA[DS/OL]. ScienceData Bank, 2022. (2022-08-08). DOI: 10.11922/sciencedb.j00001.00411. [MA N, LYU W R, GUO Z C. MedicalQA: a dataset of medical domain for machine reading comprehension[DS/OL]. ScienceData Bank, 2022. (2022-08-08). DOI: 10.11922/sciencedb.j00001.00411.]
- [2] 韩立钦, 卢晓彤, 刘慧聪, 等. 2022 年黄河流域旅游资源空间分布数据集[J/OL]. 中国科学数据, 2024, 9(3). (2024-09-30). DOI: 10.11922/11-6035.ncdc.2023.0013.zh. [HAN L Q, LU X T, LIU H C, et al. A dataset of spatial distribution of tourism resources in the Yellow River Basin in 2022[J]. China Scientific Data, 2024, 9(3). (2024-09-30). DOI: 10.11922/11-6035.ncdc.2023.0013.zh.]
- [3] 康建芳, 张耀南, 刘春, 等. 2015–2019 年黄河流域人文经济数据集[J/OL]. 中国科学数据, 2022, 7(4). (2022-09-09). DOI: 10.11922/11-6035.ncdc.2021.0024.zh. [KANG J F, ZHANG Y N, LIU C, et al. A dataset of human culture and economy in the Yellow River Basin during 2015–2019[J]. China Scientific Data, 2022, 7(4). (2022-09-09). DOI: 10.11922/11-6035.ncdc.2021.0024.zh.]
- [4] 朱登赞, 万福成, 国旗, 等. 针对北京旅游景点的评论文本情感标注数据集[J/OL]. 中国科学数据, 2025. (2025-02-24). DOI: 10.11922/11-6035.csd.2022.0026.zh. [ZHU D Y, WAN F C, GUO Q, et al. A dataset of sentiment annotation the review texts for Beijing tourist attractions[J]. China Scientific Data, 2025. (2025-02-24). DOI: 10.11922/11-6035.csd.2022.0026.zh.]
- [5] CATELLI R, BEVILACQUA L, MARINIELLO N, et al. Cross lingual transfer learning for sentiment analysis of Italian TripAdvisor reviews[J]. Expert Systems with Applications, 2022, 209: 118246. DOI: 10.1016/j.eswa.2022.118246.
- [6] ASGHAR N. Yelp dataset challenge: review rating prediction[EB/OL]. (2016-05-17) (2025-01-17). <https://arxiv.org/abs/1605.05362v1>.
- [7] 王双, 万福成, 杨逸民, 等. 中文小说短句序列文本复述数据集[J/OL]. 中国科学数据, 2022. (2022-03-31). DOI: 10.11922/11-6035.csd.2025.0041.zh. [WANG S, WAN F C, YANG Y M, et al. A dataset of Chinese novel short sentence sequence text paraphrase[J/OL]. China Scientific Data, 2022. (2022-03-31). DOI: 10.11922/11-6035.csd.2025.0041.zh.]
- [8] 于娟, 刘强. 主题网络爬虫研究综述[J]. 计算机工程与科学, 2015, 37(2): 231–237. DOI: 10.3969/j.issn.1007-130X.2015.02.007. [YU J, LIU Q. Survey on topic-focused crawlers[J]. Computer Engineering & Science, 2015, 37(2): 231–237. DOI: 10.3969/j.issn.1007-130X.2015.02.007.]
- [9] 李勇, 韩亮. 主题搜索引擎中网络爬虫的搜索策略研究[J]. 计算机工程与科学, 2008, 30(3): 4–6, 56. DOI: 10.3969/j.issn.1007-130X.2008.03.002. [LI Y, HAN L. Research on the search strategy of the web spider in topic-oriented search engines[J]. Computer Engineering & Science, 2008, 30(3): 4–6, 56. DOI: 10.3969/j.issn.1007-130X.2008.03.002.]
- [10] 郭志懋, 周傲英. 数据质量和数据清洗研究综述[J]. 软件学报, 2002, 13(11): 2076–2082. [GUO Z M, ZHOU A Y. Research on data quality and data cleaning: a survey[J]. Journal of Software, 2002, 13(11): 2076–2082.]
- [11] 郝爽, 李国良, 冯建华, 等. 结构化数据清洗技术综述[J]. 清华大学学报(自然科学版), 2018,

58(12): 1037–1050. DOI: 10.16511/j.cnki.qhdxxb.2018.22.053. [HAO S, LI G L, FENG J H, et al. Survey of structured data cleaning methods[J]. Journal of Tsinghua University (Science and Technology), 2018, 58(12): 1037–1050. DOI: 10.16511/j.cnki.qhdxxb.2018.22.053.]

[12] 王曰芬, 章成志, 张蓓蓓, 等. 数据清洗研究综述[J]. 现代图书情报技术, 2007(12): 50–56. [WANG Y F, ZHANG C Z, ZHANG B B, et al. A survey of data cleaning[J]. New Technology of Library and Information Service, 2007(12): 50–56.]

论文引用格式

卢保青, 万福成, 于洪志, 等. 2024 年西北五省旅游资源大模型微调数据集[J/OL]. 中国科学数据, 2025, 10(3). (2025-09-01). DOI: 10.11922/11-6035.csd.2024.0169.zh.

数据引用格式

卢保青, 万福成, 于洪志, 等. 2024 年西北五省旅游资源大模型微调数据集[DS/OL]. V3. Science Data Bank, 2025. (2025-08-31). DOI: 10.57760/sciencedb.j00001.01088.

A data of five elements of tourism resources for large-model fine-tuning in the five northwestern provinces of China in 2024

LU Baoqing^{1,2}, WAN Fucheng^{1,2*}, YU Hongzhi^{1,2}, CHEN Min^{1,2}

1. Key Laboratory of Language and Cultural Computing Ministry of Education, Northwest Minzu University, Lanzhou 730030, P.R. China

2. Gansu Provincial Key Laboratory of Intelligent Information Processing of Ethnic Language and Culture, Northwest Minzu University, Lanzhou 730030, P.R. China

*Email: wanfucheng@126.com

Abstract: With the wide application of large models in various fields, tourism industry faces a growing demand for high-quality datasets to support the enhancement of the models' capability in understanding and generating tourism-related information. This dataset focuses on textual data in the tourism domain and is designed to support fine-tuning tasks for tourism-oriented large models, aiming to enhance their ability to process and generate tourism-related information. The diversity and quality of the dataset are critical to the model's performance. To this end, this study integrates web scraping with manual annotation techniques, along with data cleaning, denoising, and stopword removal, to ensure high data quality and accuracy. Additionally, automated annotation tools are used to generate instructions and perform consistency checks on the texts. The LLM-Tourism dataset draws primarily from Ctrip and Baidu Baike, covering five northwestern provinces in China—Gansu, Ningxia, Qinghai, Shaanxi, and Xinjiang—and comprises 53,280 pairs of structured data in JSON format. The creation of this dataset not only improves the generative accuracy of tourism-focused large models, but also contribute to the sharing and application of large model datasets in the tourism field.

Keywords: textual data; tourism large model; fine-tuning; generation

Dataset Profile

Title	A data of five elements of tourism resources for large-model fine-tuning in the five northwestern provinces of China in 2024
Data corresponding author	WAN Fucheng (wanfucheng@126.com)
Data authors	LU Baoqing, WAN Fucheng, YU Hongzhi, CHEN Min
Time range	January, 2020–September, 2024
Geographical scope	Gansu Province, Ningxia Hui Autonomous Region, Qinghai Province, Shaanxi Province, and Xinjiang Uygur Autonomous Region
Data volume	46.65 MB
Data format	*json
Data service system	https://doi.org/10.57760/sciencedb.j00001.01088
Sources of funding	Gansu Province Basic Research Innovation Group Project (24 JRRA154); Lanzhou Youth Science and Technology Talent Innovation Project (2024-QN-167); Gansu province university graduate student ' innovation star ' (2025CXZX-243).
Dataset composition	The dataset is composed of a text file stored in JSON format, containing 53,280 pairs of structured data samples.