



文献 CSTR:

32001.14.11-6035.csd.2025.0092.zh



文献 DOI:

10.11922/11-6035.csd.2025.0092.zh

数据 DOI:

10.57760/sciencedb.25099

文献分类: 计算机科学技术

收稿日期: 2025-05-19

录用日期: 2026-02-10

发表日期: 2026-02-27

AMCK: 面向大语言模型微调的汽车维护与保养知识数据集

吴文社^{1,2}, 王有齐^{1,2}, 南哲^{1,2}, 马宁^{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: 大语言模型 (LLM) 以其强大的泛化能力, 在汽车维护与保养领域具备极高的应用潜能。本研究针对该领域高质量、结构化微调数据相对稀缺的现状, 构建了面向 LLM 微调的汽车维护与保养知识数据集 AMCK (Automotive Maintenance and Care Knowledge Dataset for Large Language Model Fine-Tuning)。数据集来源于三类: 一是从国内汽车相关网站 (如汽车之家、懂车帝等) 采集真实用户发布的汽车维护咨询、汽车保养场景的问答数据, 经过筛选噪音并过滤语义, 获得有效样本语料; 二是整合汽车工程领域公开的专业语料库 AutoMaster, 包含汽车品牌类型、保养手册、维护工艺标准等知识; 三是使用数据增强方法对数据集进行扩充, 添加更多元化数据。最终筛选出 43962 条高质量样本, 以 JSON 格式存储, 包含指令、输入、输出 3 个属性。AMCK 数据集可以作为 LLM 进行汽车维护保养的微调数据, 其丰富的领域知识样本, 可以驱动 LLM 在汽车维保任务中实现更精准的语义理解和更可靠的决策推理, 提升汽车智能服务水平。

关键词: 大语言模型; 微调; 汽车维护; 汽车保养; 数据集

数据库 (集) 基本信息简介

数据库 (集) 名称	AMCK: 面向大语言模型微调的汽车维护与保养知识数据集
数据通信作者	马宁 (maning@xbmu.edu.cn)
数据作者	吴文社、王有齐、南哲、马宁
数据量	87.9 MB
数据格式	*.json
数据服务系统网址	https://doi.org/10.57760/sciencedb.25099
基金项目	国家自然科学基金 (62466052); 西北民族大学中央高校基本科研业务费专项资金项目 (31920250008)。
数据库 (集) 组成	数据集包括43,962条JSON格式数据, 每一条JSON对象包含3个属性, 分别为instruction、input和output。

引言

* 论文通信作者

马宁: maning@xbmu.edu.cn

大语言模型 (Large Language Model, LLM) 以其强大的语言理解与生成能力, 可广泛应用于诸多领域^[1]。汽车产业智能化、网联化进程正在加速推进, LLM 在

汽车维护保养方面的应用备受关注^[2]。基于大量维修保养案例及专业知识, LLM 能够为维修人员提供准确的故障诊断依据, 为车主制定个性化保养方案, 提升维修效率与保养规范性^[3-4]。

在汽车维护场景中, 模型需掌握部件结构、常见故障和解决方案; 在保养场景中, 模型要理解保养周期、项目优先级以及因车制宜的保养逻辑。但通用大模型在处理这些复杂业务时, 语义理解和推理判断方面能力明显不足。在故障诊断环节, 由于对汽车术语和复杂故障现象理解不准确, 模型可能作出错误判断; 在生成保养建议时, 又可能因无法建立使用年限、部件磨损和维护措施的合理关联, 容易出现推荐缺项或保养过度的情况, 造成这些问题的主要原因在于缺乏适配汽车维修和保养领域的高质量微调数据集。汽车维护与保养的数据专业性强, 而且知识高度集中, 当前该领域的的数据资源非常有限, 严重限制了大模型在汽车相关任务中的表现和落地能力^[5]。

监督微调数据的质量决定了 LLM 模型在特定领域的泛化性能^[6-7]。高质量数据能够引导模型学习领域专属知识结构和语言模式, 使其更好地适用于专业任务场景。本研究聚焦汽车维护与保养知识领域, 构建了大语言模型微调数据集 AMCK (Automotive Maintenance and Care Knowledge Dataset for Large Language Model Fine-Tuning)。通过多元方式收集数据, 去除错漏及敏感数据, 结合行业背景知识选择高质量数据, 并经过多重质检和人工校验。AMCK 数据集可为 LLM 提供丰富、准确且贴合汽车领域实际应用场景的训练数据, 助力模型完成更精准的语义理解以及合理有效的推理判断, 推动汽车智能化服务系统在维护保养场景的安全落地与技术创新。

1 数据采集和处理方法

1.1 数据采集

采集数据包括三部分: 汽车相关公开网站上的真实用户语料、汽车工程领域公开的专业语料库语料以及采用数据增强技术生成的语料^[8-9]。采集过程如图 1 所示。

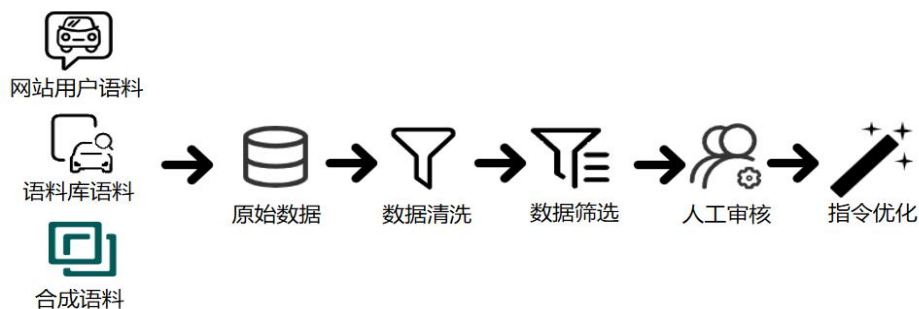


图 1 数据采集流程图

Figure 1 Data collection flowchart

1.1.1 汽车网站真实用户语料

从汽车之家、懂车帝、易车网和瓜子二手车等汽车相关网站收集数据, 先使用爬虫程序初筛, 而后结合人工细选方式。重点收集汽车维护咨询数据, 如故障发生时间、频率、表现以及车辆品牌、型号、使用年限等信息; 同时收集保养场景反馈数据, 包括保养周期咨询、维护项目建议和车辆保值等问题。数据格式如表 1 所示。

表 1 汽车网站数据示例

Table 1 Examples of automotive website data

数据类别	收集内容	示例
新能源维护咨询数据	故障发生时间、频率	故障时间：购买 3 个月出现这种情况。 故障频率：每次使用直流快充桩时必须现。
	表现特征	表现特征：快充至 80%时自动跳枪，伴有电流异响。
	车辆基础信息	品牌型号：比亚迪汉 EV 2023 款 创世版。 使用年限：行驶 3.5 万公里。 保养记录：曾更换过充电口锁止机构。
燃油车维护咨询数据	故障发生时间、频率	故障时间：购买 3 个月出现这种情况。 故障频率：每次开车外出都会出现。
	表现特征	表现特征：刹车时方向盘剧烈抖动。
	车辆基础信息	品牌型号：大众帕萨特 2022 款低配。 使用年限：行驶 1 年/2 万公里。 保养记录：一年内没进行任何维修。
保养场景反馈	保养周期咨询	混动车型长期亏电行驶，发动机机油需要提前更换吗？
	维护项目建议需求	6 万公里大保养除了换油液，电池冷却液是否必须原厂？
	老化部件问题反馈	隐藏式门把手冬季弹出困难，是否需要涂抹润滑脂？

1.1.2 整合汽车领域公开语料库

AutoMaster 是汽车领域的公开数据集，包括汽车品牌、故障出现原因和保养规范等汽车维护与保养相关内容，有 112943 条数据样本。数据以带有交互标签的原始对话文本形式存在，包含大量无信息量的寒暄语句和冗余角色标签，缺乏大模型微调需要的标准结构。语料库知识主要聚焦于传统燃油车领域，缺乏近年来普及的新能源汽车及相关的维修知识。现有的大模型微调数据集的存储格式规范，但其所包含汽车专业知识量较低，难以支撑复杂的故障诊断推理任务。AMCK 数据集在整合 AutoMaster 核心知识的基础上，针对结构化、时效性及领域深度进行了全面优化。AMCK 数据集与原始 AutoMaster 数据及通用指令数据集特性对比如表 2 所示。

表 2 数据集对比

Table 2 Dataset comparison

数据集内容	AutoMaster	Alpaca-zh	AMCK 数据集
数据形态	含有冗余标签的原始对话文本	通用问答/指令跟随	结构化指令微调数据
领域深度	垂直但碎片化	浅层（常识）	深层（覆盖故障诊断逻辑）
信噪比	低（大量重复无关字符）	高	高（经过去噪清洗，聚合核心知识）
知识时效性	滞后（侧重传统燃油车）	一般（受限于预训练）	实时（新能源三电系统等知识）
适用任务	信息检索	通用基准训练	大模型垂直领域微调与知识注入

为提高 AMCK 数据集的专业性和适配能力，本研究从中筛选出关键信息，如车型型号、保养建议、故障出现原因及处理方案，并采用规则和模型相结合的双重策略提取出关键知识。利用正则表

达式对原始文本中的车型年款、行驶里程、故障代码等具有固定格式的实体信息进行精准匹配和提取；调用 GPT-4o 大语言模型，重组结构化问答样本。设计提示模板来引导模型生成与车主问题相匹配的解决方案，并将公开语料库中的专业建议引入其中，确保输出内容的准确性。数据样本处理前后的对比如表 3 所示。

表 3 数据样本处理示例

Table 3 Examples of data sample processing

数据属性	原始数据样本	处理后数据样本
来源信息	QID: Q1 车型: 大众(进口) 高尔夫/帕萨特	(保留核心语义, 无额外数据字段)
输入部分	Question: 我的帕萨特特别的烧油怎么办怎么办?	Instruction: 以下为汽车品牌 and 车型信息: 大众(进口) 高尔夫/帕萨特, 请回答相关问题:
		Input: 我的帕萨特烧机油怎么办?
输出部分	Dialogue: 技师说: 你好...如果在保修期间...;技师说: 你好; 车主说: 嗯	Output: 针对烧机油问题, 建议首先检查车辆行驶里程。如果在保修期内, 请前往 4S 店检查。若已出保, 建议检查发动机活塞环间隙是否过大或气门油封是否老化。若每 7500 公里消耗一升机油属允许范围, 建议车内常备机油补充; 若消耗量超过两升, 建议立即进行发动机检查维修。

1.1.3 数据增强语料

采用问题和答案对的双向扩展方法构建领域生成框架, 依据汽车网站用户问题及保养手册、维修工艺标准等专业语料进行知识抽取, 利用提示工程给 GPT-4o、Chat-GLM 等基座模型输入行业模板^[10]。为保证生成内容的规范性和完整性, 提示模板包含角色设定、背景约束和输出要求等因素, 具体提示模板如图 2 所示。

你是一位资深汽车维修工程师, 精通各类汽车品牌的维修手册与故障诊断逻辑。请根据车型信息和工况情况等等, 再结合“故障现象、原因分析、解决方案”的逻辑链, 优化输入文本并分析工况并给出专业建议。1.必须包含至少 3 个具体的排查步骤; 2.尽量别使用模棱两可的词汇; 3.输出格式需严格遵循 JSON 结构 instruction、input 和 output。

图 2 提示模板示例

Figure 2 Examples of a prompt template

在此基础之上得到原始问题和答案对, 通过多维度语义扩展技术生成变体数据, 采用术语同义替换、中英双向回译实现语义相同表达的多样化; 构建结构化问题模板, 对模板变量如车型、保养周期、工况条件等进行跨品牌、跨工况组合变异, 如将“丰田卡罗拉低速抖动”扩展为“本田思域高速行驶时轮胎保养建议”。经上述方法生成的合成语料, 经领域知识匹配得分与逻辑连贯性双重校验, 再与真实用户语料、专业萃取知识进行融合, 进一步提高数据规模。具体增强示例见表 4。

1.2 数据处理

数据采集完成后, 对获取的多源数据进行清洗和结构化处理。在清洗阶段, 通过规则引擎和敏感词过滤方法, 去除包含违规词、乱码及语义缺失的噪音文本, 对缺少关键信息的样本进行人工补全或删除^[11-12]。

表 4 数据增强示例

Table 4 Examples of data augmentation

类型	原始数据	增强后的数据
Instruction	以下是汽车品牌 and 车型信息：宝马 宝马 X1（进口），请回答相关问题：	以下是汽车品牌 and 车型信息：宝马 宝马 X1（进口），请回答相关问题：
Input	2010 款宝马 X1，2011 年出厂，2.0 排量，通用 6L45 变速箱，原地换挡位 PRND 车辆闯动，行驶升降档正常，4 轮离地换挡无冲击感，更换变速箱油 12L 无改变。试过一辆 2014 年进口 X1 原地换挡位也有冲击感，这是什么情况，哪里的问题。	我的车是 2010 款宝马 X1，2011 年出厂，搭载 2.0 排量发动机，匹配的是通用 6L45 变速器。在 P-R-N-D 几个档位切换时，车辆静止状态下会产生顿挫感，但行驶过程中升档和降档都很平稳。如果将车辆架起四轮离地，再进行换挡则感觉不到冲击。更换过 12 升变速器油但没有改善。另外，2014 年款的进口 X1 也存在类似情况，请问这种问题通常是什么原因导致的？
Output	根据您的描述，2010 款宝马 X1 在原地切换档位时出现这种情况，但是在行驶过程中换挡正常、四轮离地测试无异常，换了变速箱油后还是解决不了。可能原因：变速箱阀体、电磁阀控制异常、换挡机构磨损、控制软件未更新、变速箱油型号不匹配和发动机支撑脚老化引发震动。建议：使用诊断设备读取故障码，检查变速箱阀体和换挡部件工作状态，确认油液型号，更新变速箱程序，并排查发动机悬置系统。也可以去问问其他的宝马车主提供下宝贵意见或者去专业的修车厂。	根据您的描述，您的爱车 2010 款宝马 X1 在原地切换档位时出现了顿挫感这种情况，但是在行驶过程中换挡是正常的，四轮离地测试也没有异常，换了变速器油后还是没有根除这种情况。可能存在以下的原因：变速器阀体、电磁阀控制异常、换挡机构磨损、控制软件未更新、变速器油型号不匹配和发动机支撑脚老化引发震动。建议：可以使用诊断设备读取故障码，检查变速器阀体和换挡部件工作状态是否良好，确认油液型号，检查变速器程序是否已更新，并排查发动机的悬置系统。您也可以去问问其他的同款车辆的车主来获得宝贵意见或者去专业的修理厂去维修车辆。

在结构化处理阶段，通过程序化方式提取汽车品牌、车型等关键信息并将其嵌入 instruction 字段，构建指令、输入、输出的标准化结构。所有样本转换为 JSON 格式，经三轮人工审核，确保数据的规范性和准确性。第一轮为格式校验，主要检查数据文本是否完整、字符段落是否统一；第二轮为知识验证，核对问题和答案之间逻辑一致性和专业合理性；第三轮为标签一致性检查，对 instruction、input 和 output 这三部分内容进行对比检查，保证问题和回答内容对应。经三轮质检后形成结构规范、语义准确的高质量样本集，为模型微调提供准确可靠的数据支撑^[13-14]。

2 数据样本描述

AMCK 数据集大小为 87.9 MB，包含 43962 条数据，以 JSON 格式存储。对样本进行数据类别统计，如表 5 所示。

数据包含 instruction（指令）、input（输入）和 output（输出）3 个属性：instruction 描述汽车品牌和车型相关信息；input 是车主咨询的车辆疑难问题；output 是模型根据指令和输入内容生成的回答或解决方案。数据格式见表 6。

表 5 数据集数据类别分布统计

Table 5 Dataset category distribution

一级分类	二级分类	样本数量	占比
动力系统	电池故障、发动机抖动、电机异常、动力中断、故障灯亮	18,916	43.03%
底盘传动	部件磨损、底盘异响、行驶震动、跑偏、漏油	5,416	12.32%
车身电器	车机黑屏、功能失灵、空调异常、车内异味	4,306	9.79%
日常保养	维保费用、机油更换、保养周期、火花塞/防冻液检查	15,324	34.86%
合计	——	43,962	100.00%

表 6 数据集样本示例

Table 6 Sample examples in the dataset

属性	属性说明	示例内容
instruction:	描述汽车品牌和车型相关信息	以下为汽车品牌和车型信息：本田 飞度，请回答相关问题：
input:	车主咨询的车辆疑难问题	发动机工作状态踩离合感觉有点震脚有沙子似的，是不是分离轴承的问题。车速达到 120 左右感觉车抖，是不是半轴的问题？
output:	模型根据指令和输入内容生成的回答或解决方案	根据您的描述情况，对您提出的问题作以下分析：发动机工作状态下踩离合感觉有点震脚有沙子摩擦似的的声音，可能的原因是：分离轴承的问题。分离轴承磨损或者损坏的时候踩离合会发出异响或者就像沙子摩擦一样的感觉。也有可能是离合器片的问题，离合器片磨损不均匀或者表面有杂物的话，同样会导致踩离合感觉不舒服。还有一种可能是飞轮的问题，飞轮的表面如果有问题也会造成类似的情况。需要检查下分离轴承是否正常。离合器片和飞轮的表面怎么样了？要先确定好了再去专业的汽车修理店里面去询问才行。当车速达到 120 左右时感觉车抖动。可能原因：半轴问题：半轴不平衡或损坏会引起高速行驶时车辆抖动。悬挂系统问题：悬挂系统中的一些零部件（减震器、悬挂臂）磨损、损坏也会造成抖动。轮辋问题：如果轮辋变形或者是不平衡的话也是会在高速行驶的时候出现抖动的现象。建议：检查半轴是否完好无损，或者是否达到了规定的限值。检查轮胎是否平衡，是否有损、气压如何。检查悬挂系统的各个零件是否正常。检查轮辋是否有变形或是不平衡。如果故障相对严重就建议去专业的汽修店进行全面的检查了。踩离合时会有微微震动：分离轴承、离合器片、飞轮问题都会引起该现象，建议重点关注这几个部位。---高速抖动---半轴问题有可能是其中的原因之一，同时还应注意轮胎、悬挂系统、轮辋的因素。为了安全起见和保证性能，请及时前往专业汽车修理店做详细检查并修理。

3 数据质量控制和评估

3.1 数据质量控制

采用领域语义匹配筛选法对样本的语义契合度与逻辑连贯性进行量化评估。基于 BERT 架构预训练模型，利用 AutoMaster 语料进行对比学习微调，构建领域增强编码器，并将标准维修手册等知识映射为基准向量 v_K ，从而建立领域知识基准库，将待评估数据映射到该共享语义空间，计算领域知识匹配得分（DKMS）和逻辑链关联度（LCA）两项核心指标。领域知识匹配得分计算公式如下：

$$DKMS(T,K)=\frac{v_T \cdot v_K}{|v_T| \cdot |v_K|} \quad (1)$$

其中， v_T 为数据文本的语义嵌入向量，用于捕捉“变速器异响”等专业表述的深层语义，设定 $DKMS \geq 0.7$ 以确保文本符合行业术语规范。阈值的设定基于小样本先验实验确定，随机抽取 200 条样本进行人工标注，测试不同阈值下的准确率与召回率。当 $DKMS < 0.7$ 时，样本中混入了大量非汽车领域的闲聊或噪声数据，因此将阈值设定为 0.7，确保高信噪比并保留语料丰富度。逻辑链关联度计算公式如下：

$$LCA=\frac{Sim(P,C)+Sim(C,S)+Sim(P,S)}{3} \quad (2)$$

利用余弦相似度 Sim 评估（故障现象（P），原因分析（C），解决方案（S））三元组逻辑关联。设定 $LCA \geq 0.6$ ，避免因果断裂情况的发生。故障现象描述（P）与最终解决方案（S）之间往往存在较大的语义跨度，直接相似度通常低于相邻节点。逻辑完整的有效样本 LCA 分布峰值集中在 0.65 左右。阈值设定过高，将导致大量包含复杂推理步骤但逻辑正确的长链路样本被误删；若过低则无法识别逻辑断层。因此选定 0.6 作为保留有效逻辑链的下界。

利用领域增强的大模型进行数据编码，并建立领域知识基准库，保留同时满足 $DKMS \geq 0.7$ 和 $LCA \geq 0.6$ 的样本，去除语义模糊和逻辑缺失的数据，通过对数据语义空间映射、语义逻辑校验的方式提高数据质量。将同义词“引擎/发动机”“保养周期/维护里程”等进行统一，以提高模型解析准确率、维修方案逻辑完整性以及自动化筛选率等性能^[15]。

3.2 数据评估

3.2.1 数据长度分布评估

数据评估采用自动筛选结合人工校验的方法，从专业性、逻辑性、政策上和语言上进行检验和筛选。数据集中输入文本长度在 0~100 个字符之间，平均 53.6 个字符；输出文本长度在 400~1000 个字符之间，平均 688.4 个字符。分布情况如图 3 所示。

3.2.2 实验评估及结果分析

为验证 AMCK 数据集的质量，本研究分别在 Qwen1.5-7B-Chat 模型、Qwen2-7B-Chat 模型和 Llama3.1-8B-Chat 模型上进行了训练前后的对比实验，数据集按照 6:2:2 的比例分为训练集、验证集和测试集。采用 BLEU-4、ROUGE-1、ROUGE-2 和 ROUGE-L 作为评估指标，用于衡量模型生成文本时的精确度、召回率和流畅度。实验结果如表 7 所示。

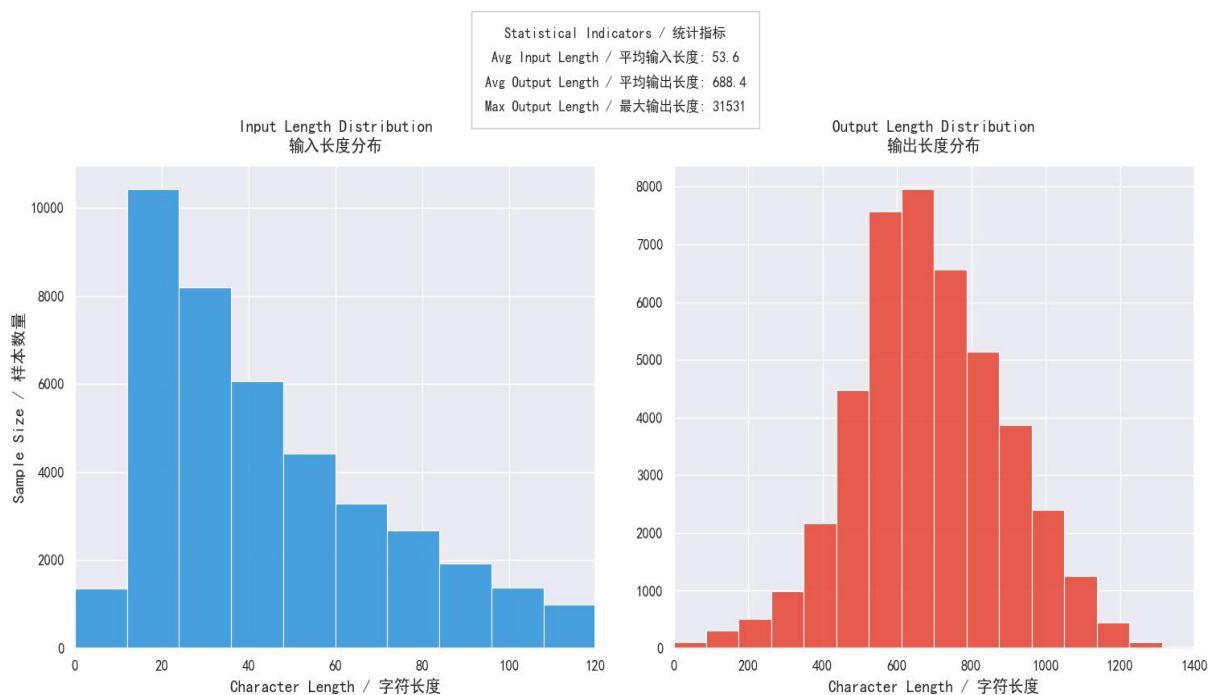


图3 属性长度分布

Figure 3 Distribution of attribute lengths

表7 模型训练前后评估

Table 7 Evaluation of model performance before and after training

评估指标	Qwen1.5-7B-Chat 训练前	Qwen1.5-7B-Chat 训练后	Qwen2-7B-Chat 训练前	Qwen2-7B-Chat 训练后	Llama3.1-8B-Chat 训练前	Llama3.1-8B-Chat 训练后
BLEU-4	14.58	29.22	22.78	29.41	8.08	30.04
ROUGE-1	44.06	50.41	47.42	50.77	30.31	49.95
ROUGE-2	15.00	23.76	18.42	23.95	9.40	23.51
ROUGE-L	25.88	34.61	30.68	34.74	19.30	34.60

观察实验结果，以 Qwen1.5-7B-Chat 为例，训练后的 BLEU-4 值提升到 29.22，ROUGE-1 值提升到 50.41，ROUGE-2 值提升到 23.76，ROUGE-L 值提升到 34.61。对于 Qwen2-7B-Chat 模型和 Llama3.1-8B-Chat 模型，模型微调后的结果也得到了较大性能提升。其中，Llama3.1-8B-Chat 模型提升的幅度相对更为明显，原因是该模型在预训练阶段以使用英文语料为主。

4 数据价值

本数据集是面向汽车维护和保养知识领域的大语言模型微调数据集，围绕汽车产业智能化需求，可作为汽车维保领域模型的训练资源。数据包含红旗、长安、比亚迪等新能源品牌，以及大众、宝马等合资品牌；知识维度包括动力系统、底盘悬挂、车身附件及日常保养等多类常见难题。对于汽车维护故障诊断任务，可以通过基于实例的微调策略进行调试，在数据集中有充分的包括故障现象、原因分析、解决方案的逻辑链路样本来引导模型学习故障诊断的推理模式；对于生成保养方案任务，

可以使用基于提示的微调策略，在 `instruction` 属性中向模型指示给定信息，规范模型以给定方式生成适合用户的个性化保养方案。

5 AIGC 使用说明

在本研究的准备过程中，作者使用了人工智能生成内容（AIGC）工具。具体情况如下：

工具与版本：我们使用了 GPT-4o 模型、Chat-GLM 模型和 Qwen 模型。

数据预处理：利用 GPT-4o、Chat-GLM 和 Qwen 模型辅助编写了用于数据清理和特征工程的 Python 代码。作者对生成的代码进行了审查、调试和优化，确保其逻辑正确。

论文撰写：使用 GPT-4o 对初稿进行了文字润色和措辞建议。所有生成文本均经过了作者的修改、事实核查和确认。

责任声明：作者仔细审阅和编辑了所有 AIGC 辅助生成的内容，并对论文的整体内容、数据真实性及学术诚信负全部责任。AIGC 工具未被列为作者。

致 谢

本数据集在构建过程中使用了 AutoMaster 数据集的开源语料，以及从懂车帝、汽车之家、易车网和瓜子二手车等相关平台获取的真实用户文本数据，特此表示感谢。使用了 Chat-GLM、Qwen 和 GPT-4o 等大语言模型进行数据合成，同时感谢相关团队提供的开放技术支持。

数据作者分工职责

吴文社（2001—），男，山东菏泽人，硕士研究生，研究方向为自然语言处理。主要承担工作：数据收集、数据预处理、质量审核、数据集构建、论文撰写。

王有齐（2001—），男，河南新乡人，硕士研究生，研究方向为自然语言处理。主要承担工作：数据收集、数据校对、质量审核。

南哲（2002—），男，河南平顶山人，硕士研究生，研究方向为自然语言处理。主要承担工作：数据收集、质量审核。

马宁（1981—），男，宁夏吴忠人，博士，教授，研究方向为自然语言处理。主要承担工作：数据集构建指导、数据质量控制、论文写作指导。

参考文献

- [1] 车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战, 机遇与发展[J]. 中国科学: 信息科学, 2023, 53(9): 1645-1687. DOI: 10.1360/SSI-2023-0113. [CHE W X, DOU Z C, FENG Y S, et al. Towards a comprehensive understanding of the impact of large language models on natural language processing: challenges, opportunities and future directions[J]. Scientia Sinica(Informationis), 2023, 53(9): 1645–1687.]
- [2] ARENA F, COLLOTTA M, LUCA L, et al. Predictive maintenance in the automotive sector: a literature review[J]. Mathematical and Computational Applications, 2022, 27(1): 2. DOI:

- 10.3390/mca27010002.
- [3] NAGY J, LAKATOS I. Predictive maintenance and predictive repair of road vehicles—opportunities, limitations and practical applications[J]. *Engineering Proceedings*, 2024, 79(1): 27. DOI: 10.3390/engproc2024079027.
- [4] ZHANG H, ZHAO Y, SUN B, et al. Large language model based intelligent fault information retrieval system for new energy vehicles[J]. *Applied Sciences*, 2025, 15(7): 4034. DOI: 10.3390/app15074034.
- [5] 朱冰, 汤瑞, 赵健, 等. 基于大语言模型的智能汽车仿真测试[J]. *汽车工程*, 2025, 47(4): 587-597. DOI: 10.19562/j.chinasae.qcgc.2025.04.001. [ZHU B, TANG R, ZHAO J, et al. Virtual Simulation Testing Method for Intelligent Vehicle Based on Large Language Model[J]. *Automotive Engineering*, 2025, 47(4): 587-597. DOI: 10.19562/j.chinasae.qcgc.2025.04.001.]
- [6] ANISUZZAMAN D M, MALINS J G, FRIEDMAN P A, et al. Fine-tuning large language models for specialized use cases[J]. *Mayo Clinic Proceedings: Digital Health*, 2024, 3(1): 100184. DOI: 10.1016/j.mcpdig.2024.11.005.
- [7] LU W, LUU R K, BUEHLER M J. Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities[J]. *npj Computational Materials*, 2025, 11: 84. DOI: 10.1038/s41524-025-01564-y.
- [8] MEDVEDIEVA K, TOSI T, BARBIERATO E, et al. Balancing the scale: data augmentation techniques for improved supervised learning in cyberattack detection[J]. *Engineering*, 2024, 5(3): 2170-2205. DOI: 10.3390/eng5030114.
- [9] CHEN H, DAN L, LU Y, et al. An improved data augmentation approach and its application in medical named entity recognition[J]. *BMC Medical Informatics and Decision Making*, 2024, 24: 221. DOI: 10.1186/s12911-024-02624-x.
- [10] CHEN J, ZHU Z, LI H, et al. A data augmentation model integrating supervised and unsupervised learning for recommendation[J]. *Scientific Reports*, 2025, 15: 4862. DOI: 10.1038/s41598-025-88858-9.
- [11] 黄硕, 闫晓东, 田金超. TN-SUM:藏文文本摘要数据集[J/OL]. *中国科学数据*, 2024, 9(4):59-67. DOI:10.11922/11-6035.csd.2024.0004.zh. [HUANG S, YAN X D, TIAN J C. A dataset of Tibetan text summaries—TN-SUM[J/OL]. *China Scientific Data*, 2024, 9(4): 59–67. DOI:10.11922/11-6035.csd.2024.0004.zh.]
- [12] 张雪英, 孙欣辰, 冯敏萱, 等. 长江流域唐宋诗词酒文化地理数据集 [J/OL]. *中国科学数据*, 2025, 10 (4). DOI:10.11922/11-6035.csd.2024.0191.zh. [ZHANG X Y, SUN X C, FENG M X, et al. A dataset of wine culture geography in Tang and Song poems of the Yangtze River Basin [J/OL]. *China Scientific Data*, 2025, 10 (4). DOI:10.11922/11-6035.csd.2024.0191.zh.]
- [13] 庞仙, 陈波, 赵小兵, 等. 基于新闻文本的汉藏新词语数据集(2017–2022)[J/OL]. *中国科学数据*, 2024, 9(4): 30–38. DOI:10.11922/11-6035.csd.2024.0001.zh. [PANG X, CHEN B, ZHAO X B, et al. A dataset of Chinese-Tibetan new words based on news texts(2017–2022)[J/OL]. *China Scientific Data*, 2024, 9(4): 30–38. DOI:10.11922/11-6035.csd.2024.0001.zh.]
- [14] 朝木尔力格, 王斯日古楞. 汉蒙双语法律领域问答语料数据集[J/OL]. *中国科学数据*, 2024, 9(4). DOI: 10.11922/11-6035.csd.2024.0031.zh. [AOMUERLIGE C, WANG S. A dataset of the Chinese-Mongolian bilingual question-answer corpus in the legal field[J/OL]. *China Scientific Data*,

2024, 9(4). DOI: 10.11922/11-6035.csd.2024.0031.zh.]

- [15] 魏崑, 丁香香, 郭梦星, 等. 文本相似度计算方法综述[J]. 计算机工程, 2024, 50(9): 18-32. DOI: 10.19678/j.issn.1000-3428.0068086. [WEI W, DING X X, GUO M X, et al. Review of Text Similarity Calculation Methods[J]. Computer Engineering, 2024, 50(9): 18-32. DOI: 10.19678/j.issn.1000-3428.0068086.]

论文引用格式

吴文社, 王有齐, 南哲, 等. AMCK: 面向大语言模型微调的汽车维护与保养知识数据集[J/OL]. 中国科学数据, 2026, 11(1). (2026-02-27). DOI: 10.11922/11-6035.csd.2025.0092.zh.

数据引用格式

吴文社, 王有齐, 南哲, 等. AMCK: 面向大语言模型微调的汽车维护与保养知识数据集[DS/OL]. V4. Science Data Bank, 2026. (2026-02-27). DOI:10.57760/sciencedb.25099.

AMCK: A dataset of automotive maintenance and care knowledge for large language model fine-tuning

WU Wenshe^{1,2}, WANG Youqi^{1,2}, NAN Zhe^{1,2}, MA Ning^{1,2*}

1. Key Laboratory of Linguistic and Cultural Computing Ministry of Education, Northwest Minzu University, Lanzhou 730000, P.R. China

2. Key Laboratory of Minzu Languages and Cultures Intelligent Information Processing, Northwest Minzu University, Gansu Province, Lanzhou 730000, P.R. China

*Email: maning@xbmu.edu.cn (MA Ning)

Abstract: Large language models (LLMs) have considerable potential for application in the field of automotive maintenance and servicing thanks to their powerful generalization capabilities. This study addresses the current shortage of high-quality structured fine-tuning data in this field by constructing the AMCK (Automotive Maintenance and Care Knowledge Dataset) for fine-tuning large language models. The data sources are classified into three types. First, automotive maintenance information requests posted by real users and question-and-answer data related to automotive maintenance scenarios were collected from domestic automotive websites (such as the AutoHome and the DongCheDi); noise was filtered out and semantic content was refined to ensure the validity of the samples. Second, the publicly available AutoMaster professional corpus in the field of automotive engineering was integrated, which includes knowledge on car brand types, maintenance manuals, maintenance process standards. Third, data augmentation techniques were used to expand the dataset, incorporating more diverse data. In the end, 43,962 high-quality samples were selected and stored in JSON format, comprising three attributes: instruction, input, and output. The AMCK dataset can be used for fine-tuning LLMs in automotive maintenance and repair tasks. Its rich domain-specific knowledge samples enables LLMs to achieve more

accurate semantic understanding and more reliable decision-making reasoning in automotive services, thereby enhancing the level of intelligent automotive services.

Keywords: large language models; fine-tuned; automotive maintenance; automotive care; dataset

Dataset Profile

Title	AMCK: A dataset of automotive maintenance and care knowledge for large language model fine-tuning
Data corresponding author	MA Ning(maning@xbmu.edu.cn)
Data author(s)	WU Wenshe, WANG Youqi, NAN Zhe, Ma Ning
Data volume	87.9 MB
Data format	*.json
Data service system	https://doi.org/10.57760/sciencedb.25099
Source of funding	National Natural Science Foundation (62466052); Fundamental Research Funds for the Central Universities of Northwest Minzu University (31920250008).
Dataset composition	The dataset includes 43,962 data entries in JSON format, with each JSON object containing three attributes: instruction, input, and output.