LinYun：林业领域智能问答与决策支持的大型语言模型研究

赵维康[1,2]，王浩宇[1]，华净[1]，刘耀兵[3]，王秀娟[1]，康孟珍[1,2,4]

（1. 中国科学院自动化研究所多模态人工智能系统全国重点实验室，北京 100190；
2. 中国科学院大学人工智能学院，北京 100049；
3. 北京江恒数智生态科技有限公司，北京 102628；
4. 澳门科技大学创新工程学院，澳门 999078）

摘　要：林业知识具有高度的专业性，涵盖范围广泛，并对法规及实践标准高度敏感。针对目前通用大语言模型在林业场景中存在的知识缺口、术语歧义以及事实性错误等问题，提出“数据合成-模型训练-系统化评测”一体化框架，基于通用基座模型进行领域指令微调，得到面向林业领域适配的模型林云（LinYun）。实验结果表明，LinYun在林业相关任务上显著优于同规模的通用模型，在部分任务上接近甚至超过更大规模模型的表现。

关键词：林业大模型；指令微调；数据合成；LoRA；人工智能
中图分类号：TP391.4
文献标志码：A
doi: 10.11959/j.issn.2096-6652.202603

LinYun: a domain-specific large language model for intelligent question answering and decision support in forestry

Zhao Weikang[1,2], Wang Haoyu[1], Hua Jing[1], Liu Yaobing[3], Wang Xiujuan[1], Kang Mengzhen[1,2,4]

1. The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
2. The School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
3. Jiangheng Digital Intelligence Ecological Technology Co., Ltd., Beijing 102628, China
4. Faculty of Innovation Engineering, Macau University of Science and Technology, Macao 999078, China

Abstract: Forestry knowledge is highly specialized and broad in scopeand is particularly sensitive to regulations and practical standards. To address the knowledge gaps, terminological ambiguities, and factual inaccuracies that general-purpose large language model (LLM) exhibit in forestry scenarios, an integrated framework of “data synthesis-model training-systematic evaluation” was proposed. Based on a general base model, domain-specific instruction fine-tuning was conducted to obtain LinYun, a domain-adapted model for forestry. Experimental results demonstrate that LinYun significantly outperforms general-purpose models of comparable scale in forestry-related tasks, and in some tasks approaches or even surpasses the performance of much larger models.

Key words: forestry large language model, instruction fine-tuning, data synthesis, LoRA, artificial intelligence

0　引言

近年来，大语言模型（large language model，LLM）在多个领域表现出卓越的语言理解与生成能力，并已应用于医学[1-3]、法律[4-5]等专业知识密集型领域。然而，将通用大模型直接应用于林业任

收稿日期：2025-09-30；**修回日期：**2026-02-02
通信作者：康孟珍，mengzhen.kang@ia.ac.cn
基金项目：国家自然科学基金项目（No.62076239）
Foundation Item: The National Natural Science Foundation of China (No.62076239)

务时，仍存在知识覆盖不足、术语歧义突出以及事实性与推理能力欠佳等问题。因此，面向特定领域的模型适配方法成为重要技术路径，即通过持续预训练或指令微调，使模型更好地适配特定领域任务。相关研究已验证了该策略的有效性[6]。

从技术实现看，上述策略在实现上主要包括3类方法：域内持续预训练（continued pretraining）、指令微调（supervised fine-tuning，SFT）、强化学习（如直接偏好优化（direct preference optimization，DPO[7]））。在生态与应急管理等领域，基于大模型的应用探索已取得一定进展[8-10]。例如，Xie等[11]提出的WildfireGPT通过整合气候预测数据与专业文献，为野火风险管理提供高相关度、高精度的辅助决策支持；Du等[12]提出的Tree-GPT融合了图像分割模型（segment anything model，SAM）、领域知识库与LLM工具链，实现了遥感影像理解、结构化参数提取及自动化代码生成，显著提升了森林遥感数据的处理效率。

尽管已有上述进展，但林业领域中基于LLM的研究仍较为有限。Tan等[13]提出的ForestryBERT对林业文本进行持续预训练，涵盖20余万篇林业相关文本，展示了域内预训练的可行性。Sun等[14]开发的ForPKG-1.0将林业政策构建为知识图谱，用于辅助检索增强生成（retrieval-augmented generation，RAG）类LLM，表明林业领域知识库与LLM的结合具有较大发展潜力。然而，上述研究仍局限于文本理解或知识结构构建，尚未形成训练−评测−应用的完整闭环，也缺乏系统性的问答生成与决策支持能力。尽管RAG在缓解幻觉和知识更新方面具有优势，但在林业实际应用中，单纯依靠检索往往难以应对高度专业化的术语歧义，且在长链条逻辑推理与复杂指令遵循方面存在局限性。同时，林业决策场景对响应时效性要求较高，RAG的多步检索过程会显著增加推理延迟。相比之下，指令微调将领域专业知识与推理模式融入模型参数，有助于提升模型的内生推理能力与响应速度，对构建深度适配林业场景的决策支持模型具有重要意义。

为弥补上述研究不足，本文提出林云（LinYun）——一款面向林业领域的大语言模型，以期为生态管理与决策支持的智能化应用提供参考。本文核心贡献包括以下3个方面。

（1）三阶段数据合成流程。以林业教材和规范文本为基础构建初始语料，采用Evol-Instruct[15]方法对初始指令进行多轮进化，提升任务的复杂性与覆盖度；随后结合通用模型生成问答对，并采用CRITIC（criteria importance through inter-criteria correlation）权重方法进行多维度质量加权筛选，构建内容丰富、质量可靠的林业专用指令微调数据集。

（2）测试集构建。在评测环节，从公开网络中收集了多类型客观题（包括单选、多选、填空与判断），构建了Lin-MMLU测试集，用于考察模型的知识掌握与推理能力。同时，从合成语料中经人工筛选形成Lin-QA开放问答测试集，用于评估模型在综合理解与生成表达方面的表现，从而形成覆盖客观与主观任务的双测试集体系。

（3）多维度评测体系。基于上述双测试集，建立了覆盖客观题与开放问答两类任务的多维度评测框架：在客观题任务中，采用多种不同指标衡量预测准确性；在开放问答任务中，引入大模型评审机制，从相关性、完整性、逻辑性等维度进行综合评分。该体系从精度、专业性与生成能力3个方面对模型进行系统性评估，为林业领域模型的性能对比与迭代提供统一基准。

1 方法

1.1 三阶段的林业数据合成

为获取高质量的林业训练数据，本文采用三阶段数据合成流程，系统化地构建林业领域指令微调数据集，如图1所示。整个流程从林业教材和规范文本出发，经过清洗、演化和生成3个环节，逐步得到可用于模型训练的高质量语料。

三阶段数据合成流程具体介绍如下。

首先，如图1（a）所示，以数十本林业教材与规范文件为主要数据源，文本按章节或节段切分后得到初始候选片段。考虑到林业教材与规范文本中往往存在大段重复或相似度极高的内容，本文采用基于局部敏感哈希（LSH）的近似重复检测，将Jaccard相似度阈值设为0.8，从而去除高度重复的片段。在此基础上，进行去噪处理，清除页眉页脚、广告、水印以及编码乱码等非结构化内容，以保证语料的纯净性与可用性。为减轻数据集在主题覆盖上的偏倚，本文引入多样性采样机制，使保留片段不仅在数量上充足，而且覆盖森林经营、病虫害防治、生态管理与法律法规等方面。

然后，基于第一阶段得到的多样化语料片段，本文结合任务模板构造初始指令集合（图1（b））。

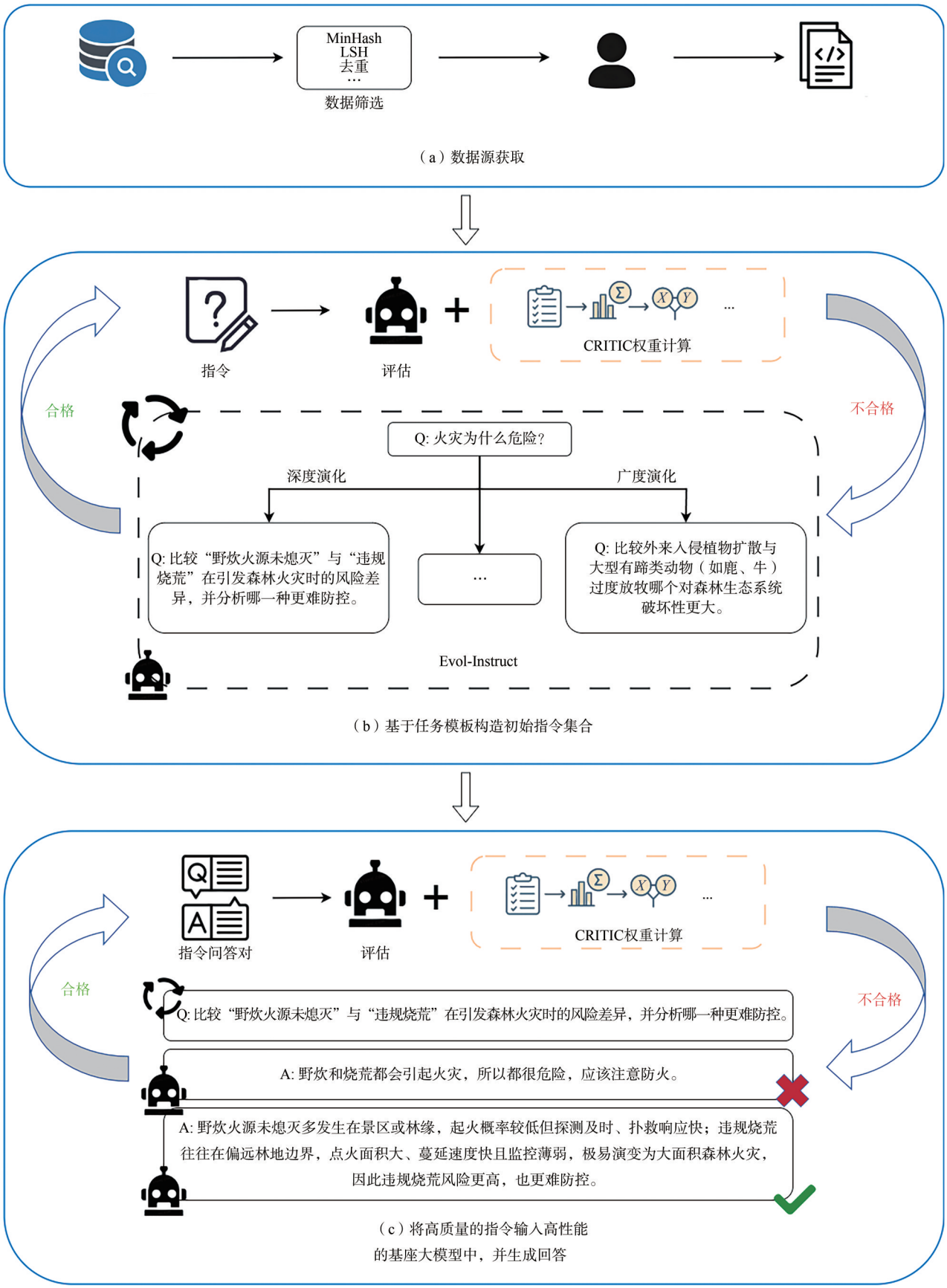


图 1　三阶段数据合成流程

这些指令的复杂度与层次性仍有限，难以充分支撑模型在推理与生成任务上的表现。为此，本文引入Evol-Instruct方法[15]对指令进行多轮进化。其核心思想是利用大语言模型对原始指令不断改写和扩展，使指令在复杂性与覆盖度上逐步提升。进化过程包括两大类：一是“深度进化”，通过增加条件约束、细化概念定义、延长推理链条、具体化输入场景以及复杂化输入形式等方式，使生成的任务更具挑战性与专业性；二是“广度进化”，在保持林业领域相关性的同时，构造与原始指令差异度更大的新问题，以提升语料在任务分布上的多样性。每轮演化之后，通过自动化评分模型对指令进行评估，维度涵盖一致性、相关性、教育价值与可评测性等。若某条指令在这些维度上的得分偏低，则进入下一轮演化，直至达到设定的质量阈值。该迭代机制保证了最终指令数据在覆盖广度、复杂度与专业性上均满足要求。

最后，在问答生成阶段，本文将高质量的指令输入高性能的基座大模型（如DeepSeek-v3[16]）并生成对应的答案（图1（c））。针对生成的问答对，本文设计了多维度质量评分体系，重点考察事实准确性、内容完整性、语言可读性以及引用充分性等。得分不足的问答对被标记为低质量样本，并通过自动重写或再生成进行修正。此外，从生成结果中抽取不少于5%的样本进行人工复核，以确保最终数据的可靠性与可控性。通过上述流程，数据集在质量与专业性上得到了双重保障。

1.2 多维度打分体系与CRITIC权重计算

本文针对指令与问答对设计了多维度打分体系，包括一致性、相关性、教育价值、整体评价等10个评价维度。仅依靠简单的整体评价虽能提供初步参考，但由于其信息维度有限，难以全面反映样本质量的各个方面。为提高评分体系的科学性与判别力，本文参考Wang等[17]的方法，设计了适用于上述多维度评分的CRITIC权重计算方法，通过同时考虑各评价维度得分的差异性与维度间的相关性，客观确定各维度权重。CRITIC方法的核心思想在于：一方面，利用各维度评分的标准差度量该维度的对比强度，即区分样本差异的能力；另一方面，利用各维度间的Pearson相关系数[18]度量指标之间的信息冗余程度（维度间相关性越强，则提供的新增信息越少）。综合标准差与相关系数两方面，即可得到各评价维度的客观权重。

具体而言，首先通过计算各维度评分的标准差衡量其对比强度，即该维度区分样本差异的能力。各维度评分的标准差定义为：

$$\sigma_j=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_{ij}-\bar{x}_j\right)^2} \tag{1}$$

其中，N为样本总数，x_{ij}为第i个样本在第j个维度的得分，$\bar{x}_j$为第j个维度得分的平均值。标准差越大，该维度的区分度越强，提供的信息越丰富。

其次，为衡量不同维度之间的信息冗余程度，本文采用Pearson相关系数r_{jk}来计算任意两个维度j和k之间的线性相关性：

$$r_{jk}=\frac{\sum_{i=1}^{N}\left(x_{ij}-\bar{x}_j\right)\left(x_{ik}-\bar{x}_k\right)}{\sqrt{\sum_{i=1}^{N}\left(x_{ij}-\bar{x}_j\right)^2}\cdot\sqrt{\sum_{i=1}^{N}\left(x_{ik}-\bar{x}_k\right)^2}} \tag{2}$$

若两个维度之间高度相关，则它们提供的信息存在较大重叠，边际增益较低。

在此基础上，为量化各维度与其他维度之间的独立性，定义冲突性指标（Conflict）如下：

$$\text{Conflict}_j=\sum_{k=1}^{m}\left(1-\left|r_{jk}\right|\right) \tag{3}$$

其中，m为维度总数。若维度j与其他维度的相关性较低，则Conflict_j越大，提供的新增信息越多。

进而，综合各维度的对比强度与独立性，可得其信息量为：

$$C_j=\sigma_j\cdot\text{Conflict}_j \tag{4}$$

信息量越大，表明该维度在整体评价中越重要。

为保证各维度权重之和为1，本文对信息量进行归一化处理，得到CRITIC权重：

$$w_j=\frac{C_j}{\sum_{l=1}^{m}C_l} \tag{5}$$

该权重能综合反映各维度的区分度与独立性，减轻单一维度偏置对结果的影响。

最后，基于所得权重，对每个样本的各维度得分进行加权求和，得到综合得分：

$$S_i=\sum_{j=1}^{m}w_j\cdot x_{ij} \tag{6}$$

其中，S_i为第i个样本的最终综合评分，用于筛选与保留高质量的指令−问答对样本。上述方法提升了打分体系的客观性与判别力，为后续模型训练提供了可靠的数据质量保障。

2　评测

在测试集构建方面，本文从多种渠道系统性地收集了数据，包括林业相关的网络公开资源、行业资料以及历年官方考试试题，用于构建客观题测试集，同时纳入经人工审核的高质量开放问答样本，最终形成了 Lin-MMLU 与 Lin-QA 两类测试集。其中，Lin-MMLU 共包含 2 213 道客观题，题型分布见表 1，覆盖填空题 162 道（7.3%）、判断题 484 道（21.8%）、单选题 1 029 道（46.5%）与多选题 538 道（24.3%），用于考察模型在不同客观题型上的知识掌握与推理能力。Lin-QA 共包含 200 条经人工严格审核的开放问答样本，输入与输出的平均长度分别为 19.9 字符与 393.2 字符。上述两类测试集共同构成了面向林业任务的系统化评测基础。

表 1　测试集分布

测试集	题型	数量/道
Lin-MMLU	单选题	1 029
	多选题	538
	填空题	162
	判断题	484
Lin-QA	问答题	200

在评测指标方面，针对不同类型任务，本文采用了多维度评价方法。对于选择题，以 Accuracy 为主要指标，Accuracy 是指模型正确预测的答案所占的比例，用于衡量模型的整体准确性。对于多选题，进一步采用 Hamming-F1[19] 作为评估指标，Hamming-F1 是基于 Hamming 损失的 F1 度量，可计算预测答案与参考答案之间的准确性，具体而言，对于每个选项，Hamming 损失计算了预测答案与真实答案的不同（0 代表正确，1 代表错误），F1 值则综合考虑了预测正确的选项和预测错误的选项之间的平衡。对于填空题，本文引入 ROUGE-L[20] 作为文本匹配指标，ROUGE-L 通过计算生成文本与参考答案之间的最长公共子序列量化模型输出的质量，从而反映生成内容与参考答案的相似性。对于开放问答题，本文采用 Qwen3-32B 作为评审模型，从相关性、完整性和逻辑性等维度对生成答案进行综合打分。上述多层次的评测设置能够较全面、客观地检验 LinYun 在各类林业任务上的表现。

3　实验设置

本文以 Qwen2.5-7B-Instruct[21] 与 Qwen3-8B[22] 为基座模型，在配备 NVIDIA 4×A6000（48 GB 显存）的硬件环境下进行指令微调，得到林业领域模型 LinYun-7B 与 LinYun-8B。为更全面地检验其性能，本文额外引入 14B 级通用模型作为对照，以考察 LinYun 在不同参数规模下的相对表现。训练采用 LLaMAFactory[23] 框架，支持 LoRA、QLoRA 等指令微调方式及多 GPU 并行与灵活调度。在参数高效微调方面，本文采用 LoRA[24]（低秩适配）方法，其核心思想是冻结预训练模型的主体权重，仅在 Transformer 各层添加少量低秩矩阵并对其实施微调，这使训练过程能够极大地减少 GPU 显存使用。本文具体训练配置如下：训练时单卡 batch size 为 8，同时设置梯度累积为 8，有效 batch size 为 64。采用 AdamW 优化器，初始学习率为 2×10^{-5}，并采用余弦退火学习率调度策略。训练共进行 3 轮，确保模型在任务上获得充分微调但不过拟合。所有实验均采用 LoRA，参数配置为 rank = 16，$\alpha = 32$，以平衡可训练参数规模与正则化强度。上述配置在资源利用与训练效果之间取得平衡：LoRA 显著降低显存需求，适配多 GPU 并行训练，同时保留模型泛化与细节学习能力。

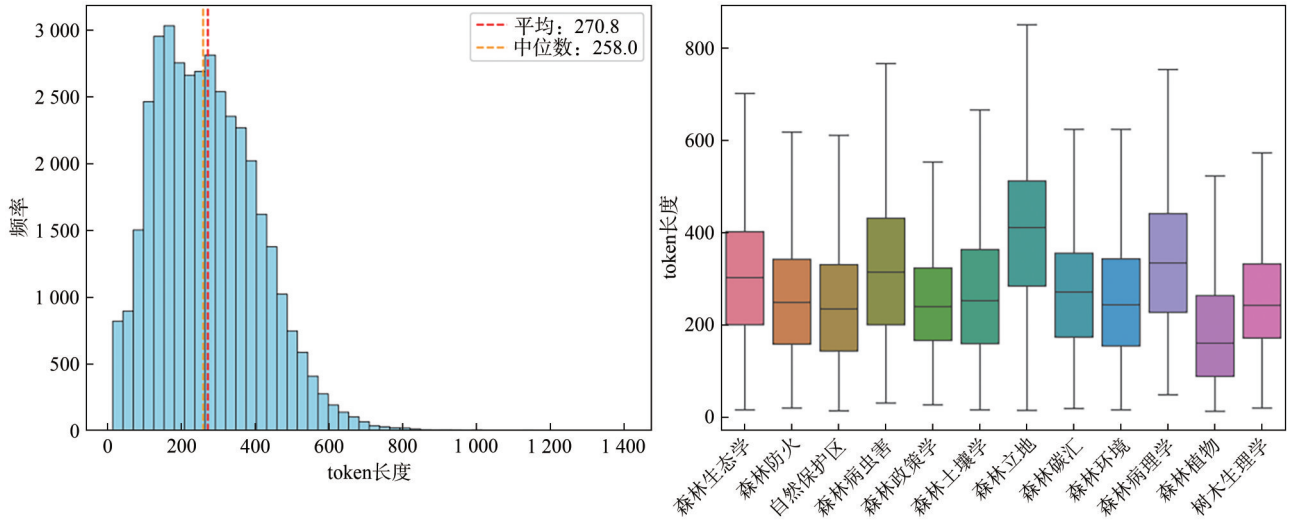
4　结果分析

4.1　数据统计与特征分析

从整体规模来看，本文基于公开林业教材等资料合成了约 7 万条指令−问答对，经去重、指令演化与质量筛选等多阶段流程后，最终保留约 4 万条高质量数据，覆盖林业多个核心领域。

从文本长度分布（图 2）来看，数据整体呈现出明显的长尾特征。训练样本的平均 token 长度（含输入和输出）为 270.8，中位数为 258.0。进一步分析表明，约 90% 的样本长度低于 456 token。这种分布特征说明，绝大多数样本处于中等规模的文本区间，可在保证知识承载量的同时避免过长文本带来的训练负担，兼顾效率与信息量。

在主题覆盖（图 3）方面，数据集涵盖 13 个林业核心子领域，分布较为均衡。其中，森林防火（17.8%）、森林土壤学（17.1%）、森林生态学（12.8%）与森林环境（12.4%）占比较高，体现了训练语料对林业核心任务与生态管理的侧重；森林



（a）整体token分布　　（b）各领域token分布

图2　林业训练数据token长度分布

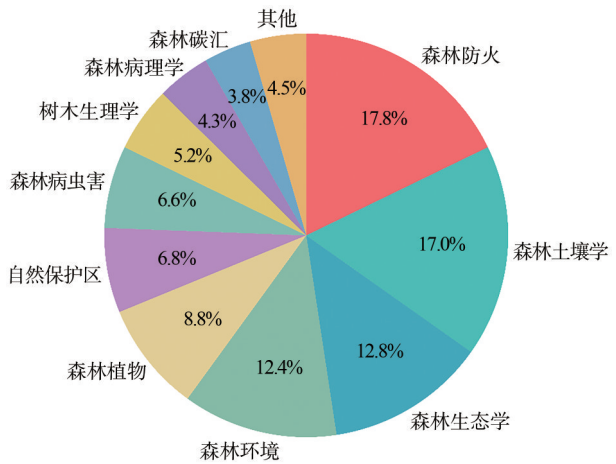


图3　林业训练数据主题覆盖分布

植物（8.8%）、自然保护区（6.8%）、森林病虫害（6.6%）等也有较充分的数据支撑，有利于模型在多类任务上的适应性；树木生理学（5.2%）、森林病理学（4.3%）、森林碳汇（3.8%）等占比相对较低，为模型提供了重要的长尾领域知识。整体来看，主题分布既突出核心领域，又兼顾长尾覆盖。

在数据质量方面，本文对训练数据进行了系统性的抽检与一致性分析。结果表明，经CRITIC筛选后的样本在各维度上均优于筛选前。人工抽检显示，筛选前的事实错误率约为7.8%，筛选后下降至3.2%；引用缺失率从10.5%降低至4.7%，表明CRITIC权重机制提升了语料在事实性与可验证性上的质量。上述结果表明，三阶段流程与权重计算机制不仅在理论上具备合理性，也在实际操作中显著提升了数据的可靠性与可控性，为后续模型的微调与评测提供了坚实的数据基础。

4.2　不同模型在各任务上的表现

表2~表4汇总了各模型在林业任务测试集Lin-MMLU与Lin-QA上的表现。整体来看，LinYun系列在林业相关任务上普遍优于同规模通用基座模型，并在多项指标上接近甚至超过更大规模模型的水平，体现了领域指令微调的有效性。

由表2可知，在Lin-MMLU的各类客观题中，LinYun-7B与LinYun-8B均较各自基座模型（Qwen2.5-7B-Instruct与Qwen3-8B Non-thinking）取得更好成绩。LinYun系列在单选题、多选题、填空题与判断题4类任务上表现更均衡，基座模型则存在明显的题型偏差，说明领域指令微调既扩大了知识覆盖面，也增强了跨任务迁移与泛化能力。在Lin-QA开放问答任务中，LinYun-7B和LinYun-8B均优于同规模基座模型，生成内容在专业性、逻辑性和完整性方面表现更佳，表明LinYun不仅在知识检索类任务中有优势，在需要综合理解与语言组织能力的开放式生成任务中也能发挥更强的能力。

为验证领域模型相较于检索增强生成（retrieval-augmented generation，RAG）方法的必要性，本文构建了基于通用模型的RAG系统作为对比基线。RAG采用向量检索引擎，以BGE-M3[25]为嵌入模型，索引库包含全部训练语料，检索Top-K设为5。由表2可知，尽管RAG系统在知识检索能力上有所提升，LinYun系列模型在整体任务上仍保持领先。此外，从推理效率角度看，LinYun-8B单次推理时延约为1.2 s（batch size=1），Qwen3-8B+RAG

表 2　LinYun-7B/8B 在 Lin-MMLU 与 Lin-QA 上的性能对比

模型	Lin-MMLU					Lin-QA
	单选题	多选题	填空题	判断题	Overall	问答题
Qwen3-8B Non-thinking	70.5%	56.3%	20.9%	74.4%	64.2%	73.4%
Qwen3-8B RAG	71.8%	57.1%	22.4%	75.2%	65.4%	76.2%
LinYun-8B	73.3%	58.2%	23.5%	76.2%	66.6%	79.5%
Qwen2.5-7B	74.5%	56.5%	22.3%	74.4%	66.5%	71.6%
Qwen2.5-7B RAG	75.6%	57.8%	23.1%	75.8%	67.5%	73.8%
LinYun-7B	77.8%	59.4%	24.6%	77.2%	69.3%	72.4%

表 3　LinYun-7B/8B 与 14B 级通用模型在 Lin-MMLU 与 Lin-QA 上的性能对比

模型	Lin-MMLU					Lin-QA
	单选题	多选题	填空题	判断题	Overall	问答题
Qwen3-14B Non-thinking	75.1%	57.8%	22.8%	75.8%	67.2%	81.8%
LinYun-8B	73.3%	58.2%	23.5%	76.2%	66.6%	79.5%
Qwen2.5-14B-Instruct	78.0%	59.7%	17.3%	77.7%	69.0%	74.6%
LinYun-7B	77.8%	59.4%	24.6%	77.2%	69.3%	72.4%

表 4　大规模通用模型在 Lin-MMLU 与 Lin-QA 上的性能对比

模型	Lin-MMLU					Lin-QA
	单选题	多选题	填空题	判断题	Overall	问答题
Qwen3-32B Non-thinking	80.0%	61.3%	30.9%	77.9%	71.4%	83.2%
Qwen2.5-32B-Instruct	85.1%	66.2%	23.5%	80.4%	74.9%	74.6%
DeepSeek-v3	81.1%	66.2%	30.9%	78.3%	73.2%	89.2%
GPT-4o	73.5%	59.3%	25.3%	70.9%	65.9%	71.8%

因额外的检索与上下文拼接，平均时延约为 2.8 s，响应速度相对降低约 57%，凸显了专用模型在实际部署中的效率优势。

与更大规模的通用模型相比，LinYun 展现出明显的参数效率优势。由表 3 可知，在 Lin-MMLU 上，LinYun-7B 的整体表现已接近 Qwen2.5-14B-Instruct 等 14B 级模型，LinYun-8B 也与 Qwen3-14B Non-thinking 的水平相当。

虽然 DeepSeek-v3 在整体上仍领先，但与 GPT-4o[26] 相比，LinYun 在单选题（77.8% vs 73.5%）、多选题（59.4% vs 59.3%）和判断题（77.2% vs 70.9%）等多数细项上已经实现反超，在总体得分 Overall（69.3% vs 65.9%）上领先；在 Lin-QA 开放问答任务中，LinYun-8B 得分为 79.5%，明显高于 GPT-4o 的 71.8%（表 4）。同时，LinYun 的显存与算力消耗远低于这些超大模型，具有更好的性价比。

综上，通过领域定向的数据构建与微调，LinYun 能以中等参数规模在林业任务上接近大模型的专业性能，在资源成本与效果之间取得较优平衡。LinYun 在计算资源可控的前提下兼顾了性能与效率，具备在林业及相关生态任务中推广应用的潜力。

4.3　消融实验

为验证 CRITIC 方法对模型性能的贡献，本文进行了消融实验。在保持训练框架与基座模型不变的前提下，考察引入 CRITIC 权重方法对数据筛选与模型效果的影响。在相同数据规模下设置 3 组实验：采用 CRITIC 方法对多维度打分加权并以此筛选出最终训练语料；直接采用平均权重进行综合打分并筛选（average）；采用随机权重进行打分（random），其中各维度权重从[0.05, 0.15]均匀采样后归一化（权重和为 1），重复 3 次实验取平均值以降低随机性。

表 5 表明，采用 CRITIC 方法的模型在多项任务上均优于对比模型。在 Lin-MMLU 任务上，使用 CRITIC 方法的模型综合得分为 66.6%，较采用简单平均权重的基线（64.9%）提升约 1.7 个百分点，较随机权重模型（64.5%）提升了 2.1 个百分点；尤其是在多选题上提升明显，说明 CRITIC 能

更好地利用各样本质量维度间的差异。值得注意的是，随机权重策略的表现略低于简单平均策略，表明权重设置是否合理对数据质量影响很大——不恰当的权重配置不仅难以改善筛选效果，还可能引入噪声、降低整体质量。在 Lin-QA 开放问答中，主观打分结果显示，CRITIC 方法在事实一致性和完整性上更具优势。

上述结果表明，CRITIC 方法通过合理分配各质量维度的权重，使筛选出的训练样本质量更高，进而提升模型在林业任务上的表现。换言之，样本筛选机制不仅影响数据质量控制，也对最终模型性能具有直接的正向作用。

为量化 Evol-Instruct 方法中“深度进化”和“广度进化”两种策略对不同题型的独立贡献，本文设置了以下 4 组对比实验。①无进化（Baseline）：仅使用初始任务模板生成的指令，未经任何进化处理；②仅深度进化（Depth-only）：只执行深度进化策略（增加约束、细化概念等），不进行广度扩展；③仅广度进化（Breadth-only）：只执行广度进化策略（创造差异化问题），不增加复杂度；④深度+广度进化（All）：同时执行两种进化策略，即本文采用的方案。参照已有经验[27]，各组实验均进行 3 轮进化。

如图 4 所示，完整的 Evol-Instruct 流程在所有题型上均取得最优表现，该结论在不同规模基座上一致。LinYun-7B 与 LinYun-8B 在各策略下的趋势一致：相对于无进化 Baseline，深度+广度进化方案在 Lin-MMLU 上分别提升 6.1 个百分点和 5.8 个百分点，在 Lin-QA 上分别提升 3.5 个百分点和 7.3 个百分点。这种跨模型的稳定性验证了 Evol-Instruct 方法的泛化性和鲁棒性。

4.4 基于 Lin-QA 的幻觉评估

幻觉（hallucination）问题指模型生成看似合理实则包含事实错误或虚构信息的答案，是领域模型面临的核心挑战之一。针对模型可能产生的事实性错误，本文利用已有的 Lin-QA 测试集进行了幻觉评估分析。

Lin-QA 测试集包含 200 条样本，均经过林业领域专家人工标注，参考答案具有较高的可靠性。本文以 DeepSeek-v3 为错误类型判定工具，对模型在 Lin-QA 上的回答进行逐条分析，将答案错误归因为以下 4 类。

（1）事实性幻觉：模型编造不存在的事实、数据或引用，或将正确事实张冠李戴。例如将“马尾松”的特征描述为“红松”的特征，或编造不存在的法规条款。

表 5 CRITIC 权重方法消融实验对比

模型	Lin-MMLU					Lin-QA
	单选题	多选题	填空题	判断题	Overall	问答题
LinYun-8B	73.3%	58.2%	23.5%	76.2%	66.6%	79.5%
LinYun-8B average	72.1%	55.3%	21.3%	74.8%	64.9%	78.1%
LinYun-8B random	70.8%	56.7%	20.4%	74.8%	64.5%	75.4%

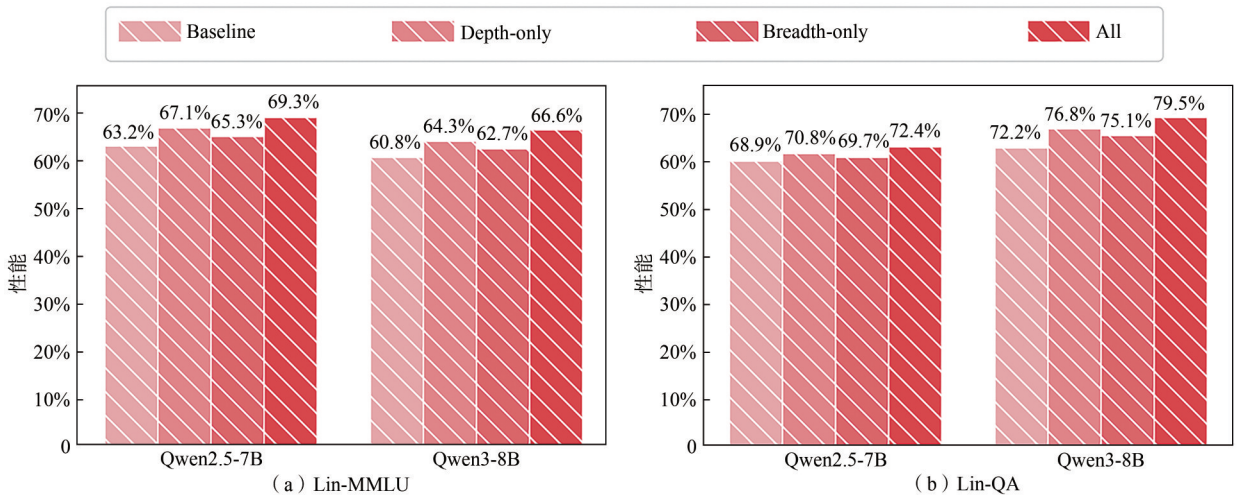


图 4 Evol-Instruct 不同进化策略的性能对比

（2）知识缺失：模型对问题涉及的专业知识掌握不足，导致答案不完整或偏离主题，但未明显编造信息。

（3）推理错误：模型具备相关知识，但在逻辑推理、因果关系判断或多步骤推理中出现错误。

（4）表达问题：模型答案内容基本正确，但表述不规范、组织混乱或存在轻微的术语使用不当。

DeepSeek-v3 根据参考答案和模型回答进行对比分析，判定错误类型并给出置信度评分（0~1）。本文统计各模型在 Lin-QA 上得分低于 75 分（满分 100 分）的样本，并分析其错误原因分布。

由表 6 可知，LinYun 系列在事实一致性上明显优于对照组。在低分样本中，LinYun-7B 和 LinYun-8B 的事实性幻觉占比分别为 15.4% 和 14.6%，较同规模通用模型（Qwen2.5-7B 为 31.0%，Qwen3-8B 为 28.7%）约降低一半，较更大规模的 GPT-4o（24.2%）约降低 9 个百分点。这一结果表明垂直领域训练与 CRITIC 方法在抑制幻觉问题上的有效性。进一步分析 LinYun 在错误类型分布上的结构性变化，其低分样本中知识缺失占比上升至 50% 以上，事实性幻觉占比显著下降，说明 LinYun 在不确定时更倾向于保守回答而非编造信息，行为模式更可靠，对林业等要求高事实准确性的应用场景尤其重要。

5　讨论

本文提出并验证了 LinYun 的构建方法及其在林业任务上的表现。实验结果表明，通过三阶段数据合成与领域指令微调，LinYun 在各类林业任务中的性能均优于同规模通用基座模型，部分任务接近甚至超过更大规模通用模型。

在方法上，本文提出的三阶段数据合成机制有效支撑了高质量林业语料的构建。相较于传统的人工构建或简单爬取，该流程融合了语义去重、指令进化与多维度质量筛选，显著提升了数据的多样性、复杂性与事实准确性。消融实验表明，基于 CRITIC 权重方法的样本筛选能更客观地评估样本质量，其引入提升了模型在多项任务中的表现，尤其在细粒度知识推理（如多选题）与完整性要求较高的开放问答中效果明显，表明数据质量的一致性和评价维度的科学性对领域模型最终性能具有关键影响。

与现有林业领域语言模型相比[13-14]，LinYun 不仅涵盖文本理解任务，还具备推理和开放生成能力。评测体系中的 Lin-MMLU 和 Lin-QA 测试集分别从客观知识掌握和主观综合表达能力两方面提供评估依据，为林业领域模型的性能衡量提供了基准。

尽管 LinYun 在林业任务中表现良好，但仍存在局限。首先，训练数据主要来源于林业教材、规范文本及公开试题，规模与覆盖范围有限，这使模型在面对跨学科推理或未见领域任务时存在瓶颈。其次，虽然通过 Evol-Instruct 与 CRITIC 权重方法提升了数据质量，但生成语料仍可能含有一定噪声与偏差，对模型的稳健性构成挑战。此外，当前评测体系主要集中在客观题与问答任务，对真实场景中的多模态输入（如遥感影像）或复杂决策模拟尚未覆盖。

未来工作可从多个方向进一步推进。第一，扩充数据来源，特别是引入科研论文、实地调研报告及多模态数据，以丰富知识广度与跨模态能力。第二，探索更多训练与推理策略，如检索增强生成（RAG）[28-29]、链式思维推理（Chain-of-Thought）[30-32] 与偏好优化等，以进一步提升模型的推理透明度与决策可靠性。第三，在应用层面上引入 Function calling[33-35]，使 LinYun 能够与外部工具联动，例如调用林业三维仿真系统，实现更强的交互式推理与任务执行能力，为应急管理与生态模拟等场景中的应用奠定基础。

表 6　Lin-QA 错误原因分析（低分样本统计）

模型	低分样本数	事实性幻觉	知识缺失	推理错误	表达问题
Qwen3-8B Non-thinking	68/200（34.0%）	28.7%	42.6%	19.1%	9.6%
LinYun-8B	48/200（24.0%）	14.6%	52.1%	22.9%	10.4%
Qwen2.5-7B	71/200（35.5%）	31.0%	40.8%	18.3%	9.9%
LinYun-7B	52/200（26.0%）	15.4%	50.0%	23.1%	11.5%
GPT-4o	62/200（31.0%）	24.2%	38.7%	25.8%	11.3%

此外，可将LinYun应用于森林经营与碳汇计算等任务[36]。例如，在森林经营中，结合遥感监测、林分动态模拟与政策约束，辅助制定更科学的森林抚育与采伐策略，提升经营活动的可持续性[37-38]。在碳汇计算方面，LinYun可与碳循环模型、遥感数据及地面调查相结合，提升碳储量估算的自动化与精度。该方向既契合“双碳”背景下的研究需求，也为林业领域提供了决策工具[39]。通过上述扩展，LinYun有望逐步发展为兼具学术研究价值与应用潜力的林业智能模型。

6 结束语

本文提出面向林业的数据合成、模型训练与系统化评测的一体化框架，并在通用大语言模型上进行领域指令微调，得到面向林业领域的大语言模型LinYun。在方法上，本文提出了三阶段数据合成流程与CRITIC权重法，并通过消融实验验证了方法的可靠性；在评测上，本文构建了Lin-MMLU与Lin-QA两类测试集，为林业模型的系统化性能检验提供了基准。实验结果显示，LinYun在林业相关的客观题与开放问答任务上的表现均显著优于同规模通用模型，并在部分任务上接近甚至超过更大规模模型的表现。本文为林业智能化提供了可行的技术路径，未来，LinYun有望进一步发展为支撑林业科研、政策制定与生态管理的综合智能模型。

参考文献：

[1] Tian Y H, Gan R Y, Song Y, et al. ChiMed-GPT: a Chinese medical large language model with full training regime and better alignment to human preferences[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2024: 7156-7173.

[2] Wang G Y, Yang G X, Du Z X, et al. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation[PP]. V1. (2023-06-16)[2025-09-30]. arXiv: arXiv.2306.09968.

[3] Chen J Y, Wang X D, Ji K, et al. HuatuoGPT-II, one-stage training for medical adaption of LLMs[PP]. V2. (2024-09-15)[2025-09-30]. arXiv: arXiv.2311.09774.

[4] Zhou Z, Shi J X, Song P X, et al. LawGPT: a Chinese legal knowledge-enhanced large language model[PP]. V1. (2024-06-07) [2025-09-30]. arXiv: arXiv.2406.04614.

[5] Huang Q Z, Tao M X, Zhang C, et al. Lawyer LLaMA technical report [PP]. V2. (2023-10-14)[2025-09-30]. arXiv: arXiv.2305.15062.

[6] 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.

[7] Ermon S, Finn C, Manning C D, et al. Direct preference optimization: your language model is secretly a reward model[C]//Proceedings of the Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 53728-53741.

[8] 王京鲁, 李杰, 冯晓川, 等. 基于人工智能的森林火灾发生预测研究进展[J]. 陆地生态系统与保护学报, 2025, 5(3): 81-89.
Wang J L, Li J, Feng X C, et al. Research progress on the prediction of forest fire occurrence based on artificial intelligence[J]. Terrestrial Ecosystem and Conservation, 2025, 5(3): 81-89.

[9] 崔晓辰, 雷一东. 基于深度学习的林业害虫智能化检测方法研究进展[J]. 世界林业研究, 2024, 37(4): 53-57.
Cui X C, Lei Y D. Research progress in deep-learning-based intelligent forest pest detection methods[J]. World Forestry Research, 2024, 37(4): 53-57.

[10] 谭晶维, 张怀清, 郭梦蕾, 等. 林草行业大模型构建思路与应用前景[J]. 林业科学, 2025, 61(7): 170-181.
Tan J W, Zhang H Q, Guo M L, et al. Construction ideas and application prospects of large models in forestry and grassand industry[J]. Scientia Silvae Sinicae, 2025, 61(7): 170-181.

[11] Xie Y, Jiang B W, Mallick T, et al. WildfireGPT: tailored large language model for wildfire analysis[PP]. V4. (2025-04-23)[2025-09-30]. arXiv: arXiv.2402.07877.

[12] Du S Q, Tang S J, Wang W X, et al. Tree-GPT: modular large language model expert system for forest remote sensing image understanding and interactive analysis[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2023, 48: 1729-1736.

[13] Tan J W, Zhang H Q, Yang J, et al. ForestryBERT: a pre-trained language model with continual learning adapted to changing forestry text[J]. Knowledge-Based Systems, 2025, 320: 113706.

[14] Sun J Y, Luo Z Z. ForPKG: a framework for constructing forestry policy knowledge graph and application analysis[PP]. V2. (2025-04-29) [2025-09-30]. arXiv: arXiv.2411.11090.

[15] Xu C, Sun Q F, Zheng K, et al. Wizardlm: Empowering large language models to follow complex instructions[C]//International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2024: 43495-43516.

[16] Liu A X, Feng B, Xue B, et al. DeepSeek-v3 technical report[PP]. V2. (2025-02-18)[2025-09-30]. arXiv: arXiv.2412.19437.

[17] Wang Y M, Luo Y. Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making[J]. Mathematical and Computer Modelling, 2010, 51(1/2): 1-12.

[18] Pearson K. Note on regression and inheritance in the case of two parents[J]. Proceedings of the Royal Society of London, 1895, 58: 240-242.

[19] Wu G, Zhu J. Multi-label classification: do hamming loss and subset accuracy really conflict with each other[C]//Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2020: 3130-3140.

[20] Lin C Y. ROUGE: a package for automatic evaluation of summaries[C]// Proceedings of the Text Summarization Branches Out. Stroudsburg: ACL, 2004: 74-81.

[21] Team Q. Qwen2 technical report[PP]. V4. (2024-09-10) [2025-09-30]. arXiv: arXiv.2407.10671.

[22] Yang A, Li A F, Yang B S, et al. Qwen3 technical report[PP]. V1. (2025-05-14)[2025-09-30]. arXiv: arXiv.2505.09388.

[23] Zheng Y W, Zhang R C, Zhang J H, et al. LlamaFactory: unified efficient fine-tuning of 100+ language models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Stroudsburg: ACL, 2024: 400-410.

[24] Hu E, Shen Y L, Wallis P, et al. LoRA: low-rank adaptation of large language models[C]//International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2022: 12513-12525.
[25] Chen J L, Xiao S T, Zhang P T, et al. M3-embedding: multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation[C]//Proceedings of the Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL, 2024: 2318-2335.
[26] HURST A, Lerer A, Goucher A P, et al. GPT-4o system card[PP]. V1. (2024-10-25)[2025-09-30]. arXiv: arXiv.2410.21276.
[27] Luo Z Y, Xu C, Zhao P, et al. WizardCoder: empowering code large language models with evol-instruct[C]//International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2024: 3190-3210.
[28] Sharma K, Kumar P, Li Y Q. OG-RAG: ontology-grounded retrieval-augmented generation for large language models[PP]. V1. (2024-12-12) [2025-09-30]. arXiv: arXiv.2412.15235.
[29] Edge D, Trinh H, Cheng N, et al. From local to global: a graph RAG approach to query-focused summarization[PP]. V2. (2025-02-19) [2025-09-30]. arXiv: arXiv.2404.16130.
[30] Bosma M, Chi E, Ichter B, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the Advances in Neural Information Processing Systems 35. New Orleans: Curran Associates, Inc., 2022: 24824-24837.
[31] Cao Y, Griffiths T, Narasimhan K, et al. Tree of thoughts: deliberate problem solving with large language models[C]//Proceedings of the Advances in Neural Information Processing Systems. New Orleans: Curran Associates, Inc., 2023: 11809-11822.
[32] Liao H R, Hu S H, Zhu Z H, et al. Forest for the trees: overarching prompting evokes high-level reasoning in large language models[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg: ACL, 2025: 1433-1453.
[33] Abdelaziz I, Basu K, Agarwal M, et al. Granite-function calling model: introducing function calling abilities via multi-task learning of granular tasks[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. Stroudsburg: ACL, 2024: 1131-1139.
[34] Zhao W K, Hua J, Wang X J, et al. ToolPlant: tool-based natural language interface for plant simulation models[C]//Proceedings of the 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Piscataway: IEEE Press, 2025: 5673-5678.
[35] Liu W, Huang X, Zeng X, et al. ToolACE: winning the points of LLM function calling[C]//Proceedings of the International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2025: 41359-41381.
[36] 陈振升, 罗陶然, 段佳丽, 等. 智慧林业科技前沿与创新发展研究[J]. 西南林业大学学报, 2025, 45(11): 199-208.
Chen Z S, Luo T R, Duan J L, et al. Research on the technological frontiers and innovative development of smart forestry[J]. Journal of Southwest Forestry University, 2025, 45(11): 199-208.
[37] Pretzsch H, Biber P, Schütze G. Forest stand growth dynamics in Central Europe: empirical results and model approaches[J]. Forest Ecology and Management, 2015, 336: 252-264.
[38] Yousefpour R, Hanewinkel M, Jacobsen J B. Modeling adaptive forest management strategies[J]. Forest Policy and Economics, 2012, 24: 16-26.
[39] Pan Y D, Birdsey R A, Fang J Y, et al. A large and persistent carbon sink in the world's forests[J]. Science, 2011, 333(6045): 988-993.

[作者简介]

赵维康（2000– ），男，中国科学院自动化研究所多模态人工智能系统全国重点实验室硕士生，主要研究方向为面向林业应用的大语言模型与智能体。

王浩宇（1984– ），男，中国科学院自动化研究所多模态人工智能系统全国重点实验室副研究员，主要研究方向为植物生长建模、智慧农业、编程语言及信息系统。

华净（1981– ），男，中国科学院自动化研究所多模态人工智能系统全国重点实验室助理研究员，主要研究方向为植物生长建模、智慧农业、编程语言及计算机图形学。

刘耀兵（1982– ），男，北京江恒数智生态科技有限公司总经理，主要研究方向为森林经理学及大模型推广应用。

王秀娟（1982– ），女，中国科学院自动化研究所多模态人工智能系统全国重点实验室副研究员，主要研究方向为平行农业和植物建模。

康孟珍（1975– ），女，中国科学院自动化研究所多模态人工智能系统全国重点实验室副研究员，主要研究方向为平行农业和计算植物。