

基于Deepseek的水土保持智能助手的开发与研究

陈 珏¹ 龚博宇^{1,2} 罗国平¹ 陈 志¹ 陈 向¹

1. 湖南省水利水电科学研究院 湖南长沙 410009

2. 湖南南方水利水电勘测设计院有限公司 湖南长沙 410009

摘要: 本研究紧密围绕《关于加强新时代水土保持工作的意见》中结合人工智能(AI)、大数据等现代技术与水土保持相关技术及政策的要求,深入剖析水土保持监管需求,精心设计水土保持智能助手架构。在系统搜集并整编水土保持相关数据资料的基础上,构建水土保持专业知识库。通过将Deepseek R1-32B 离线模型与本地知识库有机结合,部署适用于水土保持监管专业的本地离线智能助手。同时,依据实践情况,分析了现有水土保持智能助手搭建过程中在知识库解析训练、算力约束、模型运行等方面存在的问题,并针对上述各方面问题提出了相应的解决策略,为后续进一步完善水土保持智能助手提供参考依据。

关键词: 水土保持; 人工智能; AI; Deepseek

一、背景与意义

近年来,随着计算机能力的持续提升以及对大语言模型的深入开发与研究,各类智能助手已能够依托现有的训练库,开展自然语言理解、知识检索、信息处理、问题推理等活动^[1]。

在全面、常态化开展水土保持遥感监管之后,水土保持领域正构建以遥感监管为基本手段、以重点监管为补充、信用监管为基础的新型监管机制^[2]。而部分地区的水土保持监管人员对相关法律法规的学习不够深入,导致执法力度不足、对违规行为查处不及时,进而造成监管效率低下等问题。因此,为解决水土保持监管中遇到的问题,提高水土保持监管人员的监管效率,实现水土保持监管更高水平的标准化、规范化。将大数据、云计算等现代信息技术与水土保持深度融合,运用最新的AI工具辅助水土保持监管人员开展监管工作^[3, 8]的方案也被逐步提出。通过AI辅助,监管人员能够及时了解水土保持的法律法规,强化依法落实生产建设项目水土保持相关制度,加强全链条全过程监管,确保对违法行为进行标准化、规范化的查处。

基金项目: 湖南省水利科技项目-湖南省水土保持遥感影像智慧解译(XSKJ2024064-37); 湖南省水利科技项目-湖南省小流域综合治理项目碳汇核算方法和价值实现机制研究(XSKJ2024064-21)。

本文依据《关于加强新时代水土保持工作的意见》的要求,针对水土保持遥感监管中违法违规项目认定、查处、整改过程中基层水土保持监管人员提出的问题,通过将AI(人工智能)、大数据等现代技术与水土保持相关技术及政策要求相结合,利用最新的Deepseek-R1模型,根据保密要求训练本地知识库,搭建服务于水土保持监管的专用工具,为基层水土保持监管人员提供技术支持和决策依据。

二、建设目标

(一) 需求分析

本文利用最新的AI(人工智能)技术,将大数据、云计算等新技术与水土保持监督管理工作相融合,对已有的法律法规、政策、判例等资料进行训练,为负责监管的工作人员提供咨询和决策依据,为水土保持监管工作赋能^[4]。

(二) 建设目标

本文搭建的基于Deepseek的水土保持智能助手研发在完成建设后应服务于以下目标:

1、降低行政风险。通过对该助手的运用,补全水土保持监管人员对相关法律法规了解不够的短板,帮助水土保持监管人员开展执法决策,做到依法依规监管,推进水土保持监管的标准化、规范化、便利化。

2、提高监管效率。通过融合人工智能(AI)、大数据、云计算等新技术,整合多源数据,完成适用于水土保持监管的人工智能(AI)工具的构建,为水土保持监

管工作提供有力的支撑，最终实现对水土保持监管工作人员工作效率的提高。

(三) 要求

根据目标，水土保持智能助手所需大语言模型的部署需要满足以下要求。

1、保密性。由于专家库数据、用户信息等数据有一定的敏感性，因此为了保护数据的安全，水土保持智能助手需要在本地进行部署，确保所有数据处理都在用户自己的设备上进行。

2、本地部署运行。由于安全性要求，水土保持智能助手需要在部署完成后，各用户可以在没有互联网连接的情况下进行本地使用。

3、高度专业性。由于水土保持智能助手需要对水土保持行业知识有较深的理解，可以通过制定知识库和参数调整，在降低幻觉的同时，使其更适应特定的任务或领域。

4、低成本。智能助手要求运行成本低，运行算力低。

三、框架与设计

根据前文的需求，本文从硬件分析、平台选取、模型选取等3个方面对水土保持智能助手的框架进行确定

并进行初步设计。

(一) 硬件分析

由于水土保持智能助手进行训练只能在本地电脑上离线运行，根据工作要求，本文开展的人工智能助手开发用到的硬件配置具体如下表。

表1 硬件配置

序号	名称	名称	参数
1	CPU	Intel Core i9-14900KS	24核心32线程，最高频率可达6.2 GHz
2	显卡	NVIDIA GeForce RTX 4090	24GB显存
3	存储内存		64GB DDR5 5600MHz
4	运行内存		4TB, ssd

(二) 模型选取

Deepseek模型由于应用了混合专家(MoE)架构，显著降低了推理显存占用，提高了推理速度，且该模型开源、运行成本低。因此，根据实际要求，本文选择使用Deepseek R1模型开发水土保持智能助手。

Deepseek R1的参数级别分为7b, 32b, 70b, 671b(具体见下表2)。

表2 硬件配置

序号	模型类型(B)	参数量	推荐CPU要求	推荐运行内存要求	推荐存储容量要求	推荐显存要求	推荐场景
1	1.5	15亿	4核心以上	16GB(DDR4)	256GB		家用电脑
2	7	70亿	8核心以上	32GB(DDR4)	256GB	8GB	
3	8	80亿	8核心以上	32GB(DDR4)	256GB	8GB	
4	14	140亿	12核心以上	64GB(DDR4)	512GB	16GB	
5	32	320亿	16核心以上	96GB(DDR4)	512GB	40GB	高性能电脑
6	70	700亿	20核心以上	120GB(DDR4)	1TB	141GB	企业级
7	671	6710亿	48核心以上	768GB(DDR5)	2TB	1000GB	

根据已有的硬件设备，本文采用Deepseek R1-32b级别的模型来搭建本地水土保持智能助手。

(三) 水土保持智能助手框架

本文建设的基于Deepseek的水土保持智能助手按照实际需求，对水土保持智能助手的框架进行设计，主要分为支撑层、前端和后端三部分。

1、前端

(1) 应用层。应用层主要用于用户和水土保持智能助手的交互，其中包括水土保持智能助手的用户登录界面、智能助手的交互界面等。

(2) 用户层。用户层主要用于用户的信息、权限、登录状态、许可授权等功能，其为用户的登录、注销、使用提供授权认证。

2、后端

(1) 支撑层。支撑层主要是指支撑平台运行的存储设备、计算设备、网络工程、大模型支撑工具、运行环境等基础资源。

(2) 数据层。数据层包含有智能助手需要用到的相关数据，其中包括用户数据、权限数据、知识库数据、日志数据、临时存储数据，可用于智能助手模型训练、

模型调用、权限调查、搜索记录保存等功能。

(3) 训练层。训练层主要用于对本地知识库进行调用、解译、训练，获取针对水土保持的专业领域特化大模型，包含调用本地知识库调用模块、训练模型模块等。

(4) 模型层。模型层主要是智能助手的模型部分，里面负责接收用户的需求，并根据训练成果自动生成结果，并将结果以文字形式反馈至交互界面。

(5) 支撑层。支撑层主要是指支撑平台运行的存储设备、计算设备、网络工程、大模型支撑工具、运行环境等基础资源。

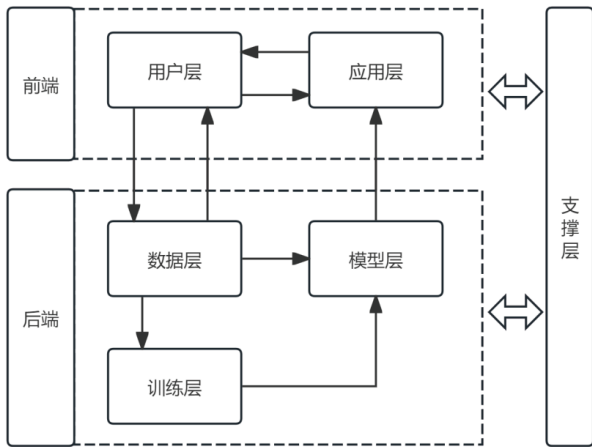


图1 水土保持智能助手

(四) 水土保持智能助手部署

1、部署Ollama与Deepseek本地模型。完成Ollama的安装后即可部署DS模型的R1版本（32B参数规模）。此步骤利用Ollama的简单命令接口，在本地电脑端部署Deepseek的R1版本模型上运行模型，无需额外服务器。

2、本地配置RAGflow。RAGFlow是一种融合了数据检索与生成式模型的新型系统架构，它能处理复杂查询，提升模型的准确性和上下文相关性，该系统主要包含数据检索模块和生成模块两个关键模块。

3、部署Docker。Docker是一组容器化引擎/Runtime产品。它基于操作系统层级的虚拟化技术，将软件与其依赖项打包为容器。Docker能够帮助开发者在轻量级容器中自动部署应用程序，并使得不同容器中的应用程序彼此隔离，高效工作。它确保RAGflow和模型在隔离环境中运行，提高稳定性。

4、修改并匹配配置文件及环境。进入RAGflow的安装目录内，进入docker子文件夹，找到.env配置文件修改部署的环境参数。

5、部署大语言模型。打开并运行Docker，并运行

Ollama，在Ollama的列表下输入Deepseek参数与地址完成Deepseek R1-32B的部署。

6、创建专业知识库。结合已经完成部署的RAGflow架构训练专业数据库。具体操作可以在RAGFlow中“创建知识库”，将相关数据库的文件上传/解析/分块并生成embedding并索引到文档引擎，完成专业知识库配置。在完成专业数据库的解析后，将训练完成的专业库信息载入到模型中。

7、用户库验证及前端结合。在完成好水土保持智能助手的训练和部署后，将相关数据和文件嵌入到前端平台上，在进行用户验证后，将用户的操作引导至水土保持智能助手的IP位置中，从而引导合法用户开展使用，同时由于Ollama、Docker及RAGflow的支持，相关的前端界面可以直接使用默认的对话界面。

四、训练与模型构建

本文在部署水土保持智能助手的关键在于模型部署、专业知识库解析以及将二者结合。本文在第三部分已经介绍了Deepseek R1模型的本地部署和专业知识库的嵌入。而关于专业知识库的搭建、解析的具体方法和步骤如下。

(一) 专业知识库的搭建

本文作者通过收集所在省份的公开公示的水土保持相关政策文件、法律法规、生产建设项目水土保持实施方案等资料构建了本地的专业库，具体情况如表3所示。

表3 专业库数据统计表

序号	类型	格式	数量	内存大小 (MB)
1	法律法规	pdf	5	100
		word	60	60
2	政策文件	pdf	15	96
		word	15	170
3	水土保持实施方案	pdf	10	248
		word	4	263
		txt	0	0
4	论文	pdf	21	184
		word	23	235
		caj	32	563
5	相关项目总结报告	pdf	10	500
		word	12	1452
总计			207	3871

(二) 专业库的解译与训练类型

在完成专业库配置后，进入RAGflow中通过Ollama调用Deepseek R1模型，再通过内置的工具对上述的文件

进行训练。训练的方式分为三种，分别为通过CPU开展训练解析、通过GPU开展训练解析以及通过Deepseek模型进行联网解析。具体情况如下：

1、CPU解析训练：通过模型调用本地电脑的CPU开展专业库解析，最终将训练专业库的成果形成本地知识库，提供给Deepseek R1模型调用更新。

2、GPU解析训练：具体步骤原理与CPU解析训练类似，但是解析专业库的硬件设备由CPU转成GPU，由于GPU具有更好的浮点运算与多线程运算的能力，GPU

解析训练的效率在理论上会优于CPU训练结果。

3、Deepseek在线解析：通过将文件上传至Deepseek的在线服务器中进行解析，该方法只能在单次对话中使用，无法将解析后的成果形成本地知识库供本地模型使用，且对于部分专业库数据存在泄密风险。

（三）专业库解析训练效果

为了对不同解译训练方法的效率、效果进行评估，本文在正式开展解译训练前利用不同类型、不同格式的脱密专业数据进行了训练，具体结果如表4所示。

表4 专业库解析训练测试结果

文件类型	文件格式	文件名	文件大小 (MB)	CPU解析时间 (秒)	GPU解析时间 (秒)	在线解析时间 (秒)
论文	pdf/caj	论文1	1.59	910	55	25
		论文2	1.12	601	31	9
政策文件	Pdf (扫描件)	生产建设项目水土保持信息化监管技术规定(试行)	3.33	失败	97	37
		湖南省生产建设项目水土保持监督管理办法	14.7	失败	77	30
各类报告	word	某工作总结报告	17.1	96	15	13
		某监管报告	178	105	29	失败

通过上表的结果可以看出：

1、三种解析方法中Deepseek在线解析的效率最高，GPU解析次之，CPU解析效率最低；

2、GPU文件解析类型最多解析效果最好，Deepseek在线解析存在文件容量的限制，CPU解析无法识别PDF扫描文件；

3、CPU和GPU的解译训练工作均在本地电脑端断网开展，解析训练后形成的本地知识库可以提供给本地模型读取调用，保密性和安全性高；

4、Deepseek在线解译必须联网将数据上传至Deepseek的服务器进行解析，每次在线解析的文件大小和数量均有限制，且解析成果只能保存在服务器端无法保存至本地，便于本地模型调用，该方法安全性、保密性均不理想。

5、三种解析方法对于Word的解析效率最高，对于文字格式的PDF解析效率次之，对于扫描的PDF文件识别解析效率最低。

因此，本文根据三种专业库解析训练的效果，从安全性、隐私性、效率等方面考虑，决定采用本地GPU解析训练专业库数据。并且为了提高效率，将尽量将相关文档资料转录成Word格式的文档后，再开展专业库的解

析训练工作。

五、结语

（一）存在的问题

本文紧扣数据的安全性、保密性需求，以Deepseek R1-32B为基础，通过搭建水土保持专业知识库，完成水土保持智能助手知识库的解析，从而部署服务于本地使用、运行的水土保持智能助手。通过对水土保持智能助手的部署和使用，本文发现了以下几个问题：

1、本地知识库的训练和解译对于图片、扫描文件识别速度较低；

2、32B参数的模型对于提出问题的解答存在幻觉；

3、水土保持智能助手训练用到的本地专业知识库的知识内容不够全面，且知识库的解析效率有待提高；

4、本地部署的服务器算力只能满足水土保持智能助手的基本运行，智能助手的反应和回答速度有待进一步提高。

（二）展望

本文所部署的水土保持智能助手是基于2025年开源的Deepseek R1-32B模型建立的，由于本地安全性的考虑，本文的水土保持智能助手要求必须部署在本地端。而受限于硬件设备的算力，本文只能采用32B参数的开

源模型进行部署，在知识库解析、对话、问题分析等方面还存在速度慢、AI幻觉问题。

而随着人工智能（AI）技术和计算机的算力正在不断地发展，今后的人工智能（AI）模型也必将更加完善。而本文通过部署水土保持智能助手，根据实践中发现的问题，得出以下经验。

1、在本地专业知识库的训练上，可以利用专业智能工具将扫描文件（PDF、图片等）先转为普通文档格式（Word、TXT等），再开始进行解析训练；

2、在知识库解析、模型运行时，可以通过代码设置，强制使用GPU开展运算，提高效率；

3、如果专业智能工具涉及安全、保密要求，则必须采用本地训练库、本地模型进行部署。如果不涉及相关要求，可以根据实际情况采用云端运行，提升运算速度，节省本地算力；

4、在部署各类型智能助手时，尽量采用专业的AI计算卡进行训练、模拟，并根据本地硬件的具体情况选择对应参数的模型，避免算力和运行内存的溢出，影响模型的运行效率。

希望上述经验可以为今后进一步完善水土保持智能助手提供参考，为人工智能、信息化技术等新技术在水土保持监管工作中的进一步利用提供帮助。

参考文献

[1]卢慧中，卞雪，金秋，等.基于深度学习的生产建设项目扰动图斑提取算法和识别策略[J].中国水土保持，

2024，（08）：23-28.DOI：CNKI：SUN：ZGSB.0.2024-08-007.

[2]赖杭，张永占，赵旭升，等.深圳市水土保持智慧监管平台研发及应用[J].水利信息化，2024，（03）：90-96.DOI：10.19364/j.1674-9405.2024.03.016.

[3]南帝，杨宇，杨植林，等.智慧水利水土保持分系统构建及应用[J].水上安全，2024，（06）：7-9.DOI：CNKI：SUN：SSXF.0.2024-06-003.

[4]陈妮，赵勇，杨凯，等.新昌县水土保持监管数字化改革探索与实践[J].中国水土保持，2023，（03）：62-65.DOI：10.14123/j.cnki.swcc.2023.0058.

[5]Abhilash S, J. A, Jaiprakash N, et al. A deep learning approach to predict the number of [formula omitted]-barriers for intrusion detection over a circular region using wireless sensor networks[J]. Expert Systems With Applications, 2023, 211 DOI:10.1016/J.ESWA.2022.118588.

[6]Matthew S, Christophe M, Cristián B. The value of text for small business default prediction: A Deep Learning approach[J]. European Journal of Operational Research, 2021, 295(2): 758-771. DOI:10.1016/J.EJOR.2021.03.008.

[7]白贺伊.基于卷积神经网络的健康大数据智能分析方法研究[J].电子设计工程，2021，29（10）：10-14. DOI：10.14022/j.issn1674-6236.2021.10.003.

[8]李刘霞.陕西清涧县水土保持监管现状与信息化技术应用[J].农业工程技术，2020，40（32）：47-48. DOI：10.16815/j.cnki.11-5436/s.2020.32.027.