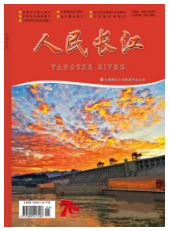

人民长江
Yangtze River
ISSN 1001-4179,CN 42-1202/TV

《人民长江》网络首发论文

题目：基于水利一张图的地理空间信息问答智能体技术

作者：明晨曦，杨鹏，张志鑫，刘哲，乔延军，李杰潘

网络首发日期：2025-03-20

引用格式：明晨曦，杨鹏，张志鑫，刘哲，乔延军，李杰潘．基于水利一张图的地理空间信息问答智能体技术[J/OL]．人民长江.
https://link.cnki.net/urlid/42.1202.TV.20250320.1506.004

基于水利一张图的地理空间信息问答智能体技术

明晨曦[1,2,4]，杨鹏[1,2,4]，张志鑫[1,2,4]，刘哲[1,2,4]，乔延军[1,2,4]，李杰潘[3]

(1.长江委网络与信息中心，湖北 武汉 430010；2. 长江委流域管理数字赋能技术创新中心，湖北 武汉 430010；3. 武汉大学测绘遥感信息工程国家重点实验室，湖北 武汉 430072；4. 长江委智慧长江创新团队，湖北 武汉 430010)

摘要：当前，生成式人工智能技术，特别是自然语言大模型的兴起，为地理空间信息获取提供了新的思路。然而，现有的大模型和搭载大模型的智能体技术，主要基于通用数据集上训练的，当应用于在地理空间信息上的问答时，容易出现幻觉，存在回答内容相关度较低、回答不准确、缺乏实时性等问题。针对该问题，本文提出基于水利一张图的地理空间信息问答大模型智能体技术框架（GLMA），该技术框架主要利用智能体，理解用户的自然语言提问，驱动大模型执行水利一张图任务，再利用大模型融合提问和任务的返回结果，生成最终的提问回复内容。为了提高任务执行的准确性与有效性，在任务分配阶段，本文使用了树形的任务分配结构，有效地提高了任务检索和参数生成的能力。此外，为了验证 GLMA 的有效性，本文构建了一套地理空间信息问答数据集，并设定了相应的评估指标。在与最新的中文开源大模型 Baichuan2、Llama3.1、ChatGLM4、Qwen2.5 等的对比测试中，GLMA 在任务分配准确率和查询结果准确率等评估指标上取得了最好的效果。本研究具有一定的扩展性，将为其它业务领域的大模型智能体研究奠定基础。

关键词：智能体；大模型；地理空间信息；水利一张图

中图分类号：TP183　　**文献标识码**：A

Research on large model agent technology of geospatial information question answering based on water conservancy map

MING Chenxi[1,2,4], YANG Peng[1,2,4], ZHANG Zhixin[1,2,4], LIU Zhe[1,2,4], QIAO Yanjun[1,2,4], LI Jiepan[3]

(1. *Network & Information Center, Changjiang Water Resources Commission，Wuhan，Hubei 430010，China*；
2. Center of Technology Innovation for Digital Enablement of River Basin Management，Wuhan，Hubei 430010，China；
3. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing，Wuhan University，Wuhan；
4. Smart Changjiang Innovation Team of Changjiang Water Resources Commission)

Abstract：At present, generative artificial intelligence technology represented by natural language large model has ushered in vigorous development, providing a new way to obtain geospatial information. However, existing large models and agents equipped with large models, most of which are trained on general datasets, are highly prone to issues such as hallucinations, such as low correlation of answer content, inaccurate answer and lack of real-time performance. To address these issues, this paper proposes a technical framework for a large language model intelligent agent for geospatial information Q&A based on “one map” of water conservancy (GLMA). This technical framework integrates the agent's natural language understanding capabilities, enabling it to accurately grasp the user's question intent and subsequently drive the large model to execute specific "Water Resources Single Map" tasks. To enhance the accuracy of the task allocating, a tree-like structure task

基金项目：国家自然科学基金（42271370），长江委流域管理数字赋能技术创新中心基金
作者简介：明晨曦，女，硕士，研究方向为自然语言处理、人工智能应用等。E-mail: mingchenxi90@163.com
通讯作者：张志鑫，男，硕士，研究方向为自然语言处理、遥感智能分析与应用。E-mail: zhixinzhang@niccwrc.cn

allocation system is used in the task dispatch stage, significantly improving the capabilities for task retrieval and parameter generation. Additionally, to validate the effectiveness of GLMA, this paper constructs a dataset for geospatial information question answering and establishes corresponding evaluation metrics. Compared with state-of-the-art Chinese open-source models such as Baichuan2, Llama3.1, ChatGLM4, and Qwen2.5, GLMA achieves the best results in terms of task allocation accuracy and query result accuracy. This research possesses significant extensibility and will lay the foundation for further studies on large model agents in other business domains.

Key words：AI agent; large language model; geospatial information; "One map" of water conservancy

0　引言

近年来，随着水利行业的数字化转型和智能化升级，水利地理空间信息作为行业发展的核心资源，其重要性日益凸显[1]。水利地理空间信息不仅涵盖了水库、河流、湖泊等自然水体的空间分布信息，还涉及水位、流量、降水量等动态监测数据，是水资源、防洪抗旱、水环境等工作的基础支撑。然而，尽管当前的地理空间信息服务系统在技术上能够支持多样化的数据检索与分析任务，却普遍存在用户界面繁杂、操作流程复杂以及对使用专业术语时的准确性要求高等问题。用户往往需要具备一定的专业知识背景，熟悉系统的各项功能和操作逻辑，精准地使用行业术语，才能有效地进行数据查询。这种技术门槛不仅限制了非专业人士的使用，也降低了信息获取和使用的效率和便捷性。因此，如何在水资源配置、水生态管理、防洪抗旱、水域空间规划等实践中帮助人们更准确、更快速、更便捷地获取和使用水利地理空间信息，提升水利管理效率、优化水资源配置、增强灾害应对能力，成为亟待解决的问题。开展水利行业地理空间信息智能问答技术的研究成为解决该问题的有效途径。

在智能问答技术上，2023 年以来，以 ChatGPT[2-3]为代表的生成式大语言大模型技术迎来了蓬勃发展，不同领域、不同参数规模的大模型层出不穷[4]。在水利行业领域中，围绕大模型技术中也有如建设水利行业大模型[5]、大模型在水利系统服务中的应用[6]等方向的研究与探索。大模型具有强大的语义理解能力，能理解对用户输入的自然语言提问，并生成相应的自然语言回答，给用户提供了更好的智能化问答体验。但大模型在问答应用中也存在着许多挑战：大模型的回答问题所依赖的知识来源，即训练数据，大部分为通用领域知识，在水利行业等专业领域内的知识量较为缺乏；并且，由于大模型的训练成本高、时间长，其训练频次往往较低，因此大模型回答问题所依赖的知识来源，即训练数据的时效性较低，尤其在水利地理空间信息方面，缺乏实时性。而当大模型面对自己知识储备以外的提问时，容易出现“幻觉”问题，即“答非所问”的不相干回答，或相关但不准确的错误答案[7]。

随后，为了解决上述问题，一系列基于大模型的技术框架应运而生，如检索增强技术[8-9]、智能体技术[10-11]等。其中，智能体技术通过大模型理解用户通过自然语言发出的指令，根据这些指令自动规划并执行相应任务，极大地提升了工作效率与用户体验。它们展现出了卓越的语义解析能力，能够精准捕捉用户意图，并生成相应的任务，同时支持多步骤、上下文关联的任务执行功能，为用户带来前所未有的智能化交互体验[12]，这为地理空间信息智能问答技术的研究提供了启发。然而，目前业内已有的大模型智能体，其支持的任务类型主要仍集中在通用领域的简单任务，如网络搜索、天气查询、图像生成等，在水利地理空间信息相关的任务上鲜有成果。

针对上述问题，本文提出了一种基于水利一张图的地理空间信息问答大模型智能体框架。水利一张图系统是针对水利信息化建设研发的综合性地理信息服务系统，它整合了河流、湖泊、水库等水利管理对象的空间特征，以及水旱灾害防御、水资源管理等水利业务的需求实现了水利地理空间信息的一体化组织、管理和应用[13-14]。本文所提出的框架将大模型智能体技术应用于水利一张图系统，通过理解用户输入的自然语言提问或查询语句，自动执行一张图系统内的相应任务，并生成最终回答，实现了地理空间信息领域的智能知识问答。

该框架主要由四个部分组成：任务构造、任

务分配、任务执行、问答生成。任务构造部分将水利一张图系统中的底层“元任务”，按照业务需求进行组合，构造出具有独立业务需求的“任务”；任务分配部分智能体通过理解用户输入的自然语言指令，逐级进行任务检索，得到和用户指令最相关的任务，并形成标准化的任务输入参数；任务执行部分根据输入参数，依次执行该任务中的元任务，每个元任务执行结束时，发起用户交互提示，接收用户反馈后继续执行，直到得到最终结果；问答生成部分由智能体利用用户指令和任务执行结果，进行总结归纳后生成答案。本文的创新点主要包括三个方面：

（1）针对生产实践中对地理空间信息的使用需求，本文提出了一种基于水利一张图的地理空间信息问答智能体框架，可提供支持自然语言输入的地理空间信息问答服务，提升了地理空间信息获取和使用的便捷性。

（2）针对大模型智能体执行任务的准确性和时效性等问题，本文提出了一种基于水利一张图的实时性任务精准分配和执行机制，包括任务参数的精准生成和支持多结果用户反馈确认的机制，有效地提升了智能体任务执行的准确性和有效性。

（3）创建了一套水利地理空间信息问答的数据集和一套面向大模型智能体效果评估的指标，在该数据集和评估指标上与最新的中文开源大模型 Baichuan[15]、ChatGLM4[16]、Llama [17] 、Qwen[18] 等进行对比，本文提出的地理空间信息问答智能体在评估指标准确率上取得了最好的效果。

1　整体流程

本文提出的地理空间信息问答智能体整体框架如图 1 所示，整个框架共分为四个部分：任务构造、任务分配、任务执行和问答生成。任务构造部分将水利一张图系统中执行任务的基本工具（本文中称为“元任务”）进行组合，得到具有独立业务意义的“任务”。任务分配部分通过理解用户输入的自然语言指令，利用树形结构任务分配系统进行逐级检索，得到和用户指令最相关的任务，并形成标准化的任务输入参数。任务执行部分根据输入参数，依次执行该任务中各个元任务，每个元任务结束时发起用户交互提示，接收用户反馈（确认或改选其他候选结果）后继续执行，直到得到最终结果。问答生成部分将用户指令和任务执行结果输入大模型，进行总结归纳后生成答案。

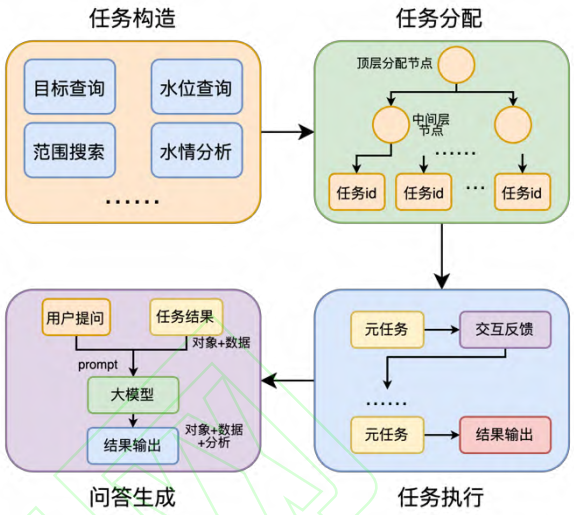

图 1　基于水利一张图的地理空间信息问答大模型智能体框架

Fig.1　Overall framework of the GLMA

1.1　任务构造

任务构造是根据业务需求，完成该需求所执行的水利一张图中的各项基本功能的组合，构造成独立的任务。任务构造部分主要有元任务标准化、任务搭建两个阶段，下面分别对其进行介绍。

（1）元任务标准化

首先，针对独立的业务需求，获取完成该需求所需的水利一张图的元任务，将其结构进行标准化，形成包括元任务 id，元任务功能、外部输入参数、内部输入参数、输出参数在内的标准化结构。

其中，外部输入表示该输入参数的最终来源是用户、智能体等外部对象，在实际运行中，将在任务分配阶段，由智能体统一生成，如地名关键词、时间信息等；内部输入表示该输入参数是一张图系统内部自定义的、并未向普通用户公开的信息，来源是水利一张图系统中其他任务，在实际运行中，将在任务执行阶段，由其前置任务的输出参数提供。

表 1 水利一张图元任务属性表

Tab.1 Attributes of meta-tasks on “one map” platform

属性	说明	要求	元任务示例 1	元任务示例 2
id	元任务的唯一标识	具有唯一性	101	102

属性	说明	要求	元任务示例 1	元任务示例 2
元任务名	元任务的名称		查询测站编码	查询测站某时刻水位
外部输入参数	通过用户输入得到的参数	应为用户可以得知的通用信息，如地名、时间等	关键词	查询时刻
内部输入参数	由其他元任务输出得到的参数	用户不可见的一张图内部参数，如编码、链接等		测站编码
输出	元任务执行结果		测站编码	测站对象，水位信息

元任务包含的属性信息如表 1 所示。表中给出了两个元任务示例，其中，元任务 101 是根据关键词查询测站编码，外部输入为关键词，输出为测站编码；元任务 102 是根据测站编码和时间范围，查询该测站的水位信息，外部输入为查询时刻，内部输入为测站编码，输出为测站对象和水位信息。

（2）任务搭建

任务搭建的核心目的，是通过有序组合不同的“元任务”，得到具有以下特点的“任务”：

1）执行任务所需的所有输入信息，均能通过用户提问的自然语言得到，即均为上文中提到的外部输入参数。

2）包含独立的业务意义，即包含一个核心地物，以及围绕该地物进行的查询和分析的结果，如查询某地物的地点位置、查询某测站的水位等。

表 2 水利一张图任务属性表

Tab.2 Attributes of tasks on “one map” platform

属性	说明	要求	任务示例
id	任务的唯一标识	具有唯一性	1
描述	任务的详细信息	使用自然语言描述任务，包含任务功能、输入参数等的详细描述	根据用户输入的地点关键词和时间，查询测站在该时刻的水位
元任务列表	任务中包含的元任务	元任务按执行顺序组成的列表，所有元任务中的内部输入参数(如果有的话)都能通过其前置任务的输出得到	[101, 102]
输入参数	任务的输入参数	元任务列表中所有元任务的外部输入参数集合	关键词，查询时刻
输出	任务的输出结果	一个或多个对象组成的列表，每个对象由两个属性组成： 1. 水利一张图中的目标地物对象； 2. 该目标地物的查询结果数据。	测站对象，水位信息

具体地，表 2 给出了一个任务包含的属性及说明，并给出了一个水位查询任务示例。该任务通过组合元任务 101、102 得到，其中执行元任务 101 仅需要外部输入关键词，而执行元任务 102 所需的内部输入，即测站编码，可通过执行元任务 101 得到。在输出方面，该任务输出一个测站对象作为核心地物，并输出相应的水位信息，具有独立的业务意义。

本文根据以上方法，构造以下了 10 类任务：

1）根据关键词，查询目标地物的位置。

2)根据关键词和类型,查询目标地物的位置。

3）根据测站名称，查询当前的水位，并分析水情。

4）根据测站名称和目标时刻，查询该时刻的水位并分析水情。

5）根据测站名称和目标时间段，查询该时间段内的水位信息，并分析水情。

6） 给定一个中心地物和目标类型及查询范围，查询位于中心地物周边范围的所有属于目标类型的地物。

7） 给定一个中心地物和目标类型及查询范围，统计位于中心地物周边范围的所有属于目标类型的地物的数量。

8）查询离目标地物最近的河流。

9）给定目标河段和目标时间段，查询该河段在该时间段的取水总量。

10）给定上下游地点关键词和目标时间段，查询上下游之间河段在该时间段的取水总量。

1.2 任务分配

任务分配是大模型智能体根据用户输入的指令，检索出和用户指令最相关的任务，并生成具体的任务输入参数的过程，智能体的任务分配结构如图 2 所示。

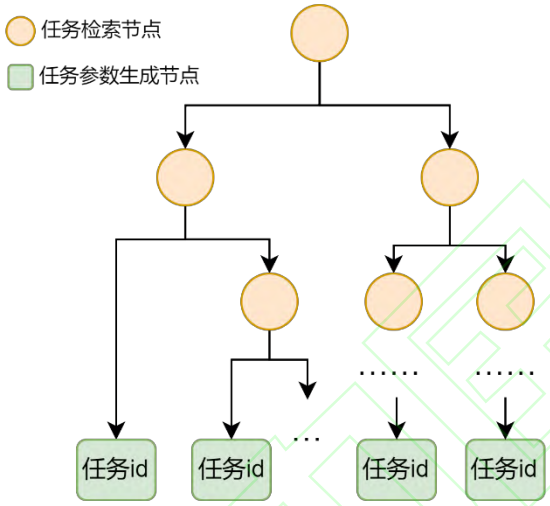

图 2　智能体任务分配结构图

Fig.2　Structure of agent task assignment

整个任务分配系统呈树形结构，每个节点中包含两部分内容：一个任务或任务集合、一个驱动大模型进行任务分配的提示词（prompt）。其中，叶子节点中包含的一个任务；非叶子节点中包含一个任务集合，具体地，为其所有子节点对应任务或任务集合的并集，表示如下：

$$T_{n_i} = \bigcup_{n_j \in C(n_i)} T_{n_j} \quad (1)$$

其中n_i为非叶子节点，T_{n_i}为节点n_i对应的任务集合，$C(n_i)$为节点n_i的子节点集合。

任务分配过程分两个阶段，第一阶段为任务检索，对应每个非叶子节点执行的操作，目的是逐级检索得到更加精准的待执行任务范围，最终确定一个和用户输入的自然语言指令最相关的任务 id；　第二阶段为任务参数生成，对应每个叶子节点执行的操作，目的是根据检索到的任务 id，结合输入指令进行自然语言理解，生成具体的输入参数。本文提出的这种两个阶段、多层树形结构的系统，通过智能体对自然语言指令进行多次理解，最终得到精准的任务信息，即任务输入参数。下面分别介绍这两个阶段。

（1）任务检索

任务检索的核心思想，是利用大模型智能体的自然语言理解能力，结合用户指令和各个任务的信息，为用户指令匹配到最相关的任务 id。

具体地，对于每一个非叶子节点，都进行一次检索操作，根据用户指令和当前节点所对应的 prompt，驱动大模型理解用户指令和任务信息，返回和用户指令最接近的一个或一类任务。这一过程主要包含两个方面：一方面基于任务功能驱动的提示词构造方法，为任务分配系统设计了一套提示词模板；另一方面基于大模型多轮次对话的生成式问答方法，联合用户指令与任务功能描述信息，引导大模型的问答生成。

基于任务功能驱动的提示词构造方法方面，针对任务功能特性构造了一套提示词模板引导大模型输出的如下提示词模板：

“你是一个帮助用户解决问题的助手，你的职责是对用户输入的问题，进行第一步的任务分配工作。

你需要将每个问题准确地分配到一个具体的任务中。注意一个问题只对应一个任务，要确保这个任务是最能解决问题的那一个。请输出对应任务的序号。

<可选任务>{task_list}</可选任务>”

针对每个非叶子节点，应用该模板，形成该节点对应的提示词。其中，task_list 为任务描述列表，长度为当前节点所有子节点个数，每个任务描述即为其对应的子节点所包含的任务或任务集合的描述。例如，根节点包含 3 个子节点，其对应的 task_list 为：

“1. 查询地物信息。

2. 查询水位信息。

3. 查询取水量信息。”

基于大模型多轮次对话的生成式问答方面，结合用户指令和每个节点的提示词，构造如下的大模型问答指令：

[{“role”:“system”,“content”:system_promp

t},

{“role”: “user”, “content”:query}]

其中，system_prompt 为对每个非叶子节点构造的任务检索提示词，设置该提示词的对话角色为“系统”（system），输入大模型；query 为用户输入的指令，设置 query 的对话角色为“用户”（user）。将大模型问答指令输入大模型，引导大模型输出当前节点的检索结果。

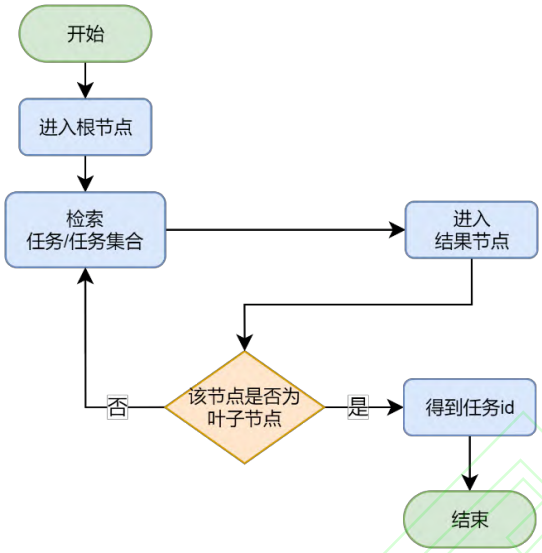

图 3 任务检索流程图

Fig.3 Task Retrieval Flowchart

在接收到用户输入的指令后，任务检索流程如图 3 所示。从根节点开始（这意味着，当前所有任务均为候选任务），依次进行当前节点的检索操作，检索结果对应当前节点的一个子节点，任务检索流程随即进入该子节点，候选任务范围缩小至该子节点所对应的任务集合。重复以上流程，直到进入的子节点是叶子节点，则表示已经确定需要执行的任务 id。此时任务检索阶段结束，进入任务参数生成阶段。

（2）任务参数生成

在任务检索阶段确定了需要执行的任务 id 后，进入任务参数生成阶段。这一阶段的核心思想，是是利用大模型智能体的自然语言理解能力，结合用户指令和目标任务的详细参数要求，从用户指令中精准提取执行目标任务的实际参数值。

具体地，对于每一个叶子节点，都根据用户指令和当前节点所对应的任务的参数要求构建大模型提示词（prompt），驱动大模型理解用户指令和任务参数要求，返回任务执行所需的实际参数值。这一过程主要包含两个方面：一方面基于任务功能驱动的提示词构造方法，为任务参数生成系统设计了一套提示词模板；另一方面基于大模型多轮次对话的生成式问答方法，联合用户指令与任务功能描述信息，引导大模型的问答生成。

基于任务功能驱动的提示词构造方法方面，针对任务的参数信息构造了一套提示词模板引导大模型输出的提示词模板。以查询经纬度任务为例，提示词如下：

“你是一个帮助用户进行根据关键词和类型，查询对象的经纬度**任务的助手，你的职责是针对用户输入的问题，提取执行任务所需的参数。**

该任务的具体内容及参数如下：

{"函数名": "经纬度查询",

"功能描述": "根据关键词和类型查询要素详细信息",

"参数":

{"关键词": {

"类型": "string",

"参数描述": "用于检索要素名称的关键词。"},

{"类型": {

"类型": "string",

"参数描述": "要素类型，请务必从枚举值中选取",

"枚举值": "水库,取水口,桥梁,河段,测站,码头"}}}}

———————— 示例 ————————

<输入>长江大桥在哪里</输入>

<输出>

请执行Action:{"action_name":"经纬度查询","args":{"关键词":"长江大桥","类型":"桥梁"}}

</输出>

———————— 示例结束 ————————”

提示词中给出了任务描述、任务各个参数的详细信息，包括参数名、参数描述，以及枚举类型的参数给出了各项枚举值等，详尽的描述使大模型尽可能地从用户输入的指令中准确提取任务所需要的参数值。示例中给出了输出的任务参数的统一格式，本文以 json 格式组织任务参数，便于后续任务执行阶段统一解析。

基于大模型多轮次对话的生成式问答方面，结合用户指令和每个节点的提示词，构造如下的大模型问答指令：

[{“role”:“system”,“content”:system_prompt},

{“role”: “user”, “content”:query}]

其中，system_prompt 为对每个叶子节点构造的任务参数生成提示词，设置该提示词的对话角色为“系统”（system），输入大模型；query 为用户输入的指令，设置 query 的对话角色为“用户”（user）。将大模型问答指令输入大模型，引导大模型输出实际的任务参数。

本文提出的这种树形结构的任务分配方式，本质是将任务和参数的特征信息进行提炼并分离，将不同的任务按照共同点分配在同一个非叶子节点中，分多次检索来识别不同的特征、逐步确定目标任务。这种方式使得智能体在每一次检索过程中，所需要区分的任务特征尽可能地聚焦，使大模型专注于对少量特征的区分，而无需关注任务中的其他细节，从而提高检索的准确性。

本文提出的树形结构从根节点至叶子节点总共包含 4 层（后文中，如无额外说明，则默认任务分配方式为 4 层树形结构）。此外，在 3.3 节的消融实验中，进行了任务分配方式的消融实验，逐渐删减树形结构的层数，分别对比了 1 层至 3 层的树形结构任务分配方式下的大模型智能体效果。其中，1 层结构即为将所有任务和参数信息放在一个 prompt 中，驱动大模型一次找到最终的任务 id 并生成相应参数。实验结果如表 9 所示。

1.3　任务执行

任务执行部分，接收任务的初始参数，结合水利一张图的信息展示功能，顺序执行任务中的各个元任务，直至得到最终结果。任务执行流程如图 4 所示。

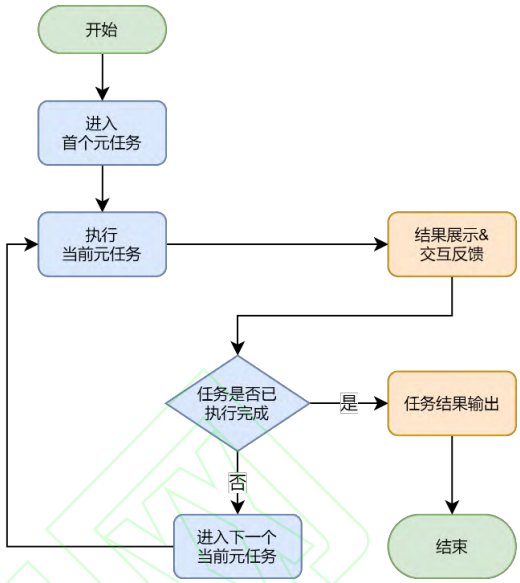

图 4　任务执行流程图

Fig.4　Task Execution Flowchart

任务执行开始时，系统接收任务的初始参数，从该任务所包含的元任务列表中的首个元任务开始，依次顺序执行每个元任务。每个元任务执行完成后将产生交互提示，并结合水利一张图的展示功能（如：定位功能、展示信息面板等），在前端向用户展示。当元任务产生多个可选结果，时，选取前 5 个结果目标在前端，并向用户提供选择按钮，供用户选择查看每一个结果目标的详情，并选择一个确认后继续执行任务，直至产生最终结果。

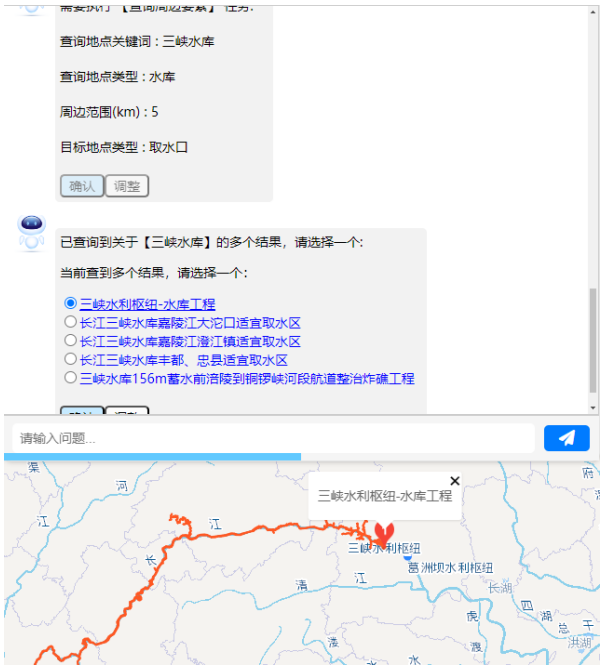

图 5 任务执行前端界面示例

Fig.5　Word vector keyword joint retrieval method Example of Task Execution Front-end Interface

图 5 为用户输入提问“查找三峡水库附近的取水口”任务执行的前端界面示例，在根据关键词“三峡水库”查找一张图系统中的三峡水利枢纽目标时，搜索出了多个目标结果，此时用户可依次查看每一个结果，查看过程中系统会自动定位到该目标，并显示该目标的信息面板，用户可根据多个目标信息，选择正确的目标。

1.4　问答生成

问答生成是利用大模型的对话生成能力，根据任务执行的结果，生成最终的针对用户指令的回答的过程。

1.3 节中所得到的任务结果为包含一个或多个对象列表的结构化数据，数据量大、可读性不强、不包含总结分析结果，对于回答用户提问并不友好。因此，本文使用了基于大模型的问答生成方法，结合用户指令和任务结果，生成由自然语言总结提炼的、易于理解的回答。具体地，构建以下的提示词模板，指导大模型输出回答：

“<指令>根据提问和已知信息，对问题进行简洁的归纳和回答。</指令>

<要求>1. 请自行解析已知信息中的结构化内容，转换成自然语言来回来问题。2. 请严格根据已知信息来回答内容，不要添加编造的内容。</要求>

<提问>{query}</提问>

<已知信息>{context}</已知信息>”

其中，“{query}”在实际使用中将被替换为用户提问内容文本，而“{context}”则被替换为任务执行结果（通常为 json 格式的数据文本），将二者融合提示词一起作为大模型的输入，驱动大模型理解问题和结构化的任务结果数据，输出最终的问答结果。

2　问答数据集

2.1　数据集来源

为了客观评估本文提出的任务执行应用的效果，本文提出了首个面向水利一张图任务的任务执行问答数据集。参考了地理空间问答数据集 GeoAnQu[19]、GeoQuestions[20]的构造方法，本文的数据集主要通过以长江流域内的码头、桥梁、取水口、测站等地物为核心，构建出围绕这些地物的提问和查询结果。问答所涉及核心地物来源于数据来源于长江委中心数据库，主要组成为各业务应用数据、水利部及长江委承担的各建设项目接入的数据以及长江流域各省份汇集的数据，数据准确、真实、可靠。通过构建数据集，能够提供关于地物目标的准确地理空间信息，为一张图任务执行应用的评估提供可靠的基础，同时也可为地理空间信息问答领域的相关研究提供参考和借鉴。

2.2　问答对构建

问答对构建流程如下：

（1）数据收集。从长江委中心数据库中收集水库、测站、码头、桥梁、河流、湖泊等地物数据信息，信息内容包括地物编码、名称、类别等，确保每一个地物有其唯一标识、类别和精准的位置信息。

（2）问题生成。按照 1.1 节中的任务类型，生成模拟用户查询时的问题语句。在这一过程中，随机选择收集到的地物数据作为问题中所涉及到的核心地物，确保问题的有效性。同时，使用大模型进行问题的批量生成，使查询语句的表达口语化、简洁化，贴近人工提问的真实表达。

（3）问答对构建。针对生成的查询问题和核心地物信息，在一张图系统中使用相应元任务进行查询，得到准确的结果数据，结果数据由以下内容构成：

1）核心地物：提问中的核心地物在一张图系统中对应的目标对象，包含名称、类别、经纬度等基本信息。

2）需执行的任务及参数：根据提问和核心地物，人工标注回答提问所需要执行的任务，及任务所需参数。

3）查询结果：根据提问和核心地物，在一张图系统中进行人工查询，得到查询结果。对于查询意图类别为数据的任务，回答中应包含数据查询结果；对于查询意图类别为其他地物的任务，回答中应包含查询结果地物在一张图系统中对应的目标对象，包含对象编码、名称、类别、经纬度等基本信息。

将提问和结果数据组织成问答对的形式，每

个问答对包括一个问题和对应的准确回答。这些问答对将用于评估任务执行问答应用的效果。

通过上述流程，本文提出了由 500 个一张图任务问答对组成的数据集，问答数据集的部分样例说明如表 3 所示。

表 3 问答数据集样例

Tab.3 Samples from The Q&A dataset

序号	问题	回答说明
1	武汉长江大桥的具体位置在哪里？	1．核心地物，武汉长江大桥一张图对象：{"objname":"武汉长江大桥","type":"桥梁","lat":114.282359,"lon":30.554291} 2．任务：地点查询，参数：{"关键词":"武汉长江大桥"} 3．查询结果，即武汉长江大桥一张图对象
2	丹江口库区现在的水位是多少？	1．核心地物，丹江口水库水位站对象：{"objname":"丹江口水库","type":"测站","lat":……,"lon":……} 2．任务：单一时刻水位查询，参数：{"关键词":"丹江口","时刻":"2024-10-15 15:00"} 3．查询结果：该水位站对象在 2024 年 10 月 15 日 15 时(提问的当日当时)的水位信息
3	三峡水库周边 5km 有取水口吗？	1．三峡水库的一张图对象，包含：{"objname":"三峡水利枢纽-水库工程","type":"水库","lat":……,"lon":……} 2．任务：查询周边指定类型的地物，参数：{"关键词":"三峡水库","目标地物类型":"取水口","范围(km)":"5"} 3．查询结果，即周边排污口对象列表：[{"objname":"堰沟水库取水坝 ","type":"取水口","lat":……},{"objname":"抱丰村碗场坪平溪沟报丰水厂取水管道","type":……},{"objname":"兴山县高桥乡洛坪村堰塘坪农村集中供水工程取水口",……}]
4	新乡自来水厂离哪条河最近？	1．新乡自来水厂对象：{"objname":"南昌县南新乡自来水厂","type":"取水口",……} 2．任务：查询最近的河流，参数：{"关键词":"新乡自来水厂"} 3．查询结果：赣江中支
5	三峡到湖口的取水口在上个月的取水总量是多少？	1．核心地物：三峡到湖口河段 2．任务：取水量查询，参数：{"关键词":"三峡到湖口","起始时间":"2024-09-01","终止时间":"2024-09-30"} 3．查询结果：该河段在 2024 年 9 月 1 日至 9 月 30 日(提问当日的上个月)的取水总量

3　实验评价

3.1　实验设置

（1）实验环境

本节的所有实验采用 python 搭建智能体服务，使用从开源模型社区 modelscope 中下载的开源大模型进行本地部署。实验中模型的部署和推理使用 5 张 GeForce RTX 4090 消费级显卡和 4 张昇腾 910B 训练卡，并且采用了单机多卡并行计算的推理方式提高计算效率。智能体问答测试服务运行在以 Linux 为操作系统的高性能计算环境中，包含 24 个 CPU 核心、256GB 运行内存、17TB 存储空间。

（2）实验方法及评估指标

大模型智能体的效果评估实验主要通过构建的问答对数据集，对智能体根据输入的信息生成的输出内容进行人工评估打分的方式完成。在评估指标上，本文参考了大模型的人工测评方法[21-22]，从提问和智能体服务的特性出发，从任务和参数准确性、查询结果准确性两个方面构造评估指标，每种指标的评价分数均为 5 级，具体的评分规则如表 4 所示。

表 4 问答模型评估标准

Tab.4　Evaluation Criteria for Question-Answering Models

分数	任务和参数准确性	查询结果准确性
5 分	模型分配的任务和提问高度相关，生成的参数完全准确	能在无任何额外交互的情况下准确查询到核心地物和最终结果

分数	任务和参数准确性	查询结果准确性
4 分	模型分配的任务和提问高度相关，生成的参数部分准确	能准确查询到核心地物和最终结果，但中间过程中存在正确选项未排在当前结果列表首位的情况，需要额外的交互反馈
3 分	模型分配的任务和提问高度相关，但生成的参数完全不准确	能查询到正确的核心地物，最终结果基本正确，但不完整或有部分错误数据
2 分	模型分配的任务和提问具有一定相关性，能部分解决问题	能查询到正确的核心地物，但最终结果不准确
1 分	模型分配的任务和提问相关性低，无法解决问题	无法查询到正确的核心地物

3.2　对比实验

本文将提出的智能体同最新的中文大模型进行对比实验。实验从两个方面进行，一方面使用智能体对比不同型号的大模型，在直接输入数据集中的提问时的回答效果。在模型选择上，由于不同型号的模型参数规模并不完全一致，因此选择参数规模接近，范围在 7B~9B 的大模型。具体地，本文用来进行对比的大模型包括 GLM4-9B-Chat 、 Meta-Llama-3.1-8B-Instruct 、 Baichuan2-7B-Chat 、Qwen2.5-7B-Chat，以及搭载了 Qwen2.5-7B-Chat 的大模型智能体。在实验中，将问题分为地物类查询任务，如查询地物位置、查询周边地位等，和数据类查询，如查询水位、查询取水量等任务等，分别评估模型在查询结果上的准确率。

实验结果如表 5 所示。其中，只使用大模型进行问答时，对于部分地物类查询问题，核心地物是常识性知识时，如“洞庭湖的位置在哪里？”，大模型能给出基本正确的回答；数据类查询问题，如“汉口水位站昨天的水位是多少？”，由于需要查询专业机构的数据，大模型几乎不能做出有效回答。在各个大模型中，Qwen2.5-7B-Chat 整体回答准确率相对最高，并且在使用了搭载大模型的智能体后，地物类查询任务回答效果进一步明显提升，数据类查询任务也得到有效执行，整体平均得分 4.09/5。。

表 5　不同型号的模型问答对比结果

Tab.5　The evaluation result of question answering for different model types

模型	查询结果准确率		
	地物类查询	数据类查询	平均分
Baichuan	1.45	1.00	1.25
Llama	1.59	1.02	1.34
ChatGLM	1.69	1.00	1.39
Qwen	1.73	1.00	1.41
Qwen+Agent	**4.07**	**4.11**	**4.09**

另一方面，使用智能体对比相同型号、不同参数规模的大模型，具体地，本文使用了 Qwen2.5-7B-Chat 、 Qwen2.5-14B-Chat 、 Qwen2.5-32B-Chat、Qwen2.5-72B-Chat，以及搭载了 Qwen2.5-72B-Chat 的大模型智能体进行对比实验。实验结果如表 6 所示。随着模型参数规模的增加，模型对地物类查询任务的准确性逐渐增加，对数据类查询任务始终无法有效执行。在使用了智能体框架后，地物类查询任务准确率进一步提升，数据类查询任务也得到有效执行，整体平均得分 4.51/5。图 6 给出了使用智能体进行提问和回答的样例。

表 6 不同参数规模的问答模型对比结果

Tab.6　The evaluation result of question answering for models with different parameter scales

模型 (Qwen2.5 系列)	查询结果准确率		
	地物类查询	数据类查询	平均分
7B	1.73	1.00	1.41
14B	1.77	1.00	1.43
32B	1.93	1.00	1.56
72B	2.14	1.00	1.64
72B +Agent	**4.71**	**4.27**	**4.51**

图 6　智能体问答样例

Fig.6　A question answering sample of GLMA

3.3 消融实验

消融实验的作用是验证本文所提出的智能体模型的关键性能和特征。在本文所提出的智能体模型中，通过大语言模型驱动智能体进行任务分配与执行，消融实验可以在大语言模型上分别替换不同系列和规模的大模型，来评估大模型对智能体的性能影响，这有助于理解不同大模型特性的优劣，更好地选择、优化大模型和配置，提升智能体的整体决策能力。

本文针对智能体中的大模型驱动部分，分别从大模型型号、大模型规模和任务分配方式三个方向进行了消融实验。下面分别介绍这三个方向的实验。

首先，在针对模型型号的消融实验中，本文分别使用 Qwen2.5-7B-Chat、GLM4-9B-Chat、Meta-Llama-3.1-8B-Instruct 、 Baichuan2-7B-Chat 来替换智能体中的大模型，并且从任务参数准确性和执行结果准确性两个方面来评估智能体的性能。消融实验-模型型号对比结果如表 7 所示，可以看出搭载了 Qwen2.5-7B-Chat 的智能体，在任务参数准确性和查询结果准确性上均取得了最高得分，其中任务参数准确性和查询结果准确性相对 GLM4-9B-Chat 高出 3.1%、11.1%。

表 7　消融实验-基础模型对比结果

Tab.7　Ablation study on base models

模型	任务参数准确性	查询结果准确性
Baichuan+Agent	1.64	1.54
Llama+Agent	4.30	3.50
GLM +Agent	4.54	3.68
Qwen+Agent	**4.68**	**4.09**

其次，选择同系列模型，对比使用不同参数规模的模型对智能体效果的影响，对比的模型包括 Qwen2.5-7B-Chat 、 Qwen2.5-14B-Chat 、 Qwen2.5-32B-Chat、Qwen2.5-72B-Chat，从任务参数准确性和执行结果准确性两个方面来评估智能体的性能。消融实验-大模型参数规模对比结果如表 8 所示，可以看出随着大模型参数规模的增加，任务参数准确性和查询结果准确性均逐渐提升，参数规模为 72B 的大模型智能体，在任务参数准确性和查询结果准确性上相对参数规模为 7B 的智能体分别提升 5.1%、10.3%。

表 8　消融实验-参数规模对比结果

Tab.8　Ablation study on parameter scales

模型(Qwen2.5)	任务参数准确性	查询结果准确性
7B+Agent	4.68	4.09
14B+Agent	4.82	4.29
32B+Agent	4.83	4.38
72B+Agent	**4.92**	**4.51**

此外，在针对任务分配方式的消融实验中，使用在之前实验中表现最好的 Qwen2.5-72B-Chat 大模型，依次减少智能体中任务检索树形结构的层数，对比 1 至 4 层树形结构对智能体效果的影响，实验结果如表 6 所示。可以看出使用 4 层树形结构进行任务分配时，任务参数准确性和查询结果准确性均取得最好效果，相对于使用 1 层树形结构时分别提升 4.2%、5.4%。

表 9　消融实验-任务分配方式对比结果

Tab.9　Ablation study on task assaignment

层数	任务参数准确性	查询结果准确性
1	4.72	4.28
2	4.77	4.31

3	4.90	4.37
4	**4.92**	**4.51**

4 结论

本文提出的基于大模型智能体的水利一张图任务执行问答技术框架，由任务构造、任务分配、任务执行、问答生成四个部分组成，并以此为基础构建了一张图任务执行问答应用。同时，为了客观评估本文构建的问答应用的效果，还创建了一套一张图任务的问答数据集。在该数据集上，本文提出的方法在任务分配准确性上达到 4.9/5 以上，最终的查询结果准确信上显著优于最新的中文开源大模型。本文提出的水利一张图任务执行问答技术框架具有较强的通用性，可进一步扩展至其它领域的任务系统。该框架能够减少用户对任务系统的学习成本和操作时间，便捷地获取信息和问题解答，提高工作效率，具有广泛的推广价值。

虽然本文提出的技术框架取得了较好的效果，但一方面地理空间信息领域的任务对结果的准确性要求较高，另一方面系统在实际应用时所面对的提问和查询需求也存在个性化和复杂任务的情况，因此在具体落地中还需要进一步提高该技术框架的准确性和对复杂任务的支持能力。下一步，将从复杂任务的构造、大模型进行任务分配的流程和思维链等方面开展进一步的研究工作。

5 参考文献

[1] 崔东林.新时代背景下 GIS 技术在水利工程信息化中的应用[J].工程技术研究,2023,8(06):202-204.DOI:10.19537/j.cnki.2096-2789.2023.06.065.

[2] Wu T, He S, Liu J, Sun S, Liu K, Han Q L, and Tang Y. A brief overview of ChatGPT: The history, status quo and potential future development[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(5): 1122-1136.

[3] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.

[4] Sun Y, Wang S, Li Y, Feng S, Chen X, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation[DB/OL]. (2021-07-05) [2024-02-17]. https://arxiv.org/abs/2107.02137.

[5] 钱峰,成建国,夏润亮,等.水利大模型的建设思路、构建框架与应用场景初探[J].中国水利,2024,(09):9-19.

[6] 杨柳,姚葳,马辉,等.LLM 在水利政府网站公共服务中的应用研究[J].水利信息化,2024,(02):58-62.DOI:10.19364/j.1674-9405.2024.02.010.

[7] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. arXiv preprint arXiv:2311.05232, 2023.

[8] Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey[J]. arXiv preprint arXiv:2312.10997, 2023.

[9] 张志鑫,明晨曦,刘颉,等.基于 JRAG 的涉水法律法规智能知识问答技术[J/OL].人民长江,1-11[2024-10-31].http://kns.cnki.net/kcms/detail/42.1202.TV.20240628.1750.006.html.

[10] Wang, L., Ma, C., Feng, X. et al. A survey on large language model based autonomous agents. Front. Comput. Sci. 18, 186345 (2024). https://doi.org/10.1007/s11704-024-40231-1

[11] Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: A survey[J]. arXiv preprint arXiv:2309.07864, 2023.

[12] Li B, Wu P, Abbeel P, et al. Interactive task planning with language models[J]. arXiv preprint arXiv:2310.10645, 2023.

[13] 蔡阳,谢文君,程益联,等.全国水利一张图关键技术研究综述[J].水利学报,2020,51(06):685-694.DOI:10.13243/j.cnki.slxb.20200081.

[14] 韦人玮,杨鹏,乔延军,等.水利空间信息资源整合与共享平台技术框架研究[J].长江技术经济,2020,4(04):104-108.DOI:10.19679/j.cnki.cjjsjj.2020.0424.

[15] Yang A, Xiao B, Wang B, et al. Baichuan 2: Open large-scale language models[J]. arXiv preprint arXiv:2309.10305, 2023.

[16] GLM T, Zeng A, Xu B, et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools[J]. arXiv preprint arXiv:2406.12793, 2024.

[17] Dubey A, Jauhri A, Pandey A, et al. The llama 3 herd of models[J]. arXiv preprint arXiv:2407.21783, 2024.

[18] Yang A, Yang B, Hui B, et al. Qwen2 technical report[J]. arXiv preprint arXiv:2407.10671, 2024.

[19] H. Xu, E. Hamzei, E. Nyamsuren, H. Kruiger, S. Winter, M. Tomko, and S. Scheider. Extracting interrogative intents and concepts from geo-analytic questions. AGILE: GIScience Series, 1:23, 2020.

[20] Pollali M A G. The Dataset GeoQuestions1089[J]. 2024.

[21] Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-45.

[22] 智慧司法技术总师系统,浙江大学,上海交通大学,阿里云计算有限公司,科大讯飞研究院.法律大模型评估指标和测评方法（征求意见稿）[EB/OL]. (2023-08-21) [2024-02-17]. http://sias.zju.edu.cn/2023/0823/c57510a2792895/page.htm.