

研究论文

基于大语言模型的河流地貌知识抽取

刘晓璐¹, 王嘉琳^{1,2}, 吴进¹, 马晓熠^{1,2}, 秦帅³, 王曙^{1,2,4*}

摘要 随着河流地貌研究日益依赖大规模文献信息的集成与分析,如何从非结构化文本中系统抽取关键术语及其语义关系,已成为推动地貌学数字化、结构化发展的关键技术问题。提出一种基于GPT(generative pre-trained transformer)大语言模型的河流地貌知识抽取方法,设计 Few-shot 提示结构,引导模型识别文本中的地貌实体与关系,并输出标准化三元组<实体₁, 关系, 实体₂>。研究方法包括语料准备、提示结构设计、模型调用与结果输出4个环节,实验基于30篇核心英文文献,累计抽取三元组7605条,涵盖22815个实体与关系元素。评估结果表明,该方法具备较高准确率(Precision为0.834)、良好召回率(Recall为0.789)、整体性能稳定($F_1=0.862$),适用于术语密集型地学语料的自动化处理。进一步分析表明,模型在处理复杂句式、隐含关系与多义词方面仍存在一定局限,未来可通过模型微调、知识约束与多模态融合进一步优化。本研究为地貌知识的智能建构与语义集成提供了可行路径,也为地理信息科学与人工智能的交叉融合拓展了应用空间。

关键词 河流地貌;大语言模型;知识抽取;三元组结构;GPT;自然地理智能化

河流地貌是地貌学与水文学交叉形成的重要分支,主要研究河流在内外营力共同作用下形成、演化及分布的地貌形态与动力学机制^[1-2]。河流作为地球表层系统中最具活力的地貌动力因子之一,通过侵蚀、搬运与沉积等过程持续塑造地貌格局,是连接地壳运动、大气循环与生态系统演变的关键纽带^[3-4]。从微观河床形态到宏观流域演化,河流地貌不仅揭示自然过程的物理机制,也是理解景观结构与水文过程

耦合关系的理论基础^[5]。然而,随着全球气候变化与人类干扰加剧,河流系统响应日趋复杂,传统依赖经验描述的研究范式逐渐难以满足现代地理科学在机制分析、预测能力与系统集成方面的需求^[6]。特别是,对河流地貌知识的智能化获取、整合、分析与应用方面^[7-9],如何实现从非结构化文本到结构化知识的高效转化,已成为亟待突破的技术瓶颈。因此,推动河流地貌研究在理论深度、数据整合与智能分析方面的发展,既是学科进步的必然趋势^[10],也是服务水土保持、流域管理、生态修复等国家生态安全和资源管理的现实需要。

尽管已有大量关于河流地貌的研究,但传统知识抽取方法在应对术语密集、句式复杂与语义隐含的文本时仍存在明显不足。基于规则的方法依赖固定模板,虽具有可控性,却难以覆盖复杂句式且维护成本

1. 中国科学院地理科学与资源研究所地理信息科学与技术全国重点实验室,北京 100101
2. 中国科学院大学,北京 100049
3. 河北省南运河河务中心,沧州 061001
4. 江苏省地理信息资源开发与利用协同创新中心,南京 210023

收稿日期: 2025-06-11; 修回日期: 2025-10-26

基金项目: 内蒙古自治区人工智能+专项(2025KJTW002303); 国家自然科学基金面上项目(42471503)

作者简介: 刘晓璐, 博士后, 研究方向为地学知识图谱, 电子邮箱: liuxl.19b@igsnr.ac.cn; 王曙(通信作者), 副研究员, 研究方向为地理知识图谱与地理人工智能, 电子邮箱: wangshu@igsnr.ac.cn

引用格式: 刘晓璐, 王嘉琳, 吴进, 等. 基于大语言模型的河流地貌知识抽取[J]. 科技导报, 2026, 44(17): 102-113; doi:10.3981/j.issn.1000-7857.2025.06.00046

高;基于词典的方法依赖术语完备性,对新词和多义表达敏感,适应性差;统计学习方法(如支持向量机、条件随机场)虽然具备一定泛化能力,但依赖大量标注语料和人工特征工程^[11]。这些方法在通用语境下运行良好,但面对专业领域的非结构化文本时,往往因表达模糊、语义割裂而导致性能下降^[12],难以满足自动化与高适应性要求。

随着深度学习与人工智能的快速发展,大语言模型(large language models, LLMs)在自然语言处理领域取得突破,展现出强大的上下文理解与语义建模能力^[13-14]。相比传统方法,LLMs 不仅能够处理复杂结构、多样句式与隐含逻辑,还支持少样本(few-shot)甚至零样本(zero-shot)任务配置,在缺乏大规模标注数据的条件下也能实现高质量知识抽取^[15-16]。这类模型已在医疗、法律、金融等术语密集型领域得到验证,表现出良好的通用性与迁移性^[17-19]。对于河流地貌研究而言,借助大语言模型从文本中自动识别地貌实体、过程机制及其关系,并以三元组形式组织表达,不仅能够提升知识体系构建的效率与深度,也为地貌学的数字化和智能化发展提供了全新技术路径。因此,探索大语言模型在河流地貌知识抽取中的应用,具有重要的理论价值与实践意义。

为应对河流地貌文本中术语密集、语义隐含与结构复杂等问题,本研究构建了一套基于大语言模型的知识抽取方法。以 GPT 模型为核心,通过语料库构建、Prompt 设计、模型调用与识别、结果输出,建立从原始文献到结构化三元组的自动化处理流程。该方法旨在验证大语言模型在地貌学语境下的可行性与优势,为未来知识图谱构建以及语义集成提供技术基础。

1 大语言模型知识抽取方法

1.1 知识抽取任务界定

地貌学文献中广泛蕴含关于地貌形态(图 1)、演化过程、控制因子等方面的重要知识,多以自然语言形式表达,其术语密集、语义隐含、上下文依赖性强,缺乏结构标记,难以实现高效整合与复用。河流地貌中,其描述常涉及“冲积扇”“侧蚀”等术语和演化逻辑,表达方式多样,为自动化解析带来挑战。

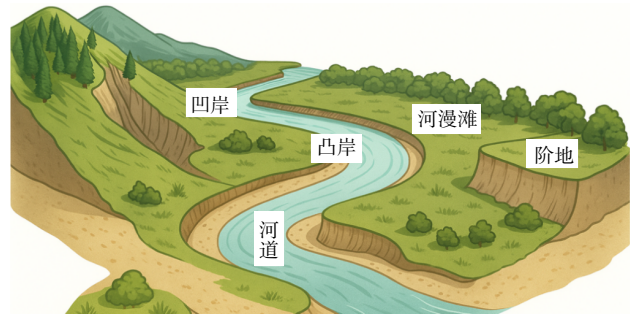


图 1 河流地貌典型单元

三元组<实体₁, 关系, 实体₂>作为知识抽取的基本单位,能够兼容各类河流地貌实体及其语义联系,包含时空关系、过程逻辑与因果结构等,且具备良好的通用性和可视化潜力,以形成结构化、可计算的知识体系。

因此,为实现自动且智能化的河流地貌知识抽取,本研究中的河流地貌知识以三元组结构为核心,开展基于大语言模型的自动化知识抽取任务,其抽取目标包括:(1) 识别地貌术语(实体),构建河流地貌知识图谱中的节点;(2) 识别河流地貌实体间的语义关系及其方向性,构建完整语义路径;(3) 输出标准化三元组,便于后续的评估分析与系统集成使用。

1.2 方法框架与关键特征

1) 研究方法。本研究采用基于大语言模型的提示驱动式知识抽取方法,以 ChatGPT 作为核心推理引擎。该方法依托预训练模型的深层语义建模能力,通过 Prompt 对任务目标与输出格式进行显式约束,使模型在少样本条件下完成实体与关系的自动识别与结构化表达。与 DeepSeek、文心一言等模型相比,ChatGPT 在英文专业文本上的句法解析与关系抽取一致性更佳;其 API 稳定、参数可控,有利于保证实验复现与结果对比。鉴于本研究语料以英文河流地貌文献为主,采用 ChatGPT 有助于提升抽取过程的稳定性与可靠性。

2) 研究框架。整体抽取流程包括 4 个核心环节:语料库构建、Prompt 设计、模型调用与识别、结果输出。首先选取具有代表性的河流地貌文献作为输入语料,并进行预处理;随后通过 few-shot 提示设计任务指令与示例,引导语言模型识别实体与关系;再利用 OpenAI 提供的 GPT API(application programming interface)进行分段调用,实现语义理解与结构生成;

最后将三元组结果规范化存储。

该方法具备 3 方面创新：一是低标注需求，仅依赖少量示例，无需大规模人工训练集；二是输出结构规范化，适配标准知识系统；三是具备良好迁移能力，可推广至其他地貌类型与自然地理文本环境。整体框架如图 2 所示，展示了从原始文本到结构化知识的完整处理路径，包括文本片段输入、指令设计、实体识别、关系抽取、三元组构建与结果保存等环节。

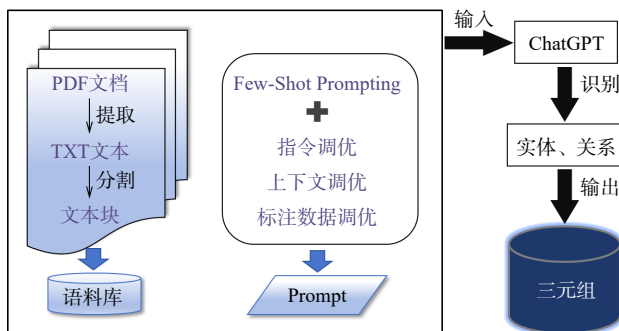


图 2 研究框架

2 实验过程

2.1 语料库构建

本研究从 Web of Science 数据库中筛选获得 30 篇英文河流地貌领域高引用文献，原始文本总量约为 18.7 万字，作为知识抽取的实验语料，语料类型多样、表述方式丰富，具有良好的知识密度与抽取潜力。为确保数据的代表性与可靠性，设置以下检索条件：(1) 数据库范围包括 Web of Science 核心合集、中国科学引文数据库 (Chinese Science Citation Database, CSCD)、Current Contents Connect 及 Derwent Innovations Index；(2) 关键词设置为“Fluvial OR River Landform”；(3) 根据相关度和引用频次排序选取文献。最终选定文献涵盖不同地貌类型与过程描述，语言表达规范，具备较高学术质量。

为便于处理，所有文献均转换为“.txt”格式文本文件，并进行预处理操作，包括去除格式空格、错误字符及非正文内容，以提升文本整洁性与模型识别准确率。所有抽取结果统一组织为标准三元组结构，格式为：<entity₁, relation, entity₂>。

在语料预处理阶段，首先对全量语料统一进行小写化，以增强实体识别的稳定性；对单位表达（如

“m³/s”“km²”等）进行标准化替换，统一为“m^{^3}/s”“km^{^2}”等格式；针对模糊或多义术语，结合《Glossary of Geology》与美国地质调查局 (United States Geological Survey, USGS) 河流分类标准，进行术语消歧与语义对齐处理；并从语料中选取少量典型术语（如“sandbar”“braided channel”“meander loop”等）进行示例性标注，作为 few-shot 提示中的参考样例，用以引导模型识别地貌实体。

2.2 Prompt 设计

为实现高质量的三元组抽取，本研究采用基于 Prompt Engineering 的任务引导机制，通过构造规范化提示词 (Prompt) 激发 GPT 的抽取能力。提示结构包括 System Message 与 User Message 2 个部分，分别承担任务设定与语料输入功能。

System Message 用于向模型明确任务目标、术语范围与输出格式要求。具体内容包括：指令性语言（如“你是一位地貌学专家，请从文本中抽取结构化三元组”）、地貌实体与语义关系的定义说明、三元组的标准表达形式（如“<实体₁, 关系, 实体₂>”），以及不允许的输出样式提示（如避免生成自然语言段落或模糊关系词）。通过设置专业化角色与结构化输出约束，引导模型在生成内容时遵循特定语义逻辑与结构规范。

User Message 用于输入待处理的文本片段，考虑到模型输入限制及语义完整性，每次提交一个段落或一定词长的文本。为提高模型理解能力，User Message 中嵌入了若干 few-shot 示例，即在正文前引入已标注的三元组示例（图 3），作为上下文引导模型学习抽取规律。示例的选取遵循句式多样、术语明确、语义典型的原则，以涵盖常见地貌术语及其多样关系类型。

Input: The ravine is a V-shaped valley.
Output: <ravine,is,V-shaped valley>

图 3 河流地貌实体、关系的文本标注示例

在 Prompt 调优过程中，本研究围绕指令、上下文与示例 3 方面优化提示结构，以提高大语言模型在河流地貌知识抽取任务中的输出质量和结构一致性。调优过程遵循控制变量原则，在调整某一提示要素时保持其余部分不变，以确保对比结果的有效性。

1) 指令调优：最初的任务提示为“从 User Message

中提取 2 个河流地貌实体及其关系”，但实验发现，模型输出中出现了大量泛泛而谈的术语组合，语义关联性弱。借鉴提示工程相关研究，本研究在 System

Message 中增加“你是一位河流地貌领域专家”的角色设定，以强化模型的语境适应性(图 4)。通过调整模型角色，使语言模型在地貌语境下的角色理解更精准。

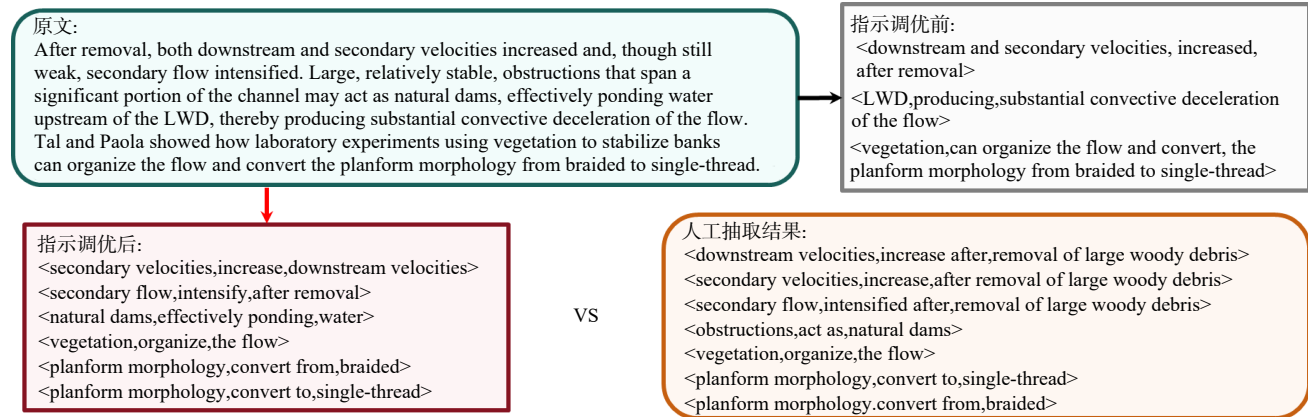


图 4 指示调优输出结果对比

2) 上下文调优: 初始提示仅包含抽取格式规范与任务说明, 模型虽能输出结构化三元组, 但部分实体不具备河流地貌语义。例如, 模型曾将“the maintenance of river anabranching”作为完整实体抽取。为此, 在上下文中补充河流地貌实体的定义与限制条

件, 引导模型判断实体是否属于预设语义范围。图 5 调优后结果显示, 提高了抽取稳定性, 实体边界更精确, 语义歧义减少, 如“anabranching”替代了含义模糊的整体短语, 而“bifurcation stability”等模糊实体仍存在, 提示模型尚存在对术语语义边界判断的局限性。

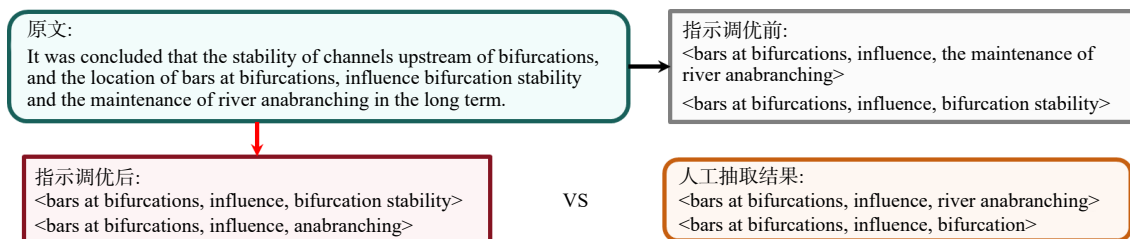


图 5 上下文调优输出结果对比

3) 标注数据调优: 实验初期仅提供正例三元组, 导致模型倾向于为所有句子输出结果, 即使句中并无有效地貌实体对。为引导模型判断“无效输入”情形, 添加图 6 所示空三元组示例(如“不包含有效实体关系时返回空结果”), 使模型学会拒绝输出不满足条件

的结构化结果, 从而有效提升了整体准确率。此外, 为提升模型对复杂句式 and 并列结构的识别能力, 进一步在示例集中补充此类句式的正例样本, 如图 7 所示增强了模型对多实体关系与句法嵌套结构的适应性, 输出效果亦随之改善。

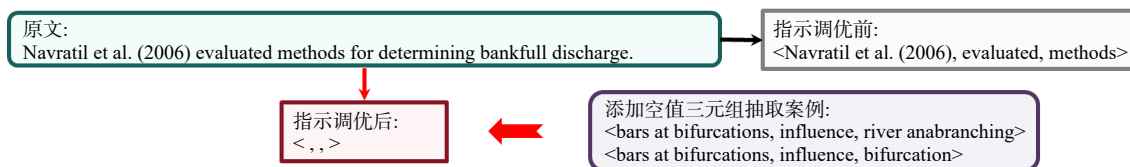


图 6 添加空值的输出结果对比

综上, Prompt 通过 3 方面调优显著提升了模型在河流地貌语境下的知识抽取性能(图 8)。通过系统性

的任务指令优化、语境限定与示例强化, 模型在实体边界识别、关系词选取及句法适应性方面均表现出

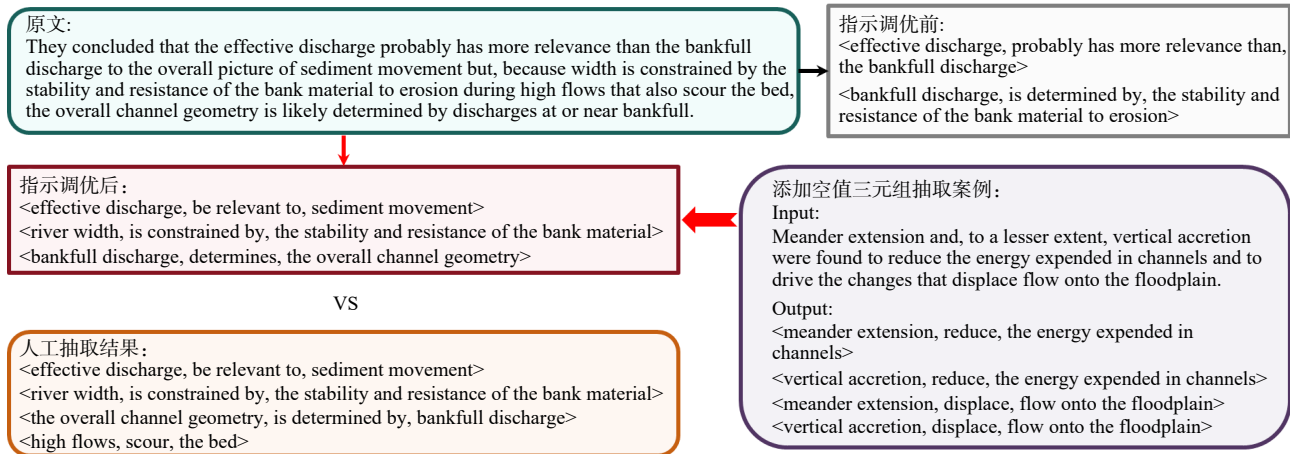


图7 标签数据补充的输出结果对比

You are a geomorphology expert. From the given text segment, identify two entities related to fluvial geomorphology and determine their semantic relationship. Geomorphic entities include, but are not limited to: river channel, terrace, alluvial fan, sandbar, floodplain, sediment, catchment, and fault. Typical relations describe genetic, process-based, or structural links, such as: controls, influences, causes, forms, results in, and is part of. Refer to the following examples to learn the extraction pattern:

Example 1:
Text: River terraces are formed by repeated incision of the river.
Output: <river terraces, formed by, river incision>

Example 2:
Text: The paper reviews previous studies on river morphology.
Output: []

Example 3:
Text: The sediment transported by floods affects channel stability and bank erosion.
Output: <sediment transport, affects, channel stability>, <sediment transport, affects, bank erosion>

Use the standardized triplet format: <entity₁, relation, entity₂>.
If no valid geomorphic entities or relationships are found, return an empty result "[]".
Do not include any natural language explanations or comments — only output the triplet structure.

图8 调整后的 Prompt

更高一致性与稳定性, 为未来大规模抽取奠定了基础。

2.3 模型调用与识别

本研究采用 OpenAI 提供的 GPT 语言模型作为抽取工具, 完成河流地貌文献中的语义三元组自动抽取任务。GPT 具备优越的上下文理解与语言生成能力, 特别适用于术语密集、表达复杂的专业文本环境^[20]。其 few-shot 能力可在无需大规模人工标注的条件下, 通过示例引导完成知识识别与结构化表达, 契合本研究对抽取效率与精度的双重需求。

实验通过调用 OpenAI 官方应用程序编程接口 (application programming interface, API) 接口进行模型接入(图9), 基于 Python 语言开发抽取流程脚本, 主要依赖 OpenAI 库实现与模型的交互调用。运行环

Function AskGPTToAnswer(base_url, api_key, model, system_prompt, user_input):

```
# Step 1: 创建 OpenAI GPT 客户端
Create OpenAI client with:
- base_url = base_url
- api_key = api_key
- http_options: allow_redirects = true

# Step 2: 构造 Chat 请求
response = client.chat.completions.create with:
- model = model
- messages = [
    { role: "system", content: system_prompt },
    { role: "user", content: user_input }
]

# Step 3: 返回 GPT 回答内容
Return response.choices[0].message.content
```

图9 调用 GPT 的 API 接口执行知识抽取任务

境基于本地工作站, 操作系统为 Windows 10, Python 版本为 3.9, 模型调用基于稳定网络连接, 采用分段处理策略按批次提交请求。

为排除文本长度对结果稳定性的影响, 同时需要

保证语言模型在理解地貌术语与关系结构时具备完整的语义上下文,本研究在实验中采用分段输入策略,将长篇文献内容拆分为若干较短的语料块进行逐段处理。同时对单次处理文本长度限制进行针对性测试,具体以河流地貌经典文献^[21]第2章部分段落为例,选取其中若干自然段作为测试数据(参数均保持默认设置,以排除参数干扰),人工标注其所包含的标准三元组,构建真值集。在保持段落语义不变的前提下,分别设置文本长度为70词、80词、90词与100词,逐项调用GPT模型进行抽取,并对其结果与人工标注的56条(真值)三元组进行比对分析。

如表1所示,从“完全正确”来看,当词长设为80时,模型提取的三元组中共有31条与真值完全一致,是所有参数中准确率最高的一组,尽管其抽取总

数(49条)略少于70(63条)与100(53条),但其在误抽率和误配率上的控制表现更优,说明其结构判别更稳定;在“错误三元组”(即三项均错误)中,80词参数的误抽数量为9条,明显低于其他设置,尤其优于90(18条)和70(19条),表明在该参数下模型更少出现严重误判;值得注意的是,在“漏抽三元组”(即未能识别出应抽内容)中,80词设定下的漏抽数量为16条,为4组中最多,这说明在控制文本长度以确保语义完整性的同时,可能牺牲了一定的召回率。但由于其在“完全正确”与“错误提取”2项关键指标上表现最佳,整体权衡其精度与稳定性,仍可视为本任务中最优输入长度设置。因此,80词长度在本实验框架下兼具较高的准确率、较低的误抽率和可接受的漏抽水平,可作为河流地貌类文本三元组抽取任务的推荐参数设置。

表1 不同文本长度设置下三元组抽取准确性对比(单位:条)

词长	抽取总数	完全正确	实体×2正确	实体+关系正确	实体×1正确	关系正确	错误三元组	漏抽三元组
70	63	30	1	8	4	1	19	12
80	49	31	0	8	0	1	9	16
90	53	24	1	8	2	0	18	14
100	53	24	1	9	6	1	12	11

此外,文本分段标准以句子边界为主,兼顾语义完整性与段落逻辑结构。对于嵌套复杂的长句或多重修饰结构,在不影响语义理解的前提下适当拆分;而对存在跨句因果关系或句间术语延续的段落,则整体输入模型处理,确保信息不被割裂。在输入策略上,还采取了上下文完整性保护机制,即避免在术语结构、句法单位中间断句。例如,对于包含“the development of river terraces driven by sediment dynamics”此类表述的句子,完整保留该单元,避免语义断裂造成模型理解偏差。

为进一步降低语义截断造成的实体或关系割裂,本研究在分段过程中引入基于语义边界的分段候选器。该候选器以英文句法停顿和语义单元完整性为主要约束,具体包括:(1)优先在句末标点(。)或强停顿符号(; : —)处切分;(2)避免在并列或从属连词(and, or, because, which, that)内部切分;(3)对括号、引号等成对符号进行配对检测,不在对内断句;(4)对文本中的专业术语短语进行整体保护,避免在多词术语内部切分;(5)当分段长度超出模型输入限制时,在

最近的“低风险语义边界”处回退切分。该策略能够保持句法和语义的自然边界,最大限度地避免因文本截断导致的实体残缺或关系方向反转问题,从而保证输入语料的语义完整性。

在此基础上,为进一步提升领域实体识别的准确性与一致性,本研究在术语消歧与标准化之后,引入术语定位机制。该机制在分段完成后对语料进行术语识别与标签化处理,结合《地貌学辞典》《Glossary of Geology》和USGS河流分类标准,通过词典匹配与轻量级命名实体识别,将典型术语(如braided channel、terrace、riverbed等)标记为“显式已知实体”。因此,模型在Prompt语境中可优先识别并聚焦于这些术语,实现语义过滤与领域聚焦。该机制与语义边界候选器相互配合:前者保障文本切分的语义完整,后者强化术语识别的领域一致性,共同提升知识抽取阶段的精度与稳健性。

2.4 实验结果输出

模型在完成语义识别后,自动将抽取结果以结构化三元组形式输出,并保存为标准化数据文件。所有

结果均采用统一格式<entity₁, relation, entity₂>, 确保可直接导入后续分析模块。

最终共获得 7605 条三元组, 涉及 22815 个实体与关系元素。抽取出的三元组在不同文献中的数量呈现明显的差异性, 平均每篇文献抽取 253 条, 最高可达 398 条, 最低为 121 条, 主要取决于文本中与河流地貌相关的术语密度与表述风格。

3 结果与分析

3.1 评估方法与指标体系

为全面评估大语言模型在河流地貌三元组抽取任务中的性能表现, 本研究从准确性与完整性 2 个维度构建评估体系, 采用常用的 3 项评价指标: 准确率 (Precision)、召回率 (Recall) 与 F_1 分数 (F_1 -Score)。其中, 准确率用于衡量模型输出中正确三元组所占比例, 召回率反映模型从应识别出的三元组中成功提取的比例, F_1 分数作为二者的调和平均, 体现整体性能。其计算公式如下:

$$P = \frac{T_p}{T_p + F_p} \quad (1)$$

$$R = \frac{T_p}{T_p + F_N} \quad (2)$$

$$F_1 = \frac{2PR}{P + R} \quad (3)$$

式中, P 为准确率 Precision, R 为召回率 Recall, T_p 为正确抽取的三元组数, F_p 为错误抽取的三元组数, F_N 为漏掉的三元组数, F_1 为 F_1 分数。

在匹配策略上, 本研究分别设定 3 种评估层级: 严格匹配、容错匹配与语义相似度匹配。严格匹配要求三元组中的实体₁、关系与实体₂ 3 项与人工标注结果完全一致; 容错匹配允许词形变化或同义词替换 (如“sediment transport”与“sediment flow”视为等价), 适用于处理语言灵活表达下的误差评估; 语义相似度匹配基于 BERT 等嵌入模型计算三元组整体的语义距离, 判断其是否在一定阈值内可视为“语义等价”, 以捕捉隐性正确项。

此外, 为提升评估的解释力与代表性, 本研究还引入人工抽查机制, 对自动评估结果中分类为 F_p 与 F_N 的样本进行复核与注释, 分析模型抽取失误的具体原因。该过程由具备地貌学背景的研究人员完成, 主

要识别实体边界错误、关系方向混淆、句法误解等常见错误类型, 为后续方法优化与样例库建设提供反馈依据。

3.2 抽取性能评估

在完成结构化三元组的抽取后, 本研究首先对命名实体识别与关系抽取 2 个子任务进行了独立评估, 随后基于“严格匹配”标准对完整三元组结构的匹配情况进行了综合检验, 以全面衡量大语言模型在河流地貌知识抽取任务中的表现。

1) 独立子任务评估。

命名实体识别: 主要考察模型对地貌实体边界及其类别的识别能力。实验结果显示, 模型在实体识别任务上取得了 Precision 为 0.912、Recall 为 0.886、 F_1 为 0.899。该结果表明, 大语言模型能够较为准确地定位并分类河流地貌相关实体, 其语义辨识能力较强。

关系抽取: 聚焦模型对实体间语义关系的识别。关系抽取子任务中, 模型获得 Precision 为 0.873、Recall 为 0.841、 F_1 为 0.857。可见, 模型在理解上下文语义、把握实体间连贯性关系方面也具备较高水平。

2) 完整三元组结构匹配。

为进一步检验模型在端到端抽取流程中的表现, 采用“严格匹配”标准统计实体₁、关系、实体₂ 3 项元素同时正确的比例。结果为 Precision 为 0.834、Recall 为 0.789、 F_1 为 0.862。相比于独立子任务评估, 整体三元组指标有所下降, 主要归因于以下 2 方面因素。(1) 误匹配误差传递: 个别情况下, 虽然实体及关系本身识别正确, 但三者配对存在偏差; (2) 结构化输出一致性: 输出格式或顺序与标注略有不符, 也会导致严格匹配失败。

表 2 汇总了主要评估指标结果。

表 2 三元组抽取结果评估

任务类型	Precision	Recall	F_1 -Score
实体识别	0.912	0.886	0.899
关系抽取	0.873	0.841	0.857
三元组整体	0.834	0.789	0.862

综上, GPT 在河流地貌语料上的抽取任务中表现出较强的实体识别与语义关系理解能力, 整体抽取性能稳定, 为后续地貌知识系统的构建提供了可靠支撑。

3.3 抽取结果分析

1) 抽取结果分布特征。

为进一步验证模型在多类型语义关系下的识别效果,本研究依据关系词的语义功能,将抽取结果划分为4类:因果关系、空间关系、演化关系和语义关系。统计各类高频关系词的Top3见表3。

表 3 各类高频关系词统计

关系类型	示例关系词(Top 3)	说明
因果关系	controls, causes, results in	表示一个事件或过程是另一个事件或过程的原因,是驱动地理系统演化的核心动力关系
空间关系	occurs in, distributed along, located at	描述地理实体或现象在空间上的位置、包含或组成关系
演化关系	initiates, develops into, formed by	描述一个地理实体或状态随时间转变为另一种实体或状态的序列过程,强调阶段性与结果
语义关系	is/are, has/have, includes	定义概念之间的逻辑或分类学联系,用于构建领域本体的层次结构和属性描述

结果表明,在关系类型方面,出现频次最高的谓词包括“is/are”“has/have”“includes”“related to”“cause”等,除常见语义关联类关系外,模型亦能识别出具有明确语义方向性的因果与演化关系,以及具备空间约束的结构关系。尽管这些复杂语义关系的频次低于一般类,但在不同语境中均得到准确提取,说明模型具备对多层级语义逻辑的理解与抽取能力。

如表4所示,实体项中频次较高者主要包括“river”“river channel”“riverbed”“floodplain”“sediment”等,说明抽取结果聚焦于河型与过程单元的主干结构,符合地貌研究的知识重心。

表 4 Top10 实体词频统计

实体	抽取次数/次
river	208
river channel	145
riverbed	132
floodplain	128
sediment	114
bed rock	98
catchment	91
tributaries	84
terrace	79
alluvial fan	71

对于每一个大类中的词,合并相同词,并统计相应的频数。对于相似的词,如“is”和“are”,或同一词根的词,如“influence”和“influences”,则将它们频数合并统计。进一步对三元组内容进行分词与词频统计分析后发现,抽取结果呈现出明显的长尾分布特征:即3大类中每类词频数在10以上的词数占总词

数的1.68%~1.78%,Top100术语覆盖全部实体项的62%,而后500个术语仅占据不到18%的频次。这一现象表明,大语言模型能够准确识别高频核心概念,同时也具备一定的“探索能力”,能发掘部分低频、边缘化的术语,有助于扩展知识体系边界。

2) 抽取案例。如图10所示,基于GPT的河流地貌知识抽取方法在多个维度均展现了显著优势:(1)将完成时或被动语态的动词自动规范化为一般时态,并完整识别包含多重修饰的空间定位短语,保证实体边界的精确;(2)在复杂句式中能剥离冗余结构,聚焦核心因果关系,准确提取主控动词与对应实体;(3)对于地貌过程与结果映射,如“沉积过程对阶地形成的控制”,模型能恰当选用专业动词并完整保留复合术语;(4)它对含有现在分词等高级语法结构的专业术语表现出良好适应性,能准确识别“tectonic motion accelerating the uplift of the plateau”这类长短语;(5)不仅适用于传统段落文本,还能跨越格式差异,从表格标题或注释中提炼出“classified according to”类型的映射关系,体现了对多源文本格式的兼容能力。这些特性使GPT成为构建高质量河流地貌知识图谱的有力工具。

此外,为了进一步提升结果表达的直观性和可读性,本研究对典型抽取结果进行了语义网络化可视化展示。图11中节点分别表示地貌实体和过程,箭线表示模型识别的语义关系,灰色背景文字为关系标签。网络展示了由气候变化、构造隆升等控制因子,经侵蚀、搬运与沉积等地貌过程,最终形成河流、洪泛平原和阶地等地貌形态的逻辑链条,体现了模型在

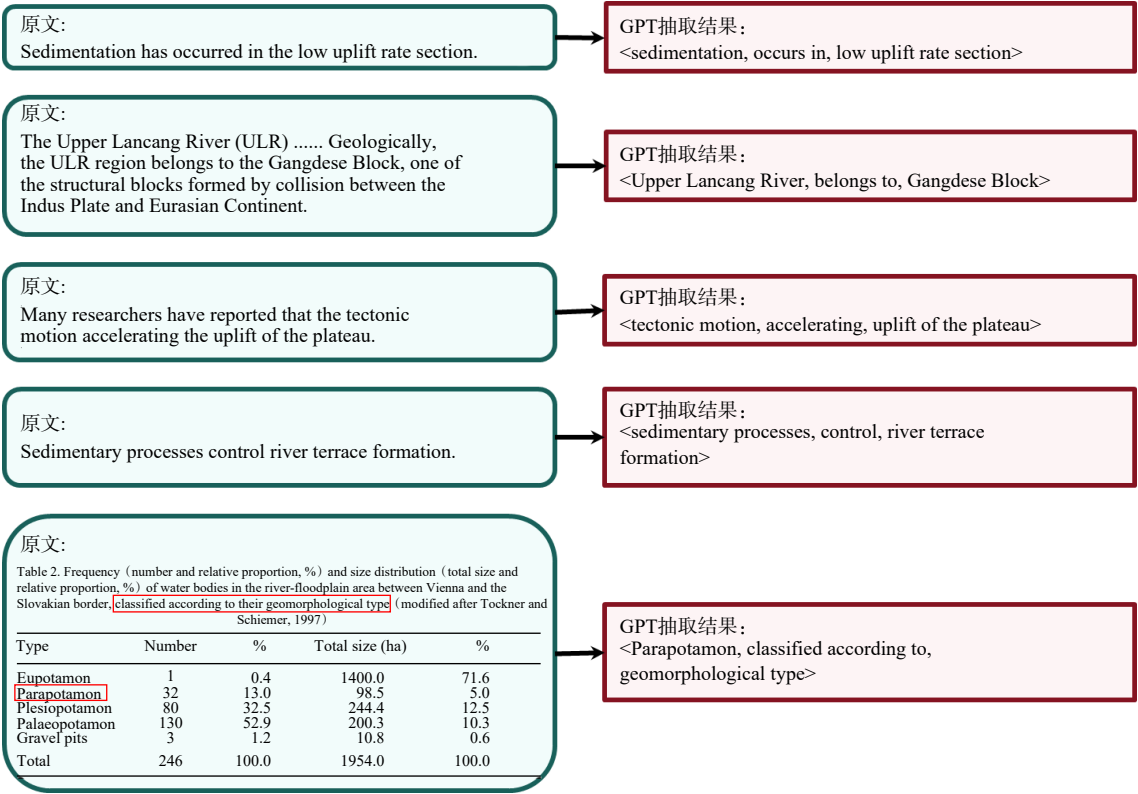


图 10 抽取结果示例

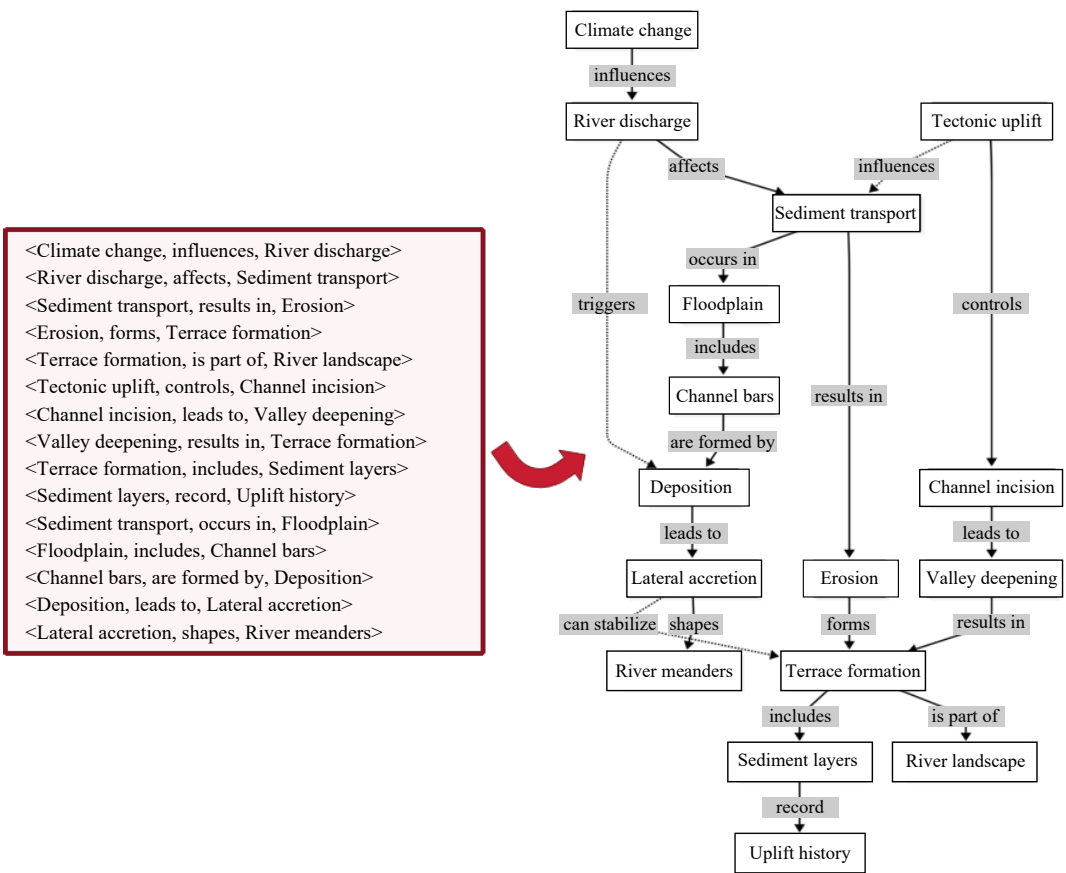


图 11 河流地貌知识抽取结果的语义网络示意

多类语义关系下的结构化理解能力,使抽取结果更加直观可读。

4 讨论

4.1 方法适用性

本研究提出的基于大语言模型的河流地貌知识抽取方法,在抽取准确率、系统适应性与运行效率方面表现出明显优势。首先,该方法具有高度自动化特性,无需依赖预设规则或复杂词表,可直接处理来自不同文献、风格各异的自然语言文本,具备良好的文本类型迁移能力。其次,得益于 GPT 在 few-shot 语境下的任务理解能力,本方法不完全依赖系统人工标注语料,仅依靠少量示例即可完成抽取任务,显著降低了时间成本。与传统基于规则或浅层学习的抽取方法相比,本方法在精度与召回率上均表现出更优性能,且具有更强的句式适应能力与关系理解能力,适合在专业领域开展大规模知识结构化工程。

在第 3.3 节的结果分析中可以看出,模型输出的关系中仍存在较多功能性谓词(如 is/are、has/have、includes、related to 等),导致知识图谱在语义解释层面信息密度不足。为提升关系抽取结果的语义表达力和结构一致性,本研究在抽取后进一步引入了关系谓词标准化与领域归并机制。该机制以河流地貌语义场为核心,构建小规模领域谓词集合,涵盖常见的高信息量关系词,如 causes、controls、influences、results in、occurs in、located at、erodes、deposits、forms、composed of、records、is type of 等。随后对功能性关系进行语义归并与规则映射:当 is/are 用于类别定义时映射为 is type of, has/have 和 includes 在组成语义下统一为 composed of 或 includes;若上下文中出现因果触发词(如 because of、due to、driven by 等),则 related to 提升为 causes 或 influences;无法确定语义方向的保持为 related_to 并标注为低信息项。映射过程中同时进行了词形归一化与方向校正,保证谓词层的语义一致性。不影响模型原始评测指标,但显著提升了关系层的可解释性与图谱结构的规范性。通过领域谓词集合的引入,知识图谱在表达因果、空间和演化等核心语义时更加清晰,也为后续基于地貌本体的关系扩展与语义推理提供了接口。

4.2 局限性分析

大语言模型在本研究的知识抽取任务中表现出较高的整体准确率,但在具体执行过程中仍存在一定局限。首先,模型对抽象关系或隐含逻辑表达(如“演化控制”“过程趋势”)的识别稳定性不足,尤其在缺乏显性动词或关系词时,容易出现漏判或误判。其次,对于长句或跨句依赖结构,模型在输入被切分时会产生语义割裂,影响关系抽取的完整性。为缓解这一问题,本研究提出了基于语义边界的分段候选器以保持语义连续,但该策略仍以规则约束为主,尚未系统比较“语义边界优先”与“固定词长优先”2 种方案的性能差异,后续可在统一语料与 Prompt 条件下开展 A/B 对照实验,评估不同分段策略对抽取准确率和语义一致性的影响。

此外,本研究未对模型生成参数(如 temperature、top_p)进行敏感性实验。由于研究重点在于非结构化抽取而非生成多样性,参数变化对主要指标影响有限,故采用固定参数以保证结果的稳定性与可复现性。未来可在模型微调或自适应生成任务中进一步探讨参数对语义一致性的潜在作用。

为进一步理解上述问题的具体表现,本研究对错误三元组进行了归类与分析,主要包括 4 类(表 5)。上述错误类型多与术语边界模糊、句法复杂性高及模型缺乏领域先验知识有关。术语表匹配虽然能显著降低噪声并提升抽取精度,但若严格限定匹配范围,可能造成新术语或复合表达的漏识别,从而降低召回率。未来可引入语义相似度匹配或动态扩展词典机制,在保持高 Precision 的同时兼顾 Recall,实现术语识别与语义泛化的平衡。

5 结论

本研究围绕河流地貌研究中的知识表达结构化问题,提出并实现了一种基于 GPT 的大语言模型知识抽取方法。通过设计专业语境下的 few-shot 提示结构,结合文本预处理与段落控制机制,实现了从地貌文献中自动抽取“实体-关系-实体”三元组的全过程。实验基于 30 篇英文核心文献,成功抽取出 7605 条结构化知识单元,形成可用于后续知识图谱构建的基础语义资源。评估结果表明,该方法在实体识

表 5 错误三元组类型

错误类型	描述	示例	说明
实体边界错误	模型截断或扩展实体范围, 丢失或包含冗余成分	原句: “Sedimentation has occurred in the low uplift rate section.” GPT抽取: <sedimentation, occurs in, low uplift> 真值: <sedimentation, occurs in, low uplift rate section>	未能完整识别“low uplift rate section”复合名词, 遗漏“rate section”
关系方向反转	主、宾位置互换或关系词方向误判	原句: “The river originates from the Jimayangzong Glacier.” GPT抽取: <Jimayangzong Glacier, originates, river> 真值: <river, originates from, Jimayangzong Glacier>	对“originates from”句法结构理解不足, 将介词“from”后对象误作主语
关系匹配错误	实体正确但关系词不准确或缺失重要介词	原句: “Because of the influence of the mountain location, the river changed course.” GPT抽取: <river, is at, mountain location> 真值: <river course, changed, due to mountain location influence>	对“changed due to”因果表达解析不全, 选用了不相关的关系词“is at”
其他	偶发格式或解析异常, 三元组缺失或无法解析	示例: 因输入文本格式或噪声, GPT返回空或非三元组结构	文本预处理或Prompt设计未能覆盖所有异常输入, 需增强格式校验机制

别、关系理解及三元组构造方面均表现出较高的准确率与稳定性, 验证了大语言模型在该任务中的可操作性与适用性。

本研究在技术路径与理论建构 2 个层面对地理知识抽取进行了拓展。一方面, 方法突破了传统基于规则或词典的抽取模式, 构建了面向地学文本的 Prompt 驱动式抽取框架, 为低标注环境下的信息结构化提供了新范式; 另一方面, 通过在河流地貌语境下的实践验证, 拓展了大语言模型在自然地理领域的实际应用场景, 丰富了人工智能与地理学科交叉融合的研究基础。

未来, 该方法可广泛应用于河流演化模拟、地貌因果链分析、语义地图构建等智能地理建模场景。同时, 通过引入领域本体、多模态数据融合与微调机制, 可进一步增强模型的专业适应能力与抽取稳定性。本研究为构建自然地理领域的语义表达体系和地学知识智能化奠定了基础, 也为相关领域的跨学科融合提供了可复制的技术路径。

参考文献 (References)

[1] Wheaton J M, Fryirs K A, Brierley G, et al. Geomorphic mapping and taxonomy of fluvial landforms[J]. *Geomorphology*, 2015, 248: 273–295.

[2] Schumm S A. The fluvial system[M]. New York: Wiley, 1977.

[3] Rosgen D L. A classification of natural rivers[J]. *Catena*, 1994, 22(3): 169–199.

[4] 王先彦, 于洋. 试论河流地貌学的新进展和趋势[J]. *地质科技通报*, 2024, 43(1): 150–159.

[5] Gilvear D J. Fluvial geomorphology and river engineering: Future roles utilizing a fluvial hydrosystems framework[J]. *Geomorphology*, 1999, 31(1/2/3/4): 229–245.

[6] 徐胜, 杨业, 张茂亮, 等. 构造-地貌-气候-生态系统动力学研究进展[J]. *地学前缘*, 2025, 32(3): 23–34.

[7] 李国杰. 智能时代呼唤新的科研方法[J]. *科技导报*, 2024, 42(10): 40–45.

[8] 程维明, 宋珂钰, 周成虎, 等. 地貌信息图谱研究述评与展望[J]. *地球科学进展*, 2022, 37(7): 661–679.

[9] Zhou C H, Wang H, Wang C S, et al. Geoscience knowledge graph in the big data era[J]. *Science China Earth Sciences*, 2021, 64(7): 1105–1114.

[10] 陈发虎, 张国友. 中国地理学的发展[J]. *科技导报*, 2020, 38(13): 12–18.

[11] 诸云强, 孙凯, 胡修棉, 等. 大规模地球科学知识图谱构建与共享应用框架研究与实践[J]. *地球信息科学学报*, 2023, 25(6): 1215–1227.

[12] 周成虎. 地理学赋能网络空间认知[J]. *科技导报*, 2023, 41(13): 1.

[13] Liang J Y, Hou S Y, Zhao A Q, et al. Design and application of a semantic-driven geospatial modeling knowledge

- graph based on large language models[J]. *Geo-spatial Information Science*, 2025, 28(6): 2927–2946.
- [14] Raiaan M A K, Mukta M S H, Fatema K, et al. A review on large language models: Architectures, applications, taxonomies, open issues and challenges[J]. *IEEE Access*, 2024, 12: 26839–26874.
- [15] 裴炳森, 李欣, 蒋章涛, 等. 基于大语言模型的公安专业小样本知识抽取方法研究[J]. *计算机科学与探索*, 2024, 18(10): 2630–2642.
- [16] 陈帅, 郭启全, 高春东, 等. 网络空间地理图谱的概念与方法[J]. *科技导报*, 2023, 41(13): 14–22.
- [17] 杨燕, 叶枫, 许栋, 等. 融合大语言模型和提示学习的数字孪生水利知识图谱构建[J]. *计算机应用*, 2025, 45(3): 785–793.
- [18] 皮乾坤, 卢记仓, 祝涛杰, 等. 基于大语言模型增强的零样本知识抽取方法[J]. *计算机科学*, 2025, 52(11): 22–29.
- [19] 陈文杰, 胡正银, 石栖, 等. 融合知识图谱与大语言模型的科技文献复杂知识对象抽取研究[J]. *现代情报*, 2025, 45(7): 14–25, 63.
- [20] 曹茹烨, 曹树金. ChatGPT完成知识组织任务的效果及启示[J]. *情报资料工作*, 2023, 44(5): 18–27.
- [21] Stott T. Fluvial geomorphology[J]. *Progress in Physical Geography: Earth and Environment*, 2010, 34(2): 221–245.

Research on river geomorphology knowledge extraction based on large language models

LIU Xiaolu¹, WANG Jiahe^{1,2}, WU Jin¹, MA Xiaoyi^{1,2}, QIN Shuai³, WANG Shu^{1,2,4*}

1. State Key Laboratory of Geographic Information Science and Technology, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. Hebei South Canal River Affairs Center, Cangzhou 061001, China
4. Jiangsu Center for Collaborative innovation in Geographical information Resource Development and Application, Nanjing 210023, China

Abstract As river geomorphology research increasingly relies on large-scale literature integration and analysis, the challenge of extracting structured knowledge from unstructured texts has become central to advancing digital and semantic approaches in geomorphological science. This study proposes a GPT-based method for automatic knowledge extraction in the field of fluvial geomorphology. A structured Few-shot prompt framework was designed to guide the model in identifying geomorphic entities and their relations, enabling the generation of standardized knowledge triples in the form of <entity₁, relation, entity₂>. The workflow consists of corpus preparation, prompt engineering, model invocation, and structured output. Applied to 30 high-quality English-language journal articles, the approach extracted a total of 7605 knowledge triples, encompassing 22815 entities and relations. Evaluation results demonstrate that the method achieves high accuracy (Precision=0.834), good recall (Recall=0.789), and stable overall performance (F1=0.862). Nonetheless, limitations remain in the handling of abstract logic, long-range dependencies, and relation directionality. Future improvements may involve domain-specific model fine-tuning, integration of geographic ontologies, and multimodal data fusion. This research provides a practical solution for intelligent knowledge construction in fluvial geomorphology and expands the methodological toolkit at the intersection of geographic information science and artificial intelligence.

Keywords fluvial geomorphology; large language models; knowledge extraction; triple structure; GPT; intelligent physical geography ●



(责任编辑 赵庆圆)