



# KNOWNET: 通过知识图谱集成 实现从大语言模型中引导式健 康信息检索

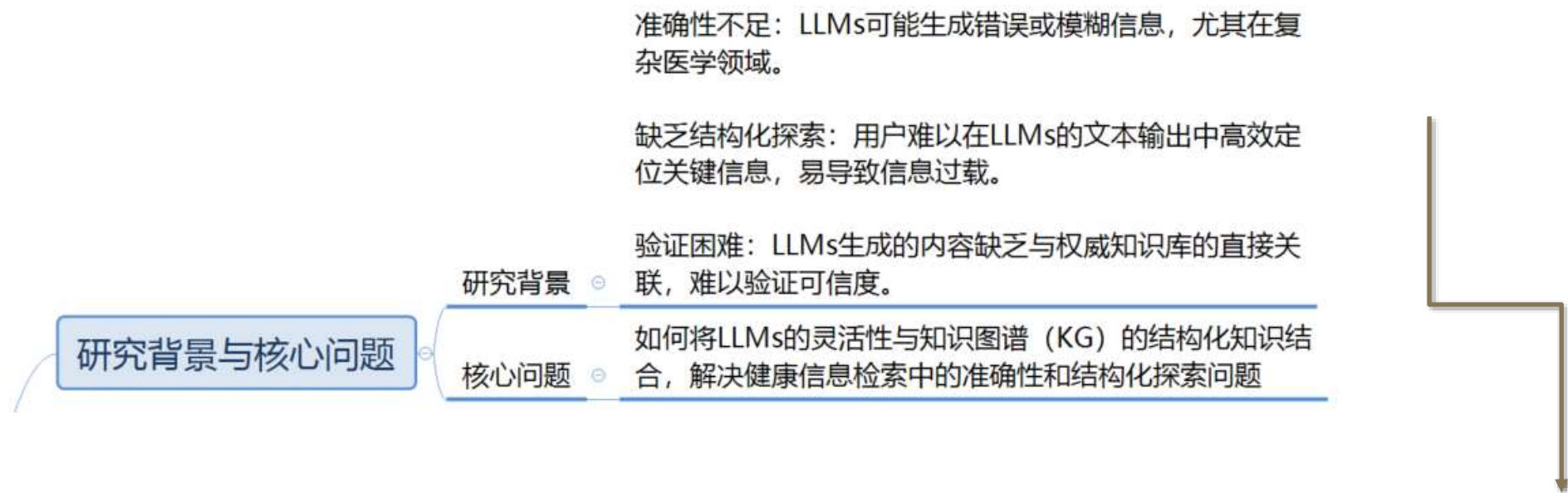
---

穆莹

# 背景

随着大型语言模型（LLMs）在信息检索中的广泛应用，尤其是在健康信息领域，**LLMs 的准确性和信息结构化的不足成为了主要问题**。LLMs 虽然能够生成丰富的文本内容，但其输出的信息往往缺乏验证，且难以帮助用户进行系统化的探索。特别是在健康领域，信息的准确性和结构化至关重要。

本文提出了 **KnowNET系统**，一个通过将 LLMs 与知识图谱（KG）结合的可视化系统，旨在提高健康信息检索的准确性和结构化探索能力。



具体目标包括：

**提高准确性：** 通过从 LLM 输出中提取三元组（实体及其关系），并将其映射到外部知识图谱中的已验证信息。

**结构化探索：** 基于知识图谱中当前探索实体的邻域，提供下一步探索建议，帮助用户全面理解主题。

# 方法

## KnowNET系统设计

### 核心方法

#### 结构化探索推荐

三元组提取：从LLM输出中提取实体及其关系（如“Omega-3脂肪酸—缓解—关节炎”）。

知识图谱映射：使用OpenAI的text-embedding-ada-002模型生成嵌入向量，通过余弦相似度匹配KG中的标准化实体和关系（阈值设为0.94）。

#### 信息提取与验证

支持证据关联：将匹配成功的实体与KG中的科学文献、临床试验数据等证据关联，增强可信度。

用户状态建模：基于n-context分析，将用户的查询历史映射为KG中的节点序列。

邻域子图构建：从用户初始查询的实体出发，提取KG中的邻接节点（如相关疾病、药物、副作用），构建探索子图（Subgraph\_Q）。

动态推荐生成：根据子图结构生成自然语言问题（如“Omega-3脂肪酸对其他疾病的影响？”），并通过用户反馈动态调整推荐池。

焦点+上下文设计：

逐步构建：用户每次查询新增的节点高亮显示，历史节点逐渐淡化。

导航器：显示探索路径和进度条，支持回溯历史查询并查看详细响应。

#### 逐步图形可视化

多模态交互：节点按类型（疾病、药物、文献）颜色编码，边表示关系类型（治疗、副作用）。

互补性融合：LLMs提供上下文和细节，KG提供结构化验证，形成“生成-验证-探索”闭环。

动态适应性：推荐系统根据用户行为实时调整，避免静态推荐的局限性。


### 技术亮点

抗信息过载：图形化界面通过渐进式展示，帮助用户聚焦当前任务。


# 知识库（信息提取与验证）

---

## 1.问题范围判断：

- 用户提出问题后，系统首先判断该问题是否在 KG 的范围内。如果问题不在 KG 的范围内，系统将像普通的 LLM 聊天一样工作，生成文本响应。
  - 如果问题在 **KG 范围内**，系统会进一步处理。
- 

## 2.三元组提取：

- LLM 生成文本输出后，系统会从中提取三元组（实体及其关系）。三元组的格式为 (实体1, 关系, 实体2)。
  - 例如，句子“鱼油含有丰富的 Omega-3 脂肪酸”可以提取出三元组 (鱼油, 含有, Omega-3 脂肪酸)。
  - 为了**提取三元组**，系统会提示 LLM 为每个实体和关系分配唯一的标识符。例如，(鱼油) 被标记为 n1，(Omega-3 脂肪酸) 被标记为 n2，(含有) 被标记为 r1。
- 

## 3.实体与 KG 节点匹配：

- 提取的三元组会与 KG 中的节点进行匹配。系统使用嵌入模型（如 OpenAI 的 text-embedding-ada-002）将实体和关系转化为**向量表示**。
- 通过计算**余弦相似度**，系统找到 KG 中与 LLM 输出相对应的节点。例如，如果 LLM 输出的实体 (鱼油) 与 KG 中的节点 (鱼油) 的余弦相似度超过某个阈值（如 0.94），则认为它们是匹配的。

# 知识库（信息提取与验证）

---

## 4.关系验证：

- 系统会检查 KG 中是否存在与 LLM 输出的关系相匹配的路径。例如，如果 LLM 输出的关系是 (鱼油, 含有, Omega-3 脂肪酸)，系统会检查 KG 中是否存在 (鱼油) 和 (Omega-3 脂肪酸) 之间的直接或间接路径。
- 如果 KG 中存在直接路径，系统会标记该关系为“支持”；如果存在间接路径（如两跳路径），则标记为“相关”；如果 KG 中找不到相关路径，则标记为“不确定”。



## 5.文献支持：

- 系统会从 KG 中提取相关的文献证据，帮助用户验证信息的准确性。例如，如果 KG 中存在 (鱼油, 含有, Omega-3 脂肪酸) 的关系，系统会提供支持该关系的文献列表。

# 探索跟踪模块（结构化探索推荐）

## 1. 用户状态建模：

- 系统根据用户的探索历史建模用户状态。**用户的当前状态由之前的查询序列表示**，例如  $\text{context}_t = (q_0, q_1, \dots, q_t)$ ，其中  $q_0, q_1, \dots, q_t$  是用户的时间步  $0, 1, \dots, t$  提出的查询。
- 通过这种方式，系统可以将用户的当前状态映射到 KG 中的节点或节点类型，作为生成下一步探索建议的基础。

## 2. 目标建模：

- 系统假设用户的初始查询反映了其主要目标。系统会**提取初始查询中提到的实体的单跳邻居，形成一个子图  $\text{Subgraph}_Q$** ，其中包含用户可能感兴趣的探索主题。
- 例如，如果用户的初始查询是“维生素 E 对阿尔茨海默病的影响”，系统会提取 (维生素 E) 和 (阿尔茨海默病) 的单跳邻居，形成一个子图。

## 3. 推荐生成：

- 系统根据用户的探索历史和 KG 的结构化邻域，生成下一步探索建议。例如，如果用户查询了“维生素 E 对阿尔茨海默病的影响”，系统可能会推荐“维生素 E 的抗氧化特性”作为下一步探索。
- 推荐生成基于规则模板，将候选查询转化为自然语言问题。例如，系统可能会生成“维生素 E 的抗氧化特性如何影响神经元健康？”这样的推荐问题。

## 4. 用户反馈机制：

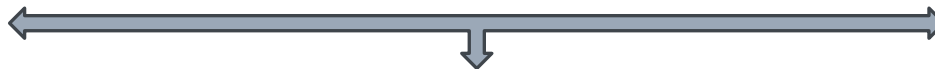
- 用户可以通过反馈机制调整推荐。如果用户对某个推荐不感兴趣，可以移除该推荐；如果用户有新的查询需求，可以手动输入新的查询。
- 系统会根据用户的反馈动态更新推荐池，移除不相关的查询，并添加新的建议。



# 用户界面 (逐步图形可视化)

---

KnowNET 的用户界面包括三个主要组件：



## 1. 文本对话：

- 显示 LLM 的文本响应，并高亮显示相关的实体和关系。例如，句子“鱼油含有丰富的 Omega-3 脂肪酸”中，(鱼油) 和 (Omega-3 脂肪酸) 会被高亮显示。
- 用户可以通过点击高亮的实体，查看其在 KG 中的详细信息。

## 2. 图形探索器：

- 将文本响应转化为图形表示，帮助用户直观地理解实体之间的关系。例如，(鱼油) 和 (Omega-3 脂肪酸) 之间的关系会以节点和边的形式显示在图形探索器中。
- 用户可以通过点击节点查看详细信息，并通过拖拽节点调整布局。

## 3. 导航器：

- 允许用户回顾之前的查询，查看探索进度，并根据推荐提交新的查询。
- 导航器使用线性布局，每个点代表一个查询。用户可以通过点击点来查看特定查询的响应和图形表示。



# 技术亮点

---

互补性融合：LLMs提供上下文和细节，KG提供结构化验证，形成“生成-验证-探索”闭环。

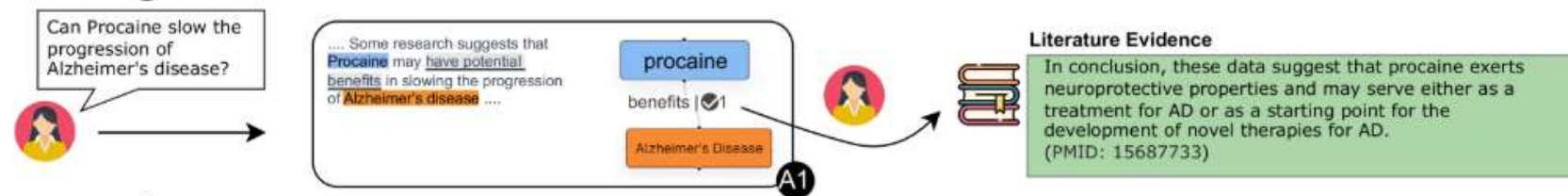
动态适应性：推荐系统根据用户行为实时调整，避免静态推荐的局限性。

抗信息过载：图形化界面通过渐进式展示，帮助用户聚焦当前任务。

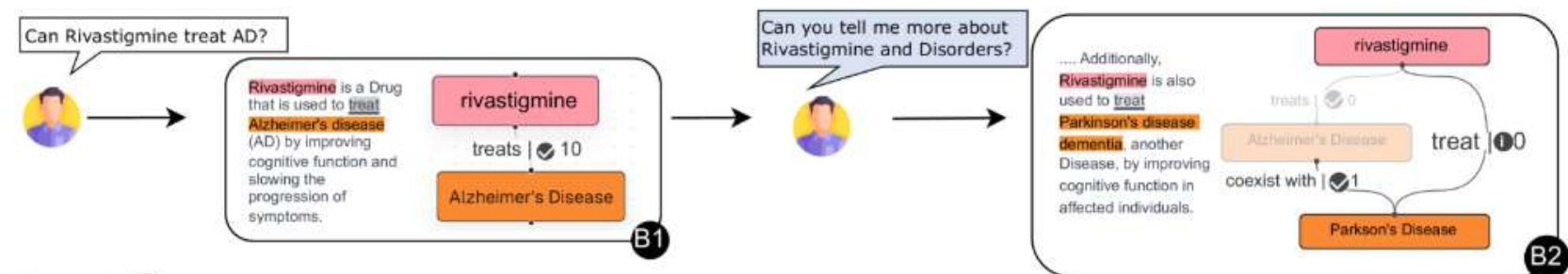
# 评估

为了评估 KNOWNET 的有效性，作者设计了 15 个使用案例 和 专家访谈，重点关注系统在 准确性、可解释性和探索引导性 方面的表现。

### Case 1 a



### Case 2 b



### Case 3 c

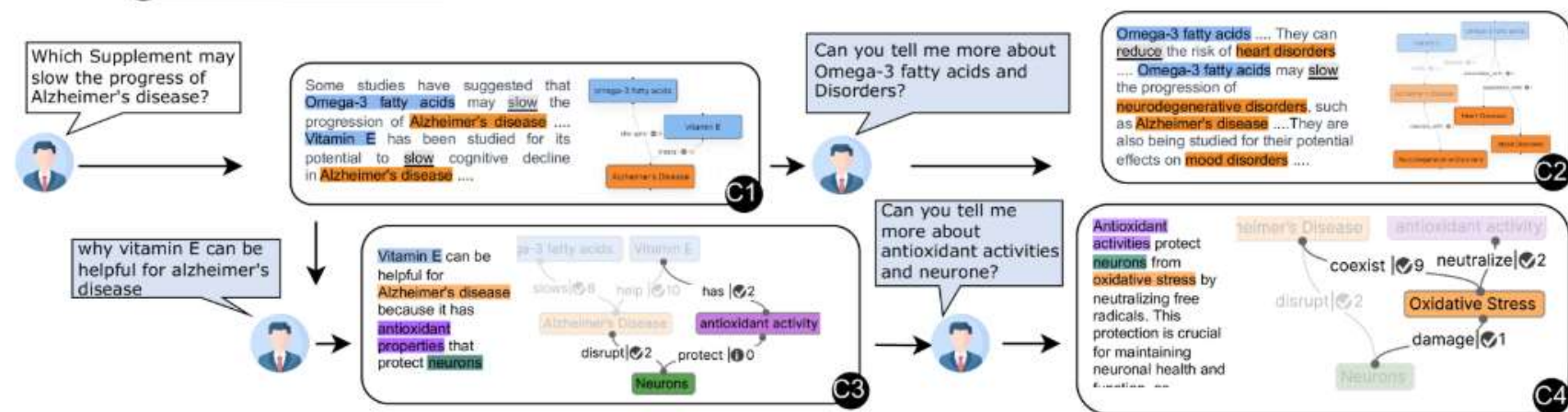
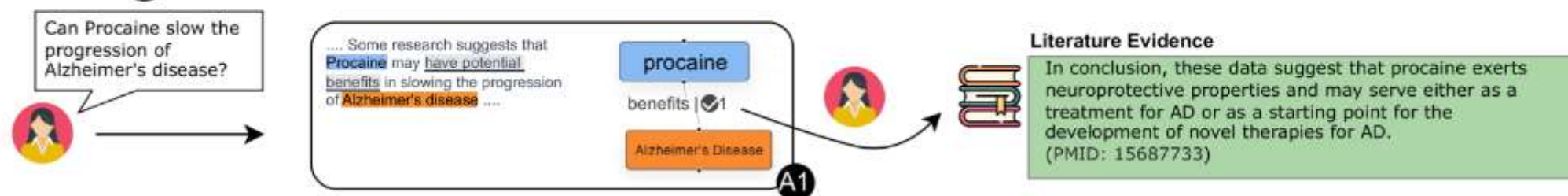


图 7：三个典型用例。蓝色聊天气泡表示推荐问题，白色聊天气泡表示免费问题。

## Case 1 a



### 案例 1：文献验证

用户询问 “普鲁卡因是否能延缓阿尔茨海默症？”

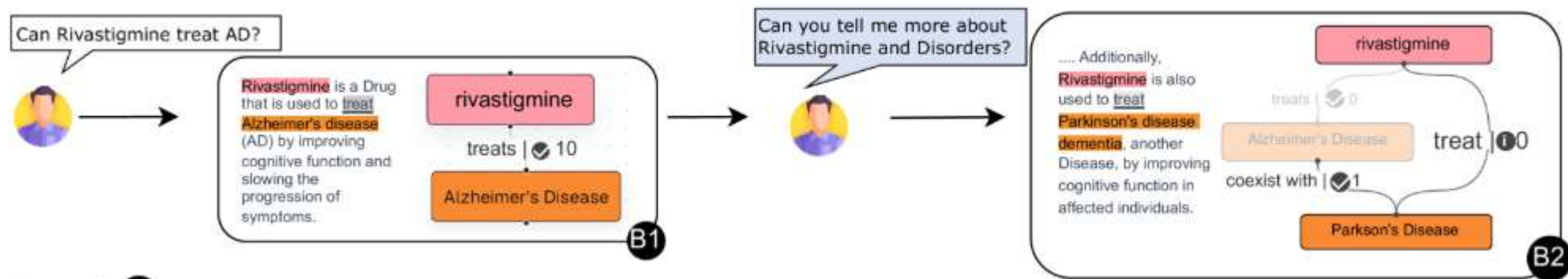
KNOWNET 返回：

LLM 回答：“研究表明普鲁卡因可能有助于减缓阿尔茨海默症的进展。”

KG 证据：该关系被 KG 文献支持，标记 ✓。

用户可点击查看文献，确认研究依据。

## Case 2 b



### 案例 2：LLM + KG 结合

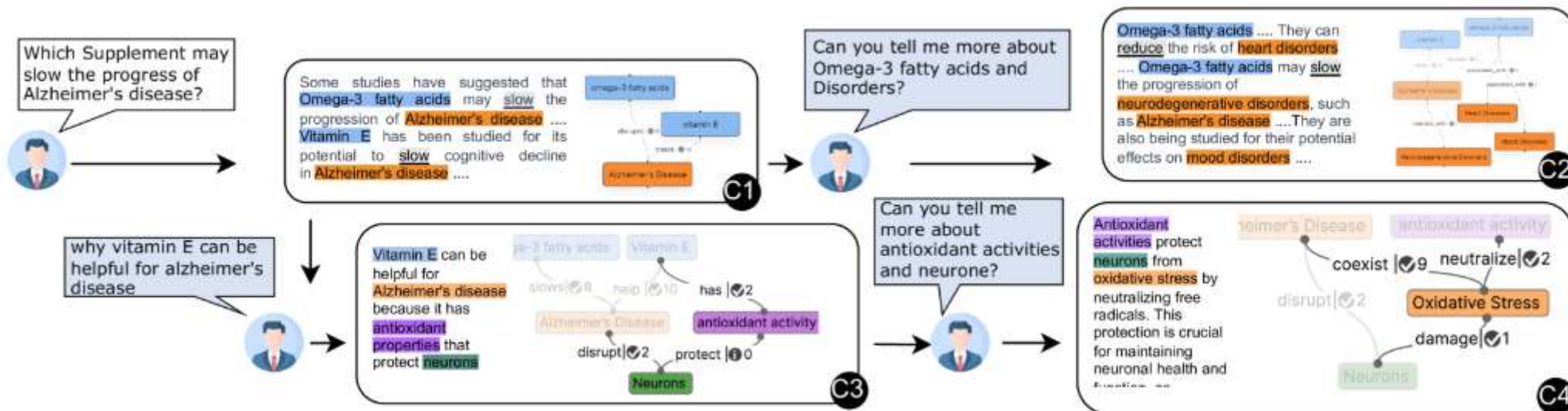
用户询问 “利伐斯的明能治疗帕金森病吗？”

KG 中未直接存储该关系，但：

LLM 预测：Rivastigmine 可用于治疗 Parkinson's Disease。

KG 发现：“Rivastigmine → 治疗 → Alzheimer's Disease → 相关 → Parkinson's Disease”  
系统标记为 ‘相关’ (i)，用户可进一步探索。

### Case 3 C



案例 3：引导式探索

用户最初问 “哪些补充剂可能有助于阿尔茨海默症？”

KNOWNET 推荐：

Omega-3 脂肪酸 和 维生素 E。

用户点击推荐问题 “为什么维生素 E 对阿尔茨海默症有帮助？”

KNOWNET 进一步展示 抗氧化作用如何保护神经元。

# 专家访谈反馈

---

- 1.系统直观易用，无需额外学习成本。
- 2.图可视化有助于信息组织，比 GPT 纯文本更清晰。
- 3.文献支持增强了可信度，有助于提高 AI 可信性。

除了我们作者团队中的三位领域专家外，我们还采访了另外两位专门从事计算健康信息学和医学研究的专家（E1，E2）。E1 是医学信息学的研究科学家，拥有博士学位，在药学和 EHR 数据分析方面拥有 8 年的经验。E2 是健康信息学博士生，拥有 4 年的研究经验。E2 拥有强大的临床背景，拥有 MBBS 学位，相当于美国的 MD。这些访谈旨在深入了解的可用性、有效性和需要改进的领域。另外两位专家不是本文的作者。每次面谈持续约 60 分钟。

简介（10 分钟）。我们首先提供了背景概述并演示了的各种组件和功能。

案例介绍（20 分钟）。在此介绍之后，我们介绍了两个特定的用例，我们还提供了 GPT-4 对相同问题的回答以进行比较。此部分旨在让专家熟悉的用法和 workflows。在演示期间，鼓励专家以问题和评论打断。

自由探索（15 分钟）。我们允许专家对系统进行 15 分钟的自由探索，在此期间鼓励他们大声思考并说出自己的想法。我们密切观察了这次互动会议，我们的团队详细记录了专家与的互动。

讨论（15 分钟）。访谈以半结构化讨论结束，在此期间，我们收集了五个关键方面的反馈：所提供信息的准确性、解释的清晰度、回复和建议的相关性、基本主题的覆盖范围以及系统的整体易用性。



# 结论

KNOWNET 通过 LLM + KG 结合的方式，提供更准确、结构化的健康信息检索体验：

- 1.提取 LLM 结构化信息，并与 KG 进行匹配校验。
- 2.提供下一步推荐，避免信息孤立，提高探索深度。
3. 采用对话+图可视化+导航结合的方式，使信息更直观可用。

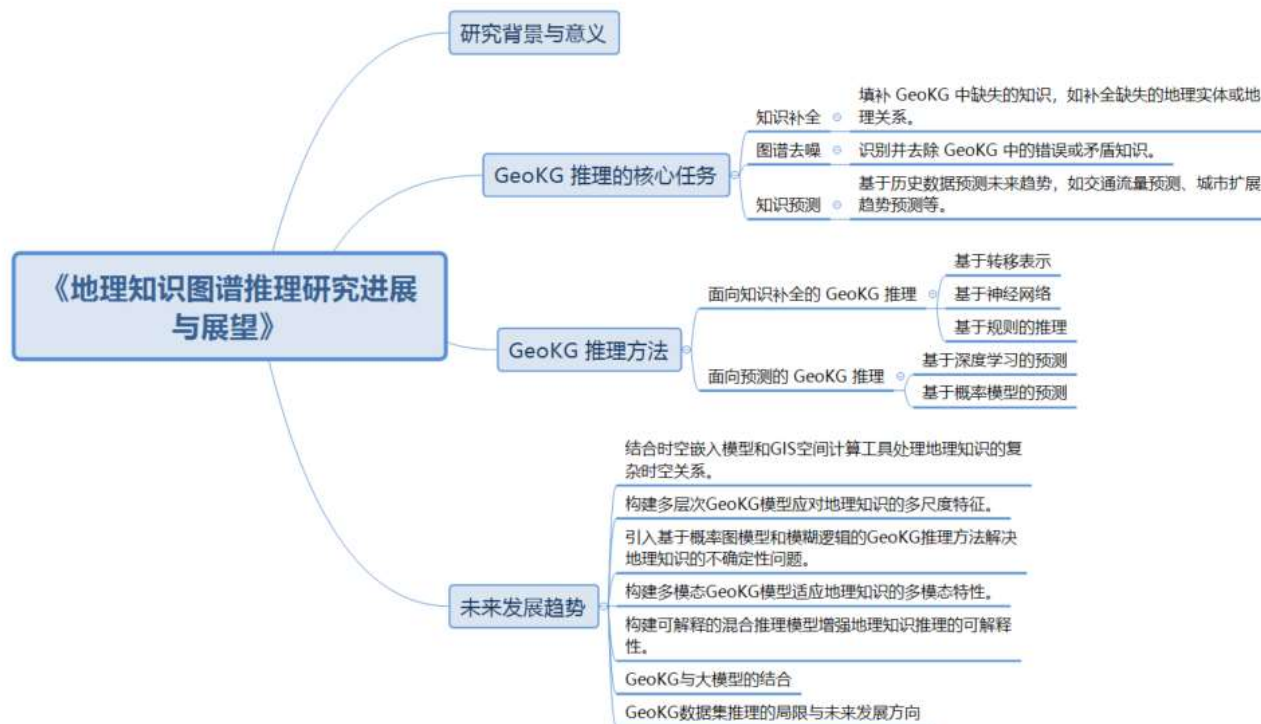
该方法可扩展到其他领域，如 医学、法律、金融等知识密集型应用，未来可进一步优化推荐算法，并支持更广泛的知识库整合。

## KNOWNET 方法解析：通俗易懂版

KNOWNET 是一个智能健康信息查询系统，它结合了大语言模型（LLM）和知识图谱（KG），帮助用户更准确地获取健康相关信息，同时提供清晰的结构化探索路径，避免信息混乱或错误。

# 《地理知识图谱推理研究进展与展望》

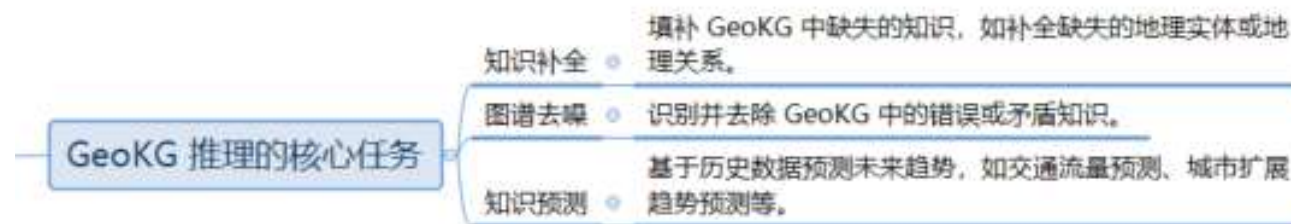
该文章《地理知识图谱推理研究进展与展望》系统地综述了地理知识图谱 (GeoKG) 的推理技术, 包括其概念、关键任务、研究进展以及未来发展趋势。



# 研究背景与意义

地理知识图谱 (GeoKG) 是知识图谱 (KG) 与地理信息科学 (GIS) 结合的产物。由于地理知识往往涉及复杂的时空特性, GeoKG 推理不仅需要补全缺失知识, 还要处理地理数据的时空动态变化。本文主要研究 GeoKG 推理的关键任务、方法及未来发展趋势, 以提高地理信息的完整性、准确性和预测能力。

# GeoKG 推理的核心任务



# GeoKG 推理的核心任务

---

GeoKG 推理主要包括三大任务：

**知识补全（GeoKG Completion）**：填补 GeoKG 中缺失的知识，如补全缺失的地理实体或地理关系。例如，补全一个城市的缺失地理属性（如人口、气候）。

**图谱去噪（GeoKG Denoising）**：识别并去除 GeoKG 中的错误或矛盾知识。例如，检测并修正错误的城市归属信息。

**知识预测（GeoKG Prediction）**：基于历史数据预测未来趋势，如交通流量预测、城市扩展趋势预测等。

# 知识补全

## 1 实体补全 (Entity Completion)

给定头节点 (head entity) 和关系 (relation) , 推理出缺失的尾节点 (tail entity) , 补全三元组 (h, r, ?)。

$$f_{\text{GeoKG}}(h, r, ?) \rightarrow t$$

已知 (长江, 流经, ?) , 推理出 (长江, 流经, 武汉)。

已知 (北京市, 是, ?) , 推理出 (北京市, 是, 中国首都)。

## 2 关系补全 (Relation Completion)

给定头节点和尾节点, 推理出它们之间的关系 (h, ?, t)。

$$f_{\text{GeoKG}}(h, ?, t) \rightarrow r$$

已知 (上海, ?, 黄浦江) , 推理出 (上海, 被流经, 黄浦江)。

已知 (广州, ?, 广东省) , 推理出 (广州, 属于, 广东省)。

## 3 三元组间关系补全 (Inter-Triple Relation Completion)

基于多个三元组的逻辑链, 推理出新的关系。

$$f_{\text{GeoKG}}\{(h_1, r_1, t_1), (h_2, r_2, t_2), \dots\} \rightarrow (h_n, r_n, t_n)$$

已知 (A, 位于, B) 和 (B, 位于, C) , 推理出 (A, 位于, C) (空间层级推理)。

已知 (台风, 影响, 福建) 和 (福建, 毗邻, 浙江) , 推理出 (台风, 可能影响, 浙江) (空间邻近推理)。

# 知识去噪

## 1 矛盾检测 (Contradiction Detection)

识别图谱中相互矛盾的三元组。

示例：

矛盾三元组：(上海, 属于, 江苏省) 和 (上海, 属于, 直辖市)。

解决方法：通过空间层级规则（如“一个城市只能属于一个省级行政区”）修正。

## 2 错误修正 (Error Correction)

基于已有知识或外部数据源，修正错误的三元组。

示例：

错误数据：(黄河, 流经, 广州)（实际不流经）。

修正方法：结合水系数据库或空间拓扑关系验证并修正。

## 3 冗余消除 (Redundancy Removal)

去除重复或冗余的三元组。

示例：

冗余数据：(北京, 是, 中国首都) 和 (中国首都, 是, 北京)。

解决方法：通过语义等价性检测去重。



# 知识预测

## 1 实体预测 (Entity Prediction)

定义：预测未来可能新增的地理实体。

$$f_{\text{EntityPrediction}}(\text{GeoKG}, \text{history}, t_{\text{future}}) \rightarrow \{h_1, h_2, \dots, h_n\}$$

示例：

预测未来10年可能新增的城市（如雄安新区）。

预测气候变化可能导致的沙漠化区域。

## 2 关系预测 (Relation Prediction)

定义：预测未来实体间可能出现的新关系。

$$f_{\text{RelationPrediction}}(h, t, t_{\text{future}}) \rightarrow r$$

示例：

预测未来某条高速公路是否会连接两个城市。

预测台风路径是否会经过某个地区。

## 3 时空属性预测 (Spatio-Temporal Attribute Prediction)

定义：预测地理实体的属性随时间或空间的变化。

示例：

预测某城市未来5年的人口密度变化。

预测某河流未来3个月的水位变化。

$$f_{\text{AttributePrediction}}(h, t_{\text{future}}) \rightarrow x$$

# 总结

核心任务	子任务	目标	典型应用
知识补全	实体补全	填补缺失的尾节点	POI补全、道路网络完善
	关系补全	推断实体间关系	空间层级推理（如城市→省→国家）
	三元组间关系补全	基于逻辑链推理新关系	跨区域影响推理（如台风路径）
知识去噪	矛盾检测	识别错误标注	修正行政区划错误
	错误修正	修正错误三元组	结合遥感数据修正水系关系
	冗余消除	去除重复数据	语义等价性检测
知识预测	实体预测	预测新增实体	城市扩张模拟
	关系预测	预测新关系	交通网络发展预测
	时空属性预测	预测属性变化	气候变化、人口密度预测

# GeoKG推理方法



# 面向知识补全的 GeoKG 推理

包括以下三类方法：

## 基于转移表示（Embedding-based）

通过将地理实体和关系转换为低维向量进行推理，例如 TransE-GDR、TransD-GDR、VerE-TransD 等。这些方法可有效捕捉地理空间特征，但对于复杂关系的处理能力有限。

## 基于神经网络（Neural Network-based）

采用图神经网络（GNN）和递归神经网络（RNN）进行推理，例如 ST-GCN（时空图卷积网络）。这类方法适用于时空依赖较强的任务，如交通流量预测，但计算复杂度较高。

## 基于规则的推理（Rule-based）

依靠预定义规则进行推理，如森林火灾分析、水稻病虫害诊断等。该方法可解释性强，但灵活性差，依赖领域专家知识。

## (1) 基于平移距离的模型

### •TransE-GDR

- 原理：扩展经典TransE模型，引入地理空间距离约束（如欧式距离）。
- 公式：  $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\| \leq d_{\text{geo}}(h, t)$
- 优点：简单高效，适合处理简单空间关系（如“相邻”）。
- 缺点：难以建模复杂关系（如“包含”“交叉”）。

### •TransR-GDR

- 改进：将实体和关系投影到不同空间，通过关系特定矩阵  $\mathbf{M}_r$  捕捉复杂语义。
- 公式：  $\|\mathbf{h}\mathbf{M}_r + \mathbf{r} - \mathbf{t}\mathbf{M}_r\|$
- 优点：可处理多对多关系（如“河流流经多个城市”）。
- 缺点：计算复杂度高，训练时间长。

### •GEUKE

- 特点：统一编码地理实体的空间结构特征（如点、线、面），确保不同类型实体嵌入的一致性。
- 应用：用于POI补全，性能优于传统TransE。

基于转移表示的方法

## (2) 基于语义的模型

### •DistMult/Complex

- 原理：利用张量分解（如三元组得分函数）建模对称/非对称关系。
- 适用场景：基于文本数据的地理关系挖掘（如从新闻中提取“城市-事件”关系）。

## (1) 图神经网络 (GNN)

### •ST-GCN (时空图卷积网络)

- 结构：结合图卷积 (GCN) 和时间卷积 (TCN)，聚合邻域节点的时空特征。
- 应用：交通流量预测、城市拥堵分析。

### •GAT (图注意力网络)

- 改进：通过注意力机制动态分配邻居节点权重，处理异构图（如道路网络+天气数据）。
- 案例：RAG-GAT模型用于自动驾驶中的实时路径规划。

## (2) 序列模型 (RNN/LSTM)

### •STSE/S-TSE

- 原理：使用LSTM编码时间序列，结合空间嵌入完成知识补全。
- 示例：预测台风路径的时空演化。



## 基于规则的方法

### (1) 空间规则

- 邻近规则：若实体A与B相邻，则可能存在关系R（如“共享边界”）。
- 层级规则：若“城市 $\in$ 省”且“省 $\in$ 国家”，则“城市 $\in$ 国家”。

### (2) 领域知识驱动

- 案例1：森林火灾知识图谱（Ge等，2022）通过规则“高温+干旱 $\rightarrow$ 火灾风险高”补全缺失属性。
- 案例2：滑坡预测（Chen等，2023）结合地质规则（如坡度 $>30^\circ$ 易滑坡）修正错误数据。

# 面向预测的 GeoKG 推理

## 基于深度学习的预测

采用 GCN、GAT、RNN 等深度学习方法，如 KG-ST-CNN、Graph-Flashback、DynAttGraph2Seq。  
适用于交通、气象等时空动态变化较大的领域。

## 基于概率模型的预测

采用隐马尔可夫模型（HMM）、动态贝叶斯网络（DBN）等，如 KG-MTKF。  
适用于预测暴雨灾害链、城市扩展等长期趋势。

## (1) 时空图模型

### •KST-GCN（知识驱动的交通预测时空图卷积网络）

- 结构：融合知识图谱与交通数据，通过GRU单元建模时间动态性。
- 效果：在交通流量预测中优于纯GCN模型。

### •KG-ST-CNN

- 创新：将知识图谱与时空卷积结合，预测城市拥堵区域的时空状态。

## (2) 动态演化模型

### •Know-Evolve

- 原理：基于循环神经网络（RNN）和点过程建模实体状态的连续变化。
- 应用：预测地理事件（如地震）的连锁反应。

## (1) 隐马尔可夫模型 (HMM)

- 示例：暴雨灾害链预测（黄晶等），计算次生灾害的发生概率。

## (2) 动态贝叶斯网络 (DBN)

### •KG-MTKF

- 特点：结合卡尔曼滤波与知识图谱，处理非线性地理系统（如水库水位变化）。

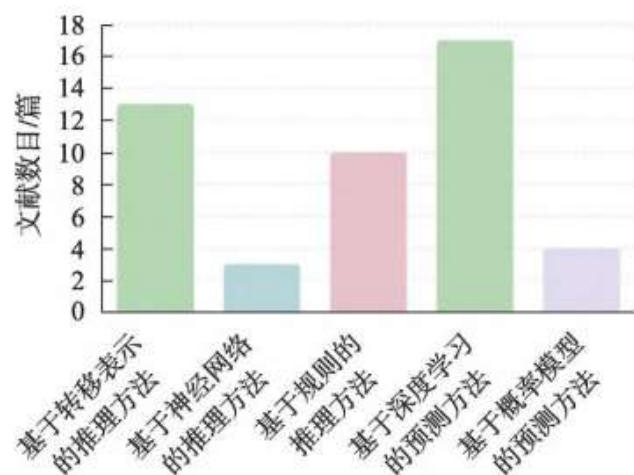


图2 地理知识图谱推理各类算法计量

# 未来发展趋势

## 未来发展趋势

结合时空嵌入模型和GIS空间计算工具处理地理知识的复杂时空关系。

构建多层次GeoKG模型应对地理知识的多尺度特征。

引入基于概率图模型和模糊逻辑的GeoKG推理方法解决地理知识的不确定性问题。

构建多模态GeoKG模型适应地理知识的多模态特性。

构建可解释的混合推理模型增强地理知识推理的可解释性。

GeoKG与大模型的结合

GeoKG数据集推理的局限与未来发展方向

## 1. 复杂时空关系的动态建模

地理知识图谱需要更强大的模型来捕捉动态时空关系，如距离衰减效应、方位关系和拓扑变化的交互影响。现有方法（如TransE-GDR）难以处理非线性演化问题（如台风路径突变），未来需结合物理机制（如流体动力学）或时序图神经网络（如ST-GNN），以实现更精准的时空推理。

## 2. 多尺度地理知识的统一表达

地理现象在不同尺度（如全球气候与城市热岛）表现出不同规律，而当前GeoKG多针对单一尺度设计，跨尺度推理易出现语义断层。未来需构建分层知识图谱，并开发跨尺度注意力机制，以支持“宏观-微观”协同分析（如疫情传播模拟）。

## 3. 地理知识的不确定性处理

地理数据常含模糊性（如城乡边界）和误差（如遥感分类错误），但传统确定性模型无法量化置信度。未来需融合概率图模型（如ProbKG）和模糊逻辑，为推理结果提供概率评估（如滑坡风险概率为70%）。

## 4. 多模态数据融合的瓶颈

GeoKG需整合文本、影像、传感器等多源数据，但异构性和时空对齐问题阻碍有效融合。未来需借助跨模态预训练模型（如GeoBERT）和统一嵌入空间，实现“卫星影像+社交媒体”的联合推理。

## 5. 模型可解释性与领域适配性

深度学习模型（如GAT）的“黑箱”特性限制了其在灾害预警等关键场景的应用。未来需发展神经符号系统（Neuro-Symbolic AI），结合规则引擎与深度学习，并生成可视化推理路径以增强可信度。

## 6. GeoKG与大模型的协同

大模型（如GPT-4）的语义理解能力可辅助GeoKG补全与验证，但需解决幻觉问题。未来应构建地理指令微调数据集（如GeoInstruct），训练领域大模型（如GeoGPT），实现“假设生成-知识验证”闭环。

## 7. 标准化评估与开放生态

当前缺乏统一的GeoKG推理基准和跨领域数据集。未来需建立开源评估平台（如GeoBench），涵盖时空、多模态和多尺度任务，并推动社区协作以加速技术落地。



本文系统梳理了 GeoKG 推理的研究进展，介绍了核心任务、方法、应用及未来发展趋势。GeoKG 推理在地理信息系统、智慧城市、环境监测等领域具有广泛的应用前景。未来研究可重点关注多模态数据融合、时空动态推理、可解释性增强等方向，以提升 GeoKG 推理的实用性和准确性。

谢谢

谢谢