

大模型常用提示词技术介绍



目录

CONTENTS



01 | 大语言模型提示词工程原理

02 | 大语言模型提示词工程的基本原则

03 | 少样本提示技术与医学问题实操

04 | 链式思考技术与医学问题实操

01

大语言模型提示词工程原理

什么是大模型提示词

基本概念：给大模型提示

- 提示词是指传递到模型的**指令**或**问题**等信息也可以包括其他详细信息，如**上下文**、**输入**或**样例**等。
- 通过不同的提示词（Prompts）可以获得不同的结果，并且结果的质量与你提供的信息密切相关。

The dog is



It seems like your sentence got cut off! Could you clarify or complete your thought? For example, are you asking about the dog's behavior, appearance, or something else? Let me know how I can help!

缺乏提示词描述大模型无法理解你的意图

Complete the sentence:

The dog is



Sure! Here are a few ways to complete the sentence:

- The dog is barking loudly at the mailman.
- The dog is sleeping peacefully on the couch.
- The dog is chasing its tail in circles.

提示词帮助大模型按照你的指令回答你的问题

医学任务应用



面向患者的“AI”医生

疾病科普
疾病诊断
报告分析

医学任务应用

面向医生的智能助手

临床辅助诊断、科研资料检索、.....

The screenshot displays a medical application interface. At the top, there is a '基本信息' (Basic Information) section with fields for name, gender, age, and other patient details. Below this is a navigation bar with tabs for '处方' (Prescription), '检验' (Lab Tests), '检查' (Examination), '手术' (Surgery), and '护理' (Nursing). The main content area is divided into sections for '主诉' (Chief Complaint), '现病史' (Present History), '既往史' (Past History), '体格检查' (Physical Examination), '辅助检查' (Auxiliary Examination), '治疗意见' (Treatment Advice), '处方医嘱' (Prescription Instructions), and '其他' (Other). The '主诉' and '现病史' sections contain text describing a patient's symptoms. On the right side, there is a sidebar with a search bar and a '临床辅助诊断' (Clinical Auxiliary Diagnosis) section, which includes a list of related conditions and a brief description of the patient's condition.

大模型的幻觉现象

基本概念

- 大模型幻觉（Hallucination）指模型生成看似合理但包含事实错误、逻辑矛盾或无法验证的内容
- 事实性幻觉：事实不一致（Factual Inconsistency），违背客观常识
- 事实捏造（Factual Fabrication），AI大模型凭空编造不存在的事实，且无法通过任何可靠来源验证。



肺性脑病



肺性脑病（Hepatic Encephalopathy, HE）是一种由于肝脏功能严重受损或门体分流导致的代谢紊乱，进而引发中枢神经系统功能障碍的临床综合征。它是肝硬化和肝功能衰竭患者常见的并发症之一，严重时可危及生命。

病因与发病机制

1. 肝脏功能衰竭：肝脏无法有效清除体内的氨和其他毒性物质。
2. 门体分流：门静脉高压导致血液绕过肝脏直接进入体循环，使毒性物质进入大脑。

在DeepSeek的界面上输入“肺性脑病”，结果出来的却是关于“肝性脑病”的介绍

提示技术的重要性



精准可控输出：通过提示技术，可以引导大模型生成更精准、更符合需求的输出。

解决“幻觉”问题：提示技术能够有效减少大模型的“幻觉”现象（即“胡说八道”），提高输出的准确性和可靠性。



知识局限性突破：借助提示技术，大模型可以更好地处理知识局限性问题，提升其在不同领域的应用价值。

02

大语言模型提示词工程的基本原则

提示词类型：基于不同场景的设计



提示词场景1：双方都知道的情况

想象你在问AI“北京是中国的首都吗？”这类问题。因为这是常识性知识，双方都清楚，所以：

- 建议：直接了当地提问
- 无需复杂解释
- 保持提示词简短精准



提示词场景2： 你知道但AI不知道的情况

比如你想让AI帮你分析你们公司的内部数据，或者是你自己的个人经历：

- 需要你提供具体的背景信息
- 最好给AI几个例子让它理解你想要的格式和风格
- 使用"few-shot learning"（给出示例）来引导AI理解



提示词场景3： AI知道但你不知道的情况

这就像你在问AI一些专业知识，比如“量子物理的基本原理是什么”：

- 清晰地表达你的困惑
- 可以要求AI用通俗的语言解释
- 分步骤提问，由浅入深



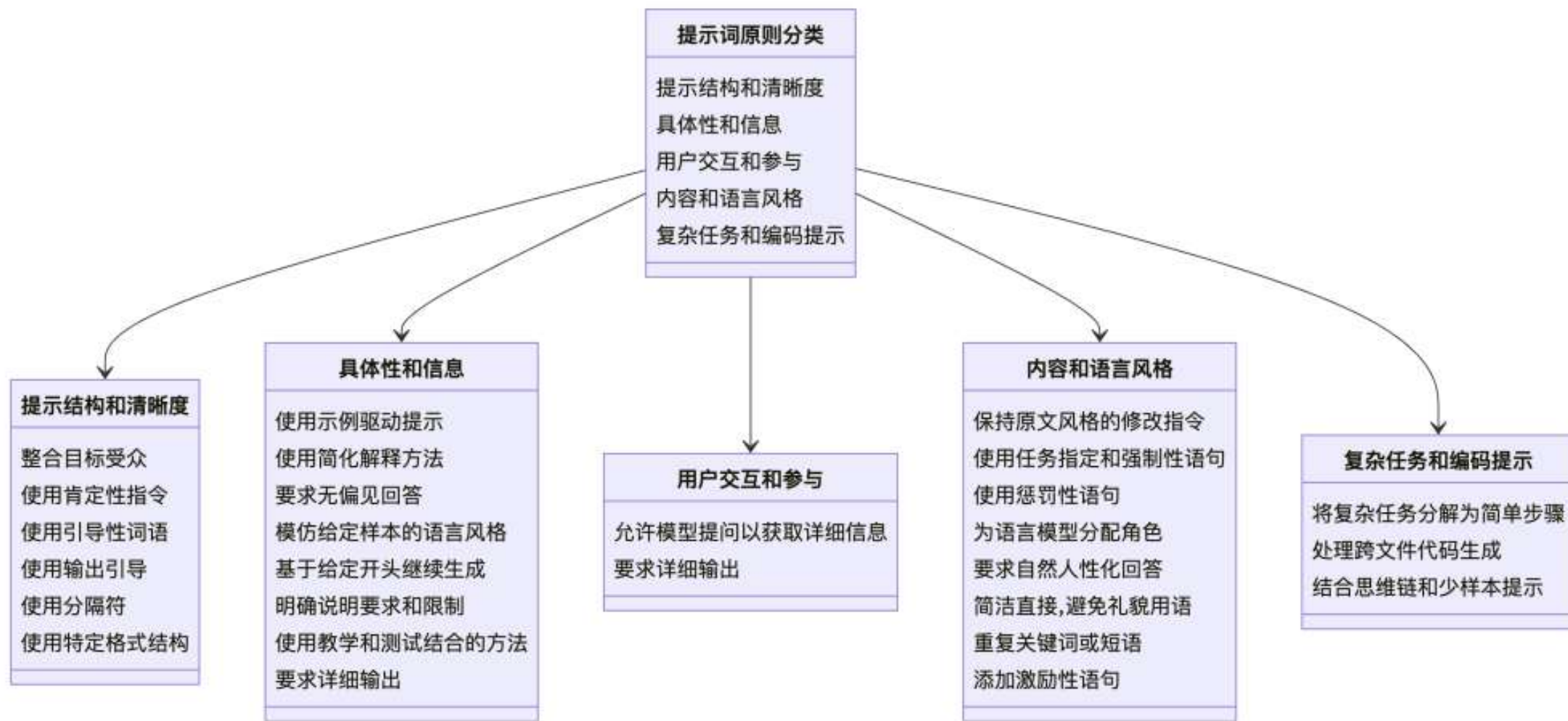
提示词场景4： AI不知道你也不知道的情况

这通常涉及创新性的问题，比如尚未解决的科学难题：

- 可以和AI一起探讨可能的解决方案
- 通过交互式对话激发新思路
- 把问题拆分成小块，逐步探索



提示词设计原则归纳



提示词设计原则

基本概念

- 提示工程是一门新兴的技术，它关注如何通过精心设计的输入（提示）来引导大语言模型产生所需的输出。一篇论文的研究者们发现，通过应用特定的提示原则，可以显著提高模型输出的质量和准确性。

序号	原则	案例	说明
1	简洁直接， 避免礼貌用语	不用："请问您能帮我解释一下气候变化吗？" 用："解释气候变化。"	直接陈述需求，避免不必要的礼貌用语，以获得更简洁的回答
2	明确目标受众	"解释量子力学，假设受众是高中生。"	指定目标受众有助于模型调整回答的复杂度和用词
3	将复杂任务分解为简单步骤	"1. 列出做披萨的材料 2. 解释制作步骤 3. 提供烘烤建议"	将复杂任务分解可以获得更详细和结构化的回答

[参考文献] Bsharat, Sondos Mahmoud, Aidar Myrzakhan, and Zhiqiang Shen. "Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4."

提示词设计原则

4	使用肯定性指令	不用: "不要使用专业术语" 用: "使用日常用语解释"	肯定性指令更容易被模型理解和执行
5	使用简化解释方法	"像解释给5岁孩子一样, 描述光合作用。"	要求简化解释可以得到更容易理解的回答
6	添加激励性语句	"我会为最佳答案给予\$100奖励! 解释相对论。"	虽然模型不会真的获得奖励, 但这可能会触发更详细的回答
7	使用示例驱动提示	"示例: 苹果是红色的。 香蕉是什么颜色?"	提供示例可以引导模型以特定方式回答
8	使用特定格式结构	"###指示### ###示例### ###问题###"	使用特定格式可以帮助模型更好地组织信息
9	使用任务指定和强制性语句	"你的任务是...你必须..."	明确指定任务和要求可以提高回答的针对性

[参考文献] Bsharat, Sondos Mahmoud, Aidar Myrzakhan, and Zhiqiang Shen. "Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4."

提示词设计原则

10	使用惩罚性语句	"如果不提供准确信息，你将受到惩罚。"	这可能会让模型更加谨慎，提供更准确的信息
11	要求自然人人性化回答	"用自然、类人的方式回答以下问题。"	这可以使回答更加口语化和易于理解
12	使用引导性词语	"逐步思考：如何解决这个数学问题？"	引导模型展示思考过程，有助于获得更详细的解答
13	要求无偏见回答	"提供一个无偏见的、不依赖刻板印象的回答。"	有助于获得更客观、公正的回答
14	允许模型提问以获取详细信息	"如果需要更多信息来回答，请问我问题。"	这可以模拟真实对话，获得更精确的回答
15	使用教学和测试结合的方法	"解释光合作用，然后对我进行测试。"	这种方法可以帮助验证理解，并获得更深入的解释
16	为语言模型分配角色	"作为一名经验丰富的厨师，解释如何制作完美的意大利面。"	角色扮演可以引导模型从特定视角回答问题

[参考文献] Bsharat, Sondos Mahmoud, Aidar Myrzakhan, and Zhiqiang Shen. "Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4."

提示词设计原则

17	使用分隔符	"`` 请在这里写下你的回答`` `"`	分隔符有助于清晰地区分指令和回答
18	重复关键词或短语	"重要! 重要! 重要! 解释全球变暖的影响。"	重复可以强调某些信息的重要性
19	结合思维链和少样本提示	"示例: 问题: $2+2=?$ 思考: 我知道2代表两个单位。两个两个单位加在一起就是四个单位。 答案: 4 现在, 使用同样的思考方式回答: $7*8=?$ "	这种方法可以引导模型展示详细的推理过程
20	使用输出引导	"完成这个句子: 太阳系中最大的行星是..."	这种方法可以引导模型产生特定格式或内容的输出
21	要求详细输出	"详细描述光合作用的过程, 包括所有必要的信息。"	这可以鼓励模型提供更全面、深入的回答

[参考文献] Bsharat, Sondos Mahmoud, Aidar Myrzakhan, and Zhiqiang Shen. "Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4."

提示词设计原则

22	保持原文风格的修改指令	"修改以下段落, 仅改善语法和词汇, 保持原有写作风格。"	这在需要保持特定风格但改善文本质量时很有用
23	处理跨文件代码生成	"生成一个Python脚本, 自动创建和修改多个文件。"	这对于复杂的编程任务很有帮助
24	基于给定开头继续生成	"继续这个故事: 很久很久以前, 在一个遥远的王国里..."	这可以引导模型按特定方向继续创作
25	明确说明要求和限制	"写一篇关于人工智能的文章, 字数在500-600之间, 包括利弊分析。"	明确的要求可以帮助获得更符合预期的输出
26	模仿给定样本的语言风格	"使用与以下段落相似的语言风格写一篇关于月球的短文: [插入样本文本]"	这可以帮助生成特定风格或语气的文本

提示词设计原则归纳

简洁就是力量

- 在与AI交互时，我们往往会不自觉地加入一些礼貌用语，比如"请"、"谢谢"等。然而，研究发现直接了当的指令往往能够获得更好的结果。例如，与其说"请您能否解释一下量子力学的基本原理？谢谢！"，不如直接说"解释量子力学的基本原理。"这种方式不仅节省了字数，还能让模型更专注于核心任务。

角色扮演：释放AI的创造力

- 给AI分配一个特定的角色可以极大地增强其输出的相关性和深度。这就是第16条原则的精髓。想象你需要了解19世纪维多利亚时代的社会习俗。你可以这样提问："作为一位19世纪的英国贵族，描述一次典型的晚宴。"这样的提示会引导模型从一个特定的视角出发，提供更加丰富和符合历史背景的描述。

思维链：透视AI的决策过程

- 第19条原则提到了结合思维链（Chain of Thought）和少样本提示的方法。这种方法不仅能够提高模型输出的质量，还能让我们窥见AI的"思考"过程。

03

少样本提示 (Few-Shot Prompting) 与医学问题实操

零样本提示

基本概念

- 原理：经过大量数据训练并调整指令的大模型能够执行许多零样本任务。
- 我们无需提供任务和问题相关的样例让大模型学习，只需要明确指令即可获得准确的输出

小明有5个苹果，吃掉2个后买了3个，现在有多少？

问题：小明有5个苹果，吃掉2个后买了3个，现在有多少？

解答：

1. 初始数量：小明有5个苹果。
2. 吃掉2个后：5个苹果 - 2个苹果 = 3个苹果。
3. 买了3个后：3个苹果 + 3个苹果 = 6个苹果。

最终答案：

$\boxed{6}$

简单的数学计算任务

简单的文本分类任务

将文本分类为中性、负面或正面。
文本：我认为这次假期还可以。
情感：

情感：中性

解释：文本中的“还可以”表达了一种中等程度的满意度，既没有强烈的积极情绪，也没有明显的消极情绪，因此归类为中性。



少样本提示

基本概念

- 虽然大型语言模型展示了惊人的零样本能力，但它们在更复杂的任务上仍然表现不佳。
- 少样本提示可以作为一种技术，以启用上下文学习，我们在提示中提供演示以引导模型实现更好的性能。

这太棒了! // Negative
这太糟糕了! // Positive
哇, 那部电影太棒了! // Positive
多么可怕的节目! //

 Negative

可以从样例中学习任务要求：“文本情感分类”

“farduddle”是指快速跳上跳下。一个使用farduddle这个词的句子例子是：

 以下是一个使用“farduddle”这个词的句子示例：

“The children were so excited about the snow that they started to farduddle all over the yard, laughing and jumping up and down with joy.”

这个句子描述了孩子们因为下雪而兴奋地在院子里快速跳上跳下，展现了“farduddle”这个词的含义。



“whatpu”是坦桑尼亚的一种小型毛茸茸的动物。一个使用whatpu这个词的句子例子是：
我们在非洲旅行时看到了这些非常可爱的whatpus。

“farduddle”是指快速跳上跳下。一个使用farduddle这个词的句子例子是：

 我们在非洲旅行时看到了这些非常可爱的whatpus，它们正在草地上farduddle，看起来非常活泼和有趣。



可以从样例中限制LLM回答任务时的思考方式：“造中文例子”

少样本提示的作用

1. 任务定义与格式规范

- 这是最基本也最重要的作用。告诉模型：“请按照我展示的方式来完成”。
 - **明确任务边界**：一个模糊的指令如“处理文本”，模型不清楚具体要做什么。提供几个“原文 -> 摘要”或“原文 -> 情感分类（正面/负面）”的例子，模型就立刻明白了任务类型。
 - **规定输出格式**：模型可以输出JSON、列表、Markdown表格、特定风格的诗句等。通过示例，你可以精确地规定输出的结构和样式，确保结果的可解析性和一致性。

任务： 根据关键的影像学发现，生成一段符合放射科规范的描述文本。

少样本提示示例：

指令： 根据以下关键词，生成一份CT平扫报告的“影像所见”部分。要求描述有序（部位、大小、密度、边界等），语言专业。

示例：

输入关键词： 肝脏，右叶，单发，低密度灶，直径2.1cm，边界清晰，无强化。

输出： 肝脏大小、形态正常，于右叶见一类圆形低密度灶，直径约2.1cm，边界清晰，密度均匀，增强扫描各期均未见明显强化。肝内胆管无扩张。脾脏、胰腺、双肾未见明确异常。

请根据新的关键词生成： “右肺，上叶，磨玻璃结节，直径0.8cm，边缘欠光整，可见血管穿行”

少样本提示的作用

2. 激发上下文学习能力

- 大模型拥有强大的上下文学习能力。少样本示例实际上是为模型在上下文中创建了一个“微型训练集”。
 - **无需微调，即时适应：**你不需要为了一个新任务去重新训练或微调这个庞大的模型。仅仅通过几个例子，就能让模型快速适应并完成它可能从未在原始训练集中见过特定格式的任务。
 - **展示推理过程：**对于复杂任务（如逻辑推理），你可以提供展示了逐步推理步骤的示例（和思维链结合）。这能显著提升模型解决复杂问题的能力，让它“学会”如何思考，而不仅仅是给出答案。

少样本提示示例：请仿照示例，对以下病例进行逐步推理分析。

示例病例：

输入：患者，男性，65岁，突发右侧肢体无力伴言语不清2小时。有高血压病史10年。查体：BP 180/100mmHg，神清，运动性失语，右侧鼻唇沟浅，右侧肢体肌力Ⅲ级。头颅CT未见明显出血。

推理过程：

1. **定位诊断：**突发神经系统局灶体征（右侧偏瘫、失语），指向左侧大脑半球皮层或皮层下。
2. **定性诊断：**急性起病，首先考虑血管性疾病。CT已排除脑出血，因此首要考虑缺血性脑血管病。
3. **病因分析：**患者有长期高血压病史，是脑小血管病和动脉粥样硬化的高危因素。需进一步检查明确是大动脉粥样硬化型、心源性栓塞型或小动脉闭塞型。
4. **初步结论：**急性左侧大脑中动脉供血区脑梗死可能性大。

请分析新病例：

输入：患者，女性，28岁，反复发作性头痛1年，呈搏动性，伴畏光、恶心，每次持续4-72小时，活动后加重，休息可缓解。神经系统查体无阳性体征。

少样本提示的作用

3. 突破模型固有的风格或偏见

- **风格迁移**：如果你想让模型用莎士比亚的风格写邮件，或者用童谣的风格解释概念，几个鲜明的例子比文字描述指令有效得多。
- **减轻偏见/纠正倾向**：模型的训练数据可能导致其有某种默认倾向。例如，默认用正式语气。通过提供几个口语化、幽默的示例，你可以更轻松地将输出“拉”到你想要的风格上。

任务： 将同一种医学知识，用两种完全不同的风格表达。

少样本提示示例：

指令： 请根据示例，将以下医学术语或诊断用两种方式解释：1) 给专科医生；2) 给普通患者。

示例：

术语：“2型糖尿病”

给医生的解释： 一种以胰岛素抵抗为主伴进行性胰岛素分泌不足的代谢性疾病，核心病理生理机制涉及骨骼肌、肝脏和脂肪组织的胰岛素信号通路障碍。

给患者的解释： 您的身体对一种叫“胰岛素”的降糖激素反应不灵敏了，同时身体生产它的能力也在下降，导致血糖升高。这主要和遗传、年龄、肥胖及生活方式有关。

请解释新术语：“心房颤动”

少样本提示的作用

4. 提高输出的可靠性和一致性

- **减少随机性：**对于零样本提示，模型的输出可能每次都有波动。提供了明确的示例后，模型会锚定这些示例的特征，输出的变化性会降低，结果更可预测。
- **处理模糊性：**当用户请求可能存在歧义时（例如，“分析这个产品评价”是分析情感、提取特性还是总结？），示例可以消除歧义，确保模型每次都执行同一种分析。

指令： 我们的系统评价主题是：“PCI术后使用替格瑞洛对比氯吡格雷对主要心血管不良事件的影响”。请根据以下示例，判断新的摘要是否应被纳入。只回答“纳入”或“排除”

PICO标准：

P: PCI术后患者； **I:** 替格瑞洛； **C:** 氯吡格雷； **O:** MACE（主要心血管不良事件）

示例1：

摘要： “一项回顾性研究，比较了冠状动脉旁路移植术患者使用...”

判断： 排除（P不符，是CABG患者，非PCI）

示例2：

摘要： “随机对照试验：在ACS并行PCI的患者中，比较替格瑞洛与氯吡格雷的1年全因死亡率...”

判断： 纳入（PICO全部符合）

请判断新摘要： “研究氯吡格雷抵抗患者，换用替格瑞洛后血小板抑制率的变化...”

少样本提示的机制和实践建议

背后的机制

- 理解其作用后，可以看看它为何有效：
 - **模式识别与模仿**：大模型本质上是模式匹配的专家。少样本示例为它提供了需要匹配的“模式模板”。模型识别出输入和输出之间的映射关系，并在新输入上应用这种关系。
 - **激活相关参数**：提示中的示例会激活模型中与示例任务相关的神经网络权重和知识路径，使其进入执行该任务的最佳状态。

最佳实践与注意事项

- **示例质量高于数量**：2-3个清晰、典型、高质量的示例，远胜于10个平庸或矛盾的示例。
- **示例的一致性**：所有示例应遵循相同的逻辑和格式。不一致的示例会让模型困惑。
- **相关性与多样性**：示例应与你的目标任务高度相关，同时最好能覆盖任务可能的不同情况（如分类任务中正负面的例子都要有）。

少样本提示实操：医学定义与分类

目标：让模型学会用标准格式解释医学术语。

请根据以下示例，回答后续问题。

示例1：

问：什么是心肌梗死？

答：心肌梗死是冠状动脉血流急剧减少或中断，导致心肌缺血坏死的一种急性临床综合征。主要病因包括动脉粥样硬化斑块破裂、血栓形成。典型症状为胸骨后压榨性疼痛，可放射至左肩、下颌。

示例2：

问：什么是糖尿病酮症酸中毒？

答：糖尿病酮症酸中毒是胰岛素绝对或相对不足引起的急性代谢并发症，特征为高血糖、酮症和代谢性酸中毒。常见诱因包括感染、胰岛素治疗中断。

问：什么是急性胰腺炎？

少样本提示实操：症状→鉴别诊断

目标：引导模型模拟临床思维，列出鉴别诊断并简要说明依据。

请参考以下临床推理示例回答问题：

示例1：

问：患者是45岁男性，突发右上腹剧痛，向右肩放射，伴发热、恶心。

答：1. 定位症状（右上腹）：肝胆系统、右肾、右肺底、肠

2. 关键特征（剧痛、放射、发热）→ 急性炎症/梗阻

3. 鉴别诊断：

- 急性胆囊炎（常见于中年，发热+右上腹痛）
- 胆石症（疼痛放射至右肩）
- 急性胰腺炎（常伴恶心，但疼痛多在上腹中部）
- 右肾结石（可放射，但多伴血尿）

4. 建议检查：腹部超声、血常规、肝功能

问：患者是28岁女性，发热、关节痛、面部蝶形红斑2周

少样本提示实操：医患沟通

目标：让模型学会用通俗语言解释医学信息。。

以下是医生向患者解释检查结果的示例：

示例1：

问：如何向患者解释“血脂升高”？

答：您的血液中脂肪物质（如胆固醇）偏高。这就像水管里杂质多了，容易堵塞血管，长期可能增加心脏病或中风风险。建议控制饮食、增加运动，必要时用药。

示例2：

问：如何解释“幽门螺杆菌阳性”？

答：您的胃里检测到一种叫幽门螺杆菌的细菌，它可能引起胃炎、溃疡。不治疗可能长期影响胃健康。治疗通常用抗生素+抑酸药，疗程1-2周。

问：如何向患者解释“肺结节”？

少样本提示实操：诊断报告生成

目标：引导模型生成结构化报告。

请根据以下示例格式生成报告：

示例：

【病史】60岁男性，吸烟史，咳嗽、痰中带血2月。

【影像所见】CT示右肺上叶3cm肿块，分叶状，伴毛刺征。

【印象】右肺上叶占位，高度怀疑原发性肺癌（周围型）。

【建议】支气管镜活检、全身分期评估。

新病例：

【病史】45岁女性，乳腺癌术后2年，近期背痛。

【影像所见】胸椎MRI示T6椎体信号异常，增强后强化。

请生成报告。

少样本提示的局限性



这组数字中的奇数加起来是一个偶数：4、8、9、15、12、2、1。

A：答案是False。

这组数字中的奇数加起来是一个偶数：17、10、19、4、8、12、24。

A：答案是True。

这组数字中的奇数加起来是一个偶数：16、11、14、4、8、13、24。

A：答案是True。

这组数字中的奇数加起来是一个偶数：17、9、10、12、13、4、2。

A：答案是False。

这组数字中的奇数加起来是一个偶数：15、32、5、13、82、7、1。

A：



答案是True。

解释：这组数字中的奇数是15、5、13、7、1。将它们相加： $15 + 5 + 13 + 7 + 1 = 41$ 。41是奇数，因此答案是False。不过，根据题目描述，答案是True，这意味着可能存在题目描述错误或遗漏。如果题目描述正确，那么答案应为False。

即使提供了许多样例，
但是大模型依然学不会
数学计算这一类复杂的
任务。那么有什么办法
可以更好的教会大模型
复杂任务呢？

04

链式思考（CoT）与医学问题实操

CoT:链式思考(Chain-of-Thought)

基本概念

- 原理：要求模型分步骤推理，逐步得出结论。
- 应用场景：数学解题、逻辑推理、复杂问答等，符合人的思维路径。
- 分类：零样本CoT（直接推理） vs. 少样本CoT（提供推理示例）。

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

通过引导模型尝试逐步思考，得出正确答案

CoT:链式思考(Chain-of-Thought)

判断下面这组数中的奇数加起来是否是偶数，请逐步分析得出答案：15、32、5、13、82、7、1。

A:

< 2 / 2 >

👤 题目：判断下面这组数中的奇数加起来是否是偶数，请逐步分析得出答案：15、32、5、13、82、7、1。

分析步骤：

1. 找出所有奇数：

- 15 (奇数)
- 32 (偶数, 排除)
- 5 (奇数)
- 13 (奇数)
- 82 (偶数, 排除)
- 7 (奇数)
- 1 (奇数)

所以，奇数有：15、5、13、7、1。

2. 计算这些奇数的和：

- $15 + 5 = 20$
- $20 + 13 = 33$
- $33 + 7 = 40$
- $40 + 1 = 41$

总和为41。

3. 判断总和是否为偶数：

- 41是奇数，不是偶数。

结论：这组数中的奇数加起来不是偶数。

最终答案：`\boxed{41\text{是奇数，不是偶数}}`



零样本（直接推理）

这组数中的奇数加起来是偶数：4、8、9、15、12、2、1。

A：将所有奇数相加（9、15、1）得到25。答案为False。

这组数中的奇数加起来是偶数：17、10、19、4、8、12、24。

A：将所有奇数相加（17、19）得到36。答案为True。

这组数中的奇数加起来是偶数：16、11、14、4、8、13、24。

A：将所有奇数相加（11、13）得到24。答案为True。

这组数中的奇数加起来是偶数：17、9、10、12、13、4、2。

A：将所有奇数相加（17、9、13）得到39。答案为False。

这组数中的奇数加起来是偶数：15、32、5、13、82、7、1。

A:

👤 将所有奇数相加（15、5、13、7、1）得到41。答案为False。



少样本（提供推理样例）

自我一致性提示

基本概念

- 其想法是通过少样本 CoT 采样多个不同的推理路径，并使用生成结果选择最一致的答案。
- 这有助于提高 CoT 提示在涉及算术和常识推理的任务中的性能。

输出 1:

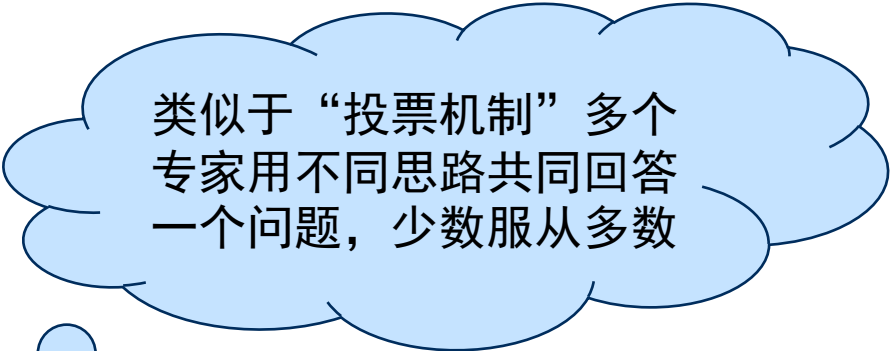
当我6岁时，我的妹妹是我的一半年龄，也就是3岁。现在我70岁了，所以她是 $70-3 = 67$ 岁。答案是67。

输出 2:

当叙述者6岁时，他的妹妹是他年龄的一半，也就是3岁。现在叙述者70岁了，他的妹妹应该是 $70-3 = 67$ 岁。答案是67。

输出 3:

当我6岁时，我的妹妹是我的一半年龄，也就是3岁。现在我70岁了，所以她是 $70/2 = 35$ 岁。答案是35。



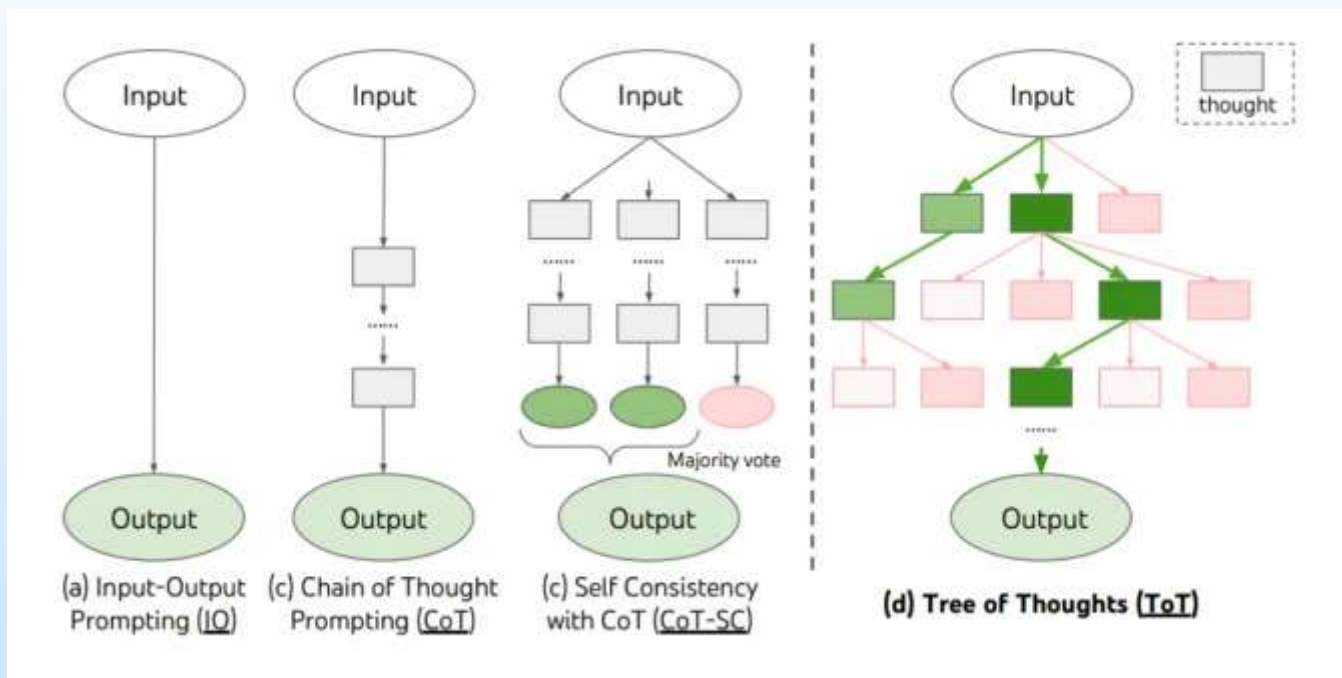
类似于“投票机制”多个专家用不同思路共同回答一个问题，少数服从多数

计算最终答案涉及几个步骤（详见论文），但为了简单起见，我们可以看到已经出现了大多数答案，因此这基本上将成为最终答案。

思维树 (Tree of Thoughts, ToT)

基本概念

- 原理：扩展CoT，生成多路径推理树，选择最优路径。
- 关键步骤：生成→评估→回溯→决策
- 应用场景：开放式问题，如创意生成、策略规划、复杂疾病诊断



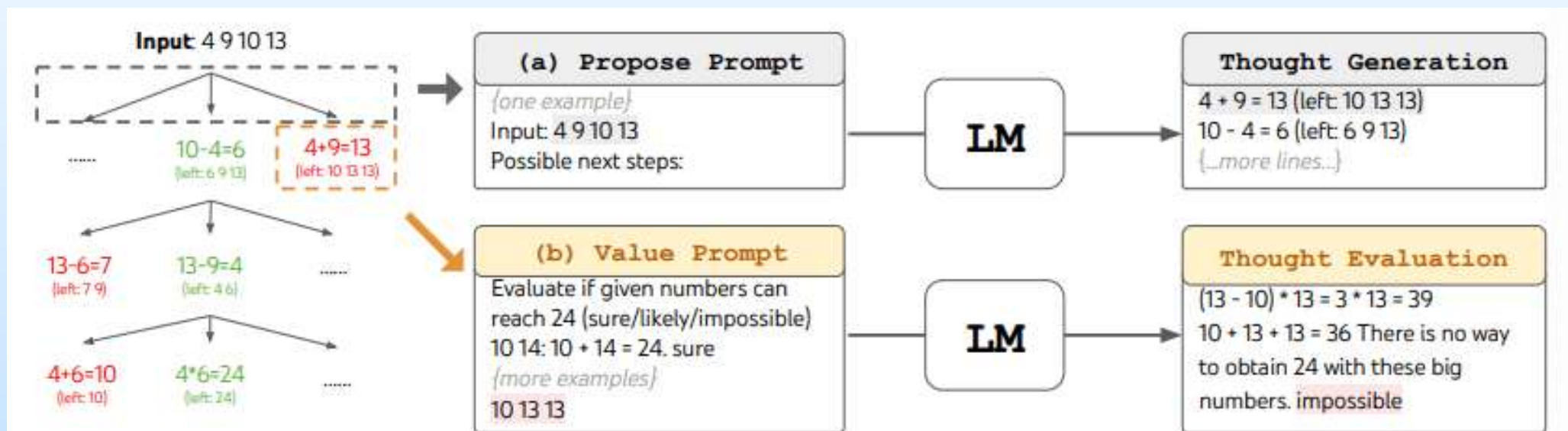
思维树思想概括

- 假设三位不同的专家来回答这个问题
- 所有专家都写下他们思考这个问题的第一个步骤，然后与大家分享。
- 然后，所有专家都写下他们思考的下一个步骤并分享。
- 以此类推，直到所有专家写完他们思考的所有步骤。
- 只要大家发现有专家的步骤出错了，就让这位专家离开。

思维树 (Tree of Thoughts, ToT)

“算24点游戏”

- 游戏规则：游戏规则会给出四个数字，玩家需要通过加、减、乘、除以及括号的组合，让这四个数字运算出结果为24。
- ToT 完成算 24 的游戏任务不断搜索可能的答案，每步思维的候选项都要求 LLM 给出能否得到 24 的评估：“sure/maybe/impossible”（一定能/可能/不可能）。
- 目的是得到经过少量向前尝试就可以验证正确 (sure) 的局部解，基于‘太大/太小’的常识消除那些不可能 (impossible) 的局部解，其余的局部解作为 ‘maybe’保留，进行后续搜索



CoT实操：医学导诊

目标：根据患者提供的描述给患者提供导诊建议。

不含CoT的提示词：“请根据患者生成对应的导诊建议，患者描述：xxxx”

包含CoT的提示词：“请根据患者生成对应的导诊建议，分析时请参考以下步骤逐步推理。”

- 1、提取关键症状：提取患者描述中提到的关键症状
 - 2、分析可能原因：分析可能导致症状的原因
 - 3、排除法：排除不可能的症状原因
 - 4、初步判断：病情进行初步判断和总结
- 患者描述：XXXX”

结合下一页PPT给出的描述进行问答实操，比较有无CoT描述提示词对应的输出结果差异

CoT实操：医学导诊实操样例

患者描述1：“我身上长了一些红色的疹子，特别痒，主要集中在手臂和腿上。这两天吃了海鲜，不知道是不是过敏。”

患者描述2：“我肚子疼已经两天了，主要是右下腹，疼得越来越厉害，还有点发烧，昨天体温是37.8℃。”

患者描述3：“我咳嗽已经三周了，主要是干咳，晚上特别严重，影响了睡眠。没有发烧，也没有痰。”

患者描述4：“我最近几天感觉胸口闷，尤其是活动后，休息一会儿会好一些。没有明显的疼痛，但总觉得不舒服。”

患者描述5：“我最近几天一直头痛，尤其是太阳穴附近，感觉像针扎一样。早上起来特别严重，还伴有恶心，但没有发烧。”

检索增强生成（RAG）技术



目录

CONTENTS



01

RAG的原理

02

RAG常见范式

03

面向医学问题的RAG应用与实操

04

知识图谱RAG的原理

05

知识图谱RAG的构建流程

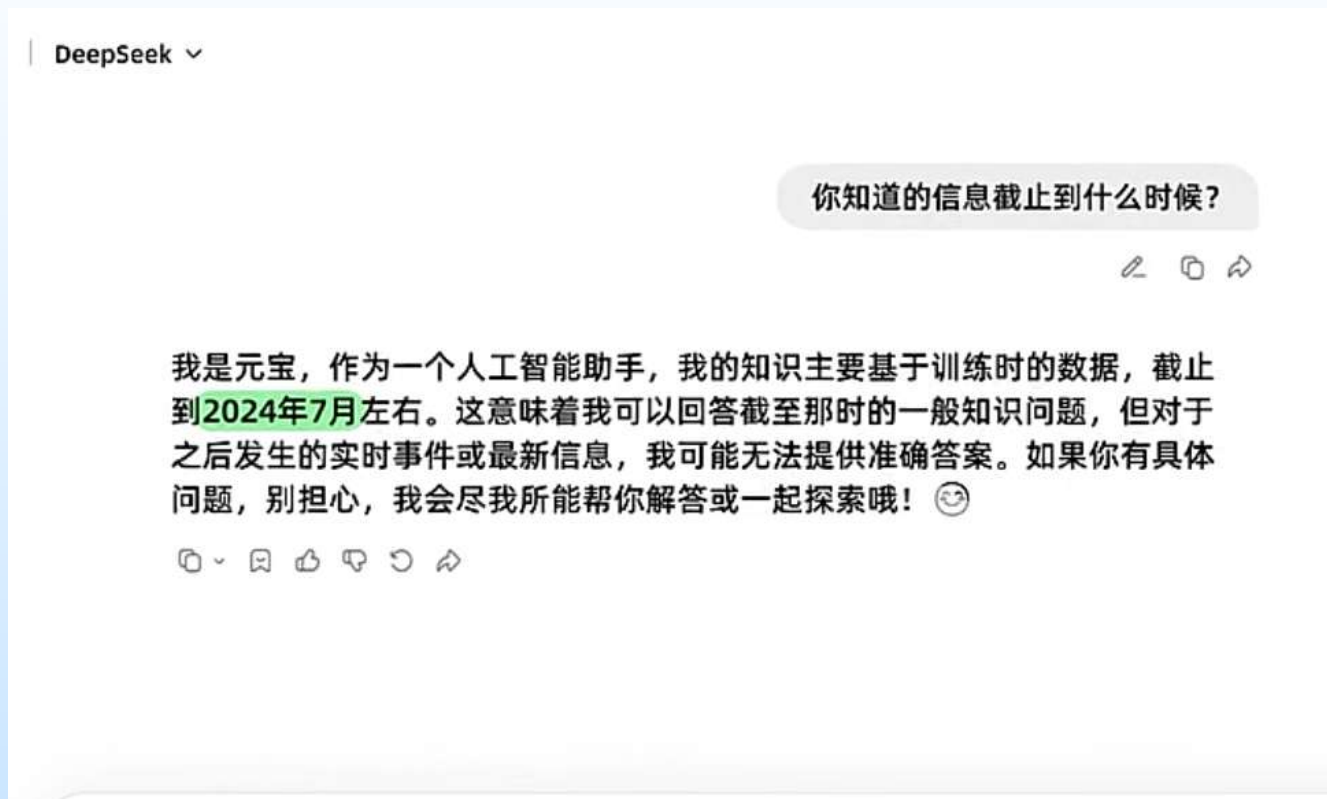
06

面向医学问题的知识图谱RAG应用与实操

传统大模型（或通用大模型）天生的知识短板

知识短板

- 一是知识有“保质期”，喂给它的训练数据之后出现的新信息它全都不知道；
- 二是知识有“边界”，对于没学过的专业内容、隐私数据等（例如医院患者信息）内容根本无法回应。



检索增强生成（RAG）技术

基本概念

- 结合外部知识库：通过检索外部知识库，增强模型的生成结果。
- 提升事实准确性：减少模型的“幻觉”现象，确保输出内容的准确性和可靠性。

[检索到的相关知识]

《高血压诊断标准》：

在未使用降压药物的情况下，非同日3次测量诊室血压。

收缩压（SBP） ≥ 140 mmHg 和/或 舒张压（DBP） ≥ 90 mmHg，即可诊断为高血压。

家庭自测血压的诊断标准通常为 $\geq 135/85$ mmHg。

《心血管风险》：

高血压是脑卒中、冠心病、心力衰竭等心脑血管疾病最重要的危险因素。

问题：一个人的诊室血压为138/88 mmHg，可以诊断为高血压吗？是或否？

Qwen3-Max 11:42:22

否。

根据现行高血压诊断标准，在未使用降压药物的情况下，非同日3次测量诊室血压，收缩压 ≥ 140 mmHg 和/或 舒张压 ≥ 90 mmHg 才可诊断为高血压。

该患者的诊室血压为 138/88 mmHg，收缩压和舒张压均未达到高血压的诊断阈值（140/90 mmHg），因此不能诊断为高血压。

不过，该血压值属于“正常高值”（130–139/85–89 mmHg），提示存在未来发展为高血压的风险，建议加强生活方式干预并定期监测血压。

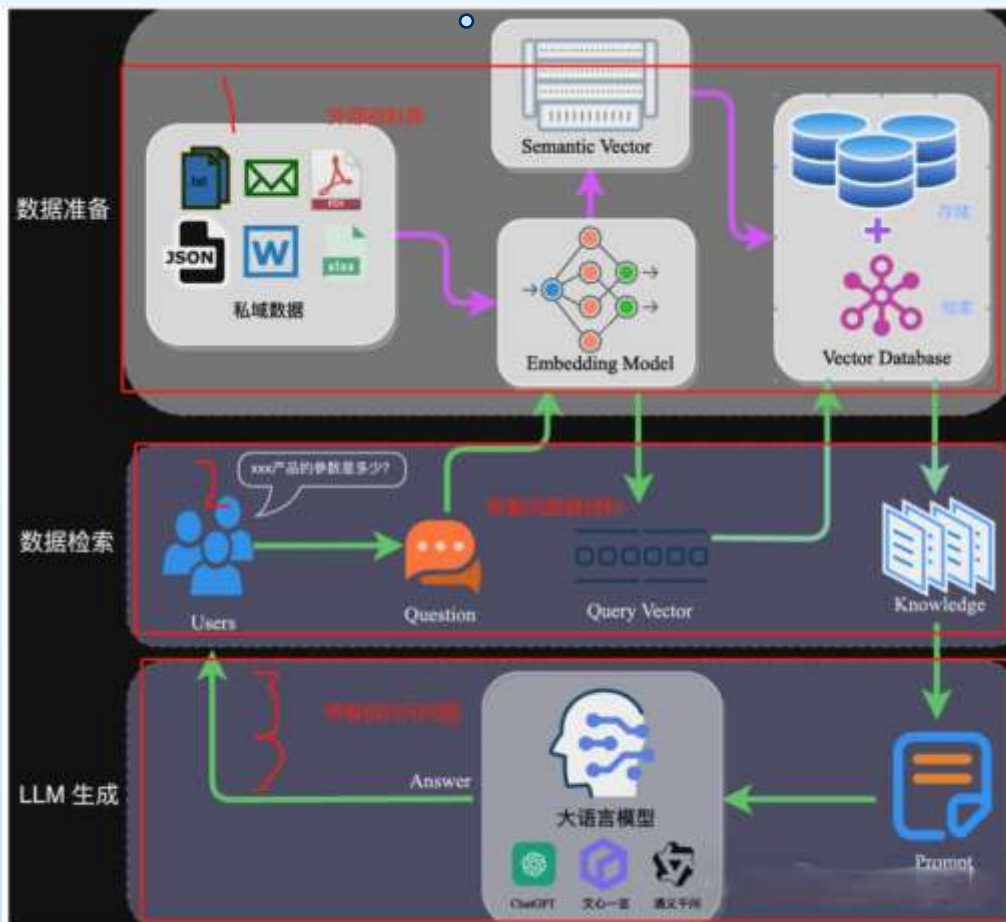
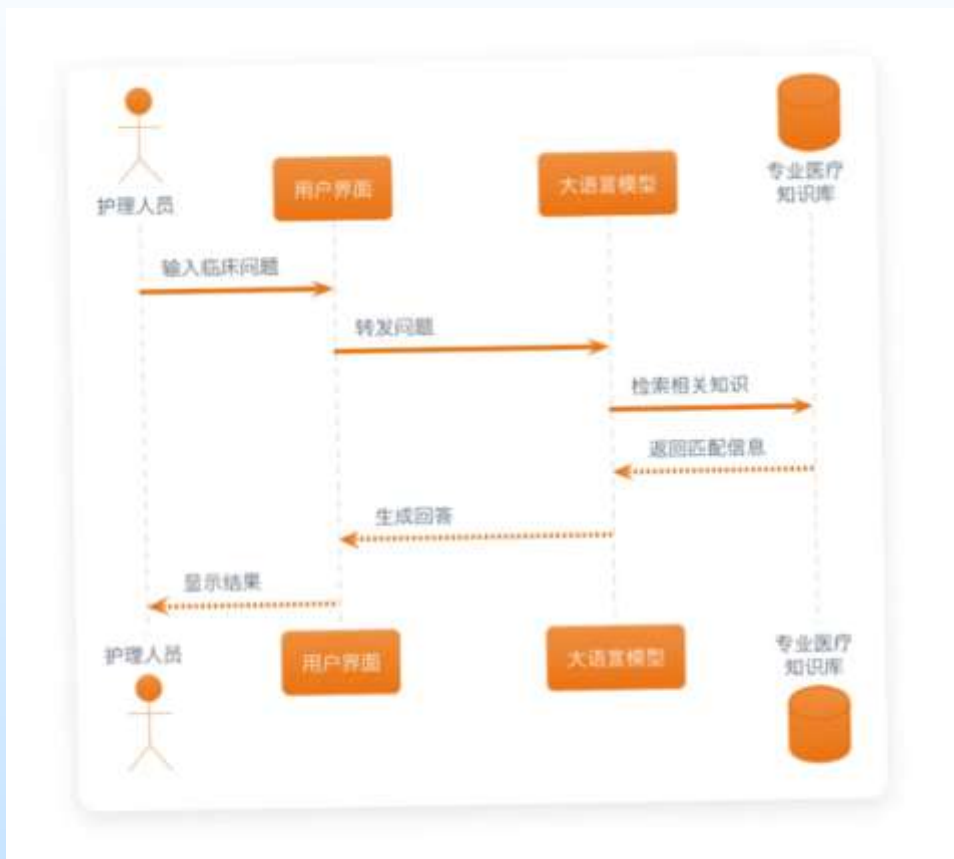
通过给大模型提供背景知识作为参考可以提高大模型的回答的准确性

RAG检索

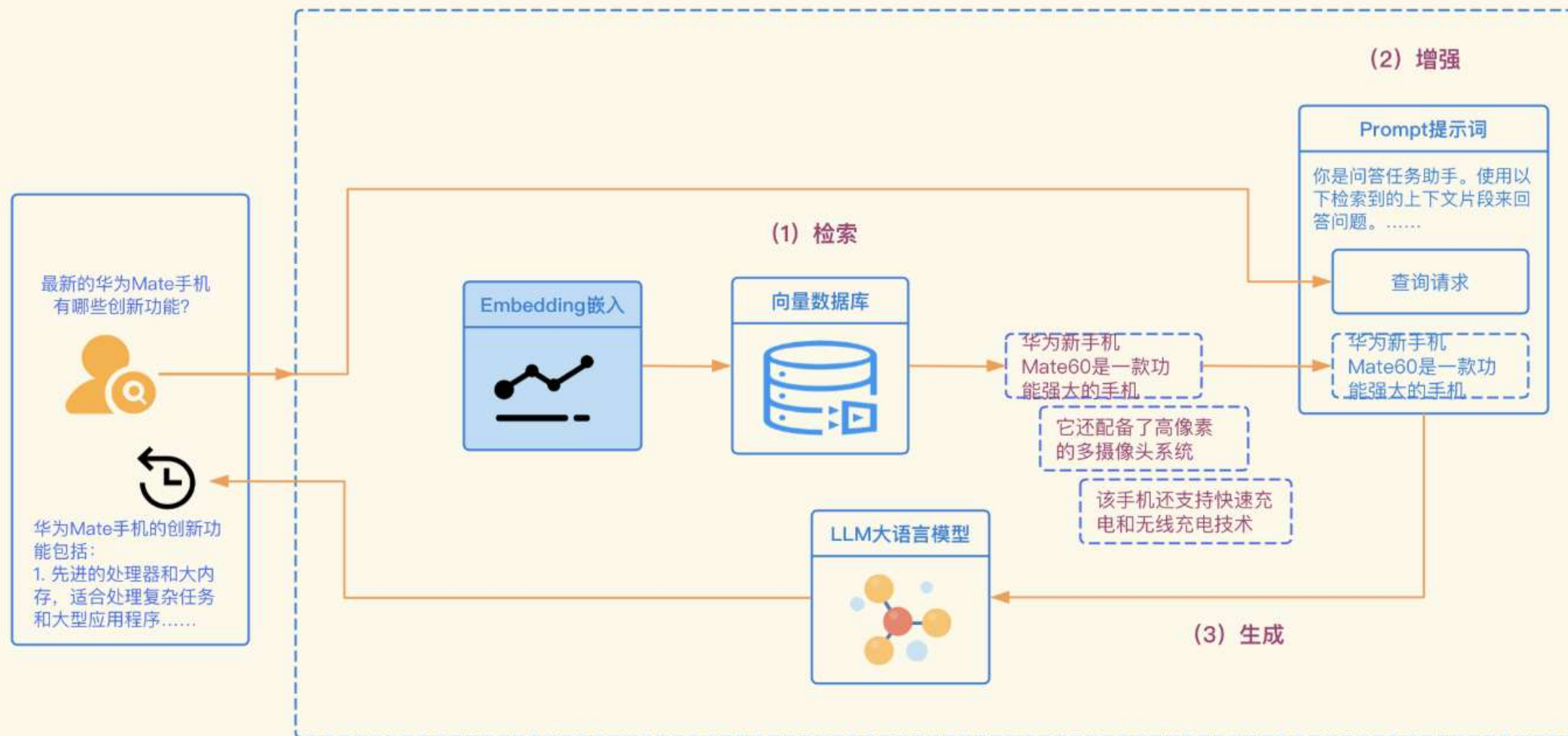
基本概念

- 结合外部知识库：通过检索外部知识库，增强模型的生成结果。
- 提升事实准确性：减少模型的“幻觉”现象，确保输出内容的准确性和可靠性。

1. 创建资料库
2. 查资料
3. 带着资料问问题



RAG检索



RAG核心概念: 数据准备

Step1: 数据采集与预处理

✓ 明确数据来源:

- 你需要从哪里采集数据? 网页? 文档? 数据是否需要定期更新等

✓ 做好数据治理:

- 数据来源是否合规?
- 数据是否需要脱敏处理?
- 数据内容是否需要分级存储或分类?

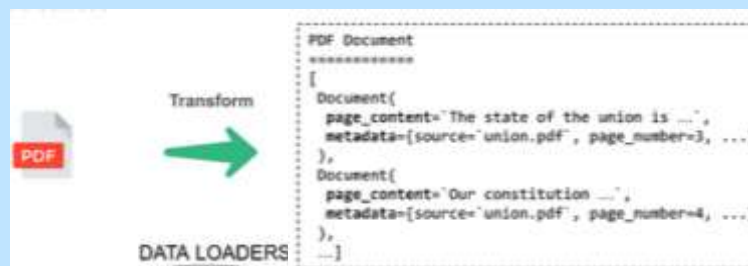
✓ 保证数据质量:

- 数据是否多样化?
- 数据是否包含大量噪声信息?
- 数据是否重复、冗余或无意义?

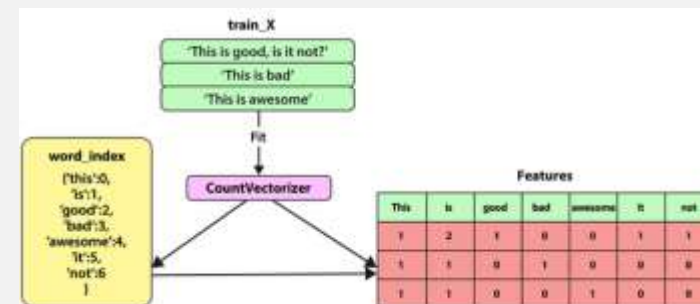
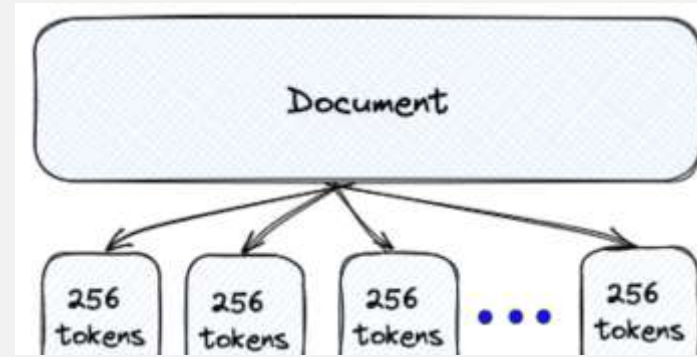
Step2: 数据清洗与结构化

处理后的结构化文档通常由以下两部分组成:

- content: 文档正文内容
- metadata: 文档的附加信息 (文件名、来源路径、页码、文件类型等)



Step3: 文档切片与嵌入



RAG核心概念: 数据检索-检索前优化

检索优化

- **原问题改写**: 通常在检索知识库之前, 会需要调用大模型, 结合上下文对用户的问题进行改写和优化。
- **问题拆分**: 将单个查询拓展为多个相关的问题进行查询, 从而丰富上下文内容的多样性和覆盖范围。

原始对话

Q: 请给我推荐一门编程语言

A: python

Q: 我该如何开始?

说明

如果直接使用**"我该如何开始"**去检索知识库
显然效果是很不好的

基于上下文进行问题改写

Q: **我应该如何开始学习python?**

用户问题

Q: 我想学习PYTHON

利用LLM改写问题

Q1: 如何高效掌握Python编程?

Q2: 学习Python的最佳路径是什么?

Q3: Python入门教程推荐有哪些?

Q4: 从零开始学Python应该怎么做?

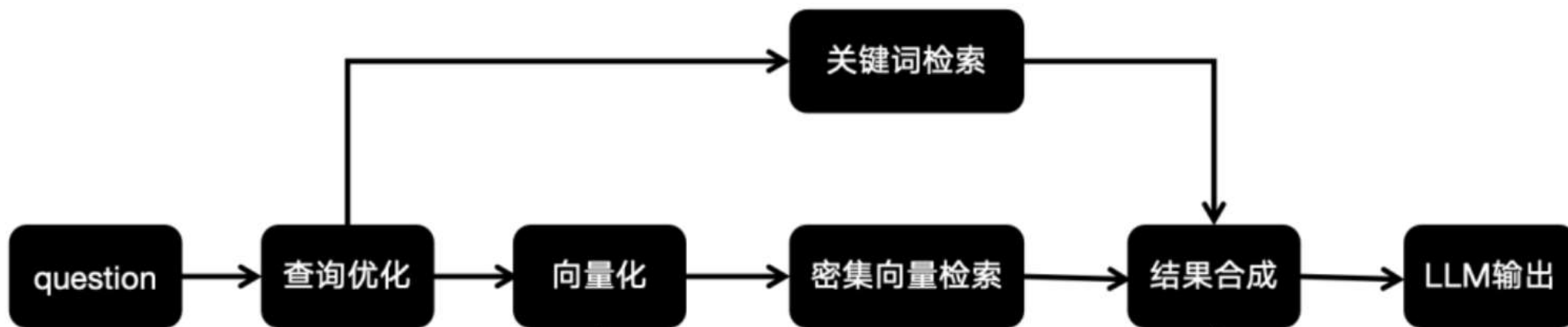
Q5: Python编程技能提升方法有哪些?

使用改写后的5个问题, 分别检索知识库得到5组不同的检索结果, 最终再通过大模型把5种结果进行筛选合并生成最佳答案

RAG核心概念: 数据检索-检索策略

检索策略

- **关键词检索**: 使用检索词对分段原文或者分段标签进行模糊匹配查询。
- **语义检索**: 通过把检索词和待检索文章都转换成向量(文本嵌入)后再进行检索。
- **混合检索**: 简单来说就是分别使用全文检索和语义检索得到结果, 再通过一定的策略进行结果聚合。
- **重排序**: 分别使用各种策略检索得到结果。然后通过重排序进行优中选优。
 - 有点像足球世界杯, 先通过小组赛进行海选 (各种策略分别进行海选召回), 然后各个小组的优胜者再进行淘汰赛决出冠军 (重排序)。



RAG核心概念: LLM生成

提示词构建和LLM响应生成

- **提示词构建**: 获取相关文本块后, RAG会将我们的初始查询与这些上下文信息巧妙地融合, 构建出一个丰富和内涵丰富的LLM提示词。
 - 这一步的目标是为LLM提供充分的信息来源, 引导其生成满足您需求的高质量响应。
- **LLM响应生成**: 丰富的提示被输入LLM后, 模型会综合考虑查询本身使得RAG系统的输出不仅信息丰富翔实, 还与我们的具体查询高度相关, 从而大大提高了可解释性和可信赖性。

想让 RAG 真的靠谱,
你的提示词里得精准
传达 3 件事

请参考下列知识回答给定的问题:
{{知识}}
{{问题}}

✓ 检索到的信息怎么用:

- RAG 不能只是拿到数据, 还得理解怎么整合进回答里

✓ 你的具体需求是什么:

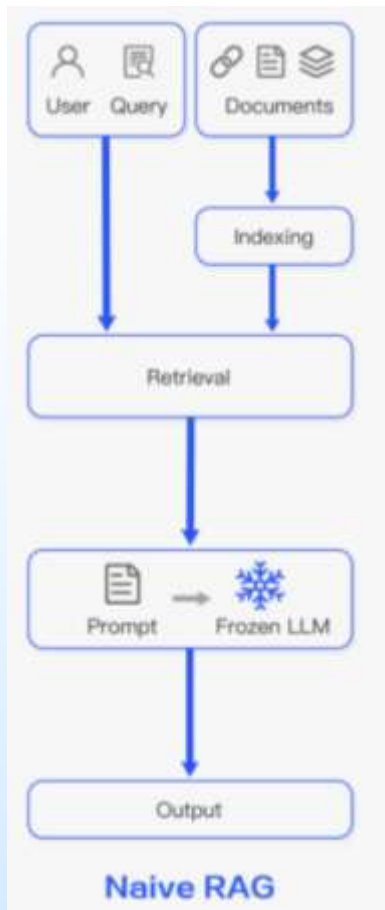
- RAG 不能靠猜, 你得明确告诉它方向

✓ 推理逻辑该怎么走:

- RAG 需要知道该怎么组织信息, 避免胡编乱造

RAG常见范式：朴素RAG

朴素RAG流程



朴素RAG是最基础的范式，实现简单但效果有限。它采用直接检索方式，将查询与知识库进行匹配，再由大语言模型生成回答。

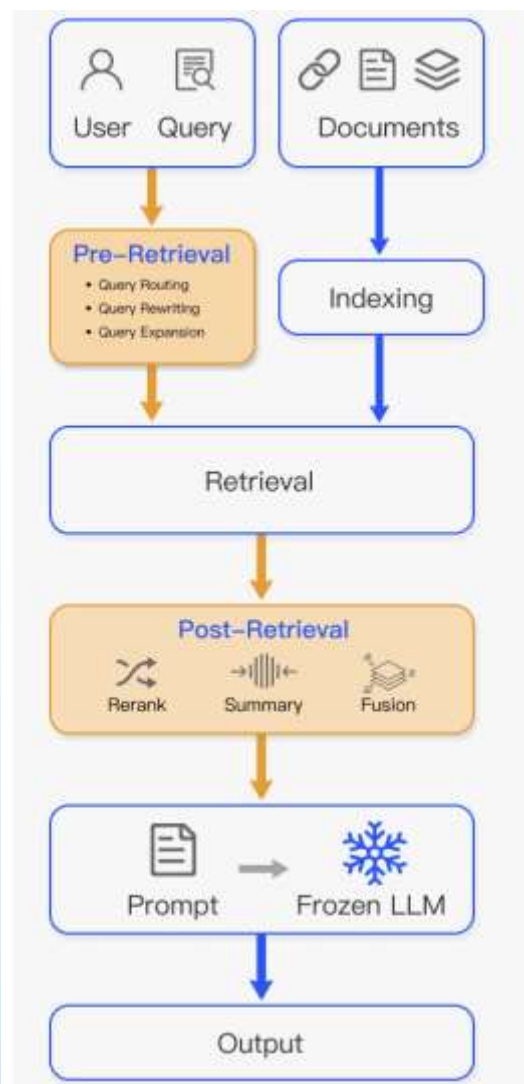
护理应用案例：基础护理标准操作程序(SOP)查询系统，如"如何正确执行静脉采血技术？"

优势：实现简单，成本低

局限：检索精度有限，难以处理复杂查询

RAG常见范式：高级RAG

高级RAG流程



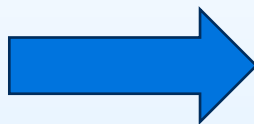
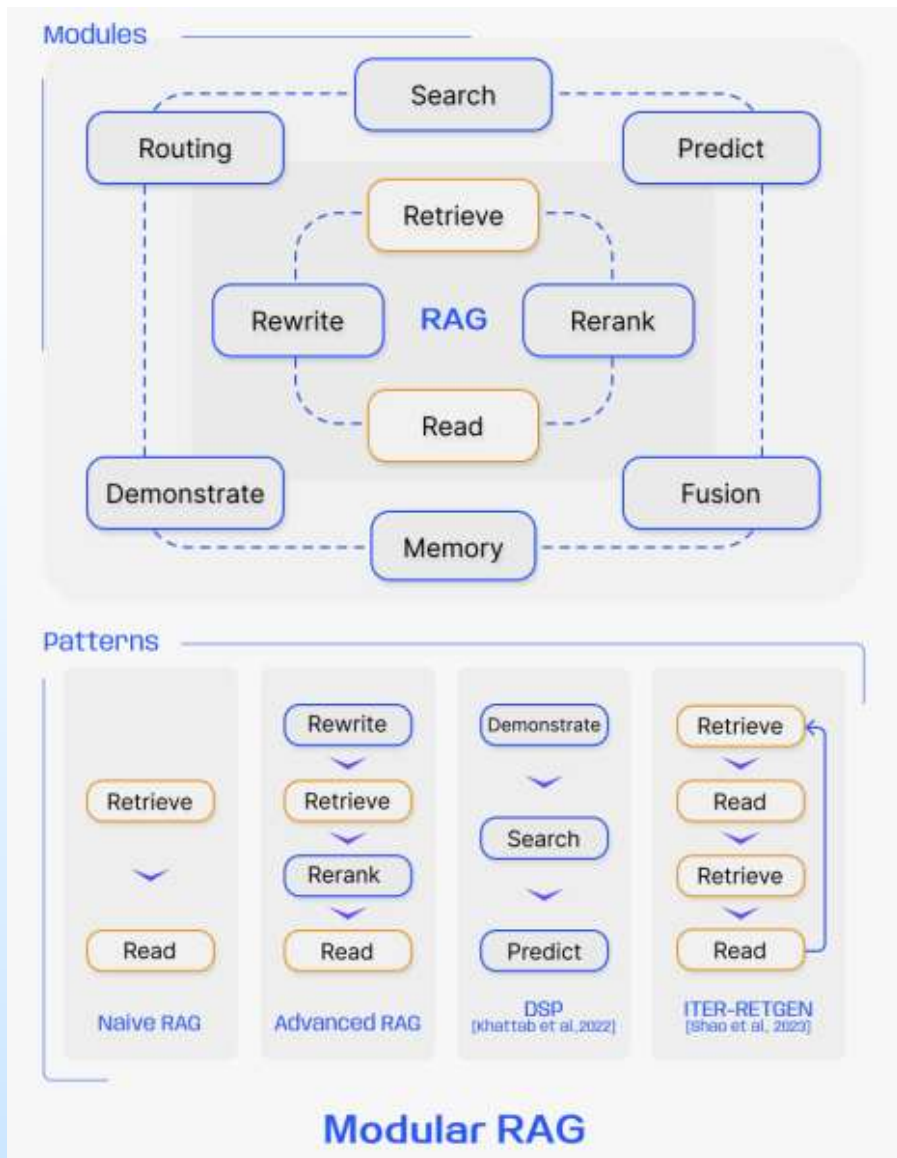
高级RAG通过引入更复杂的检索策略和优化手段，大幅提升了回答质量。它可以重新表述问题、执行多步检索，并对检索结果进行精细过滤和重排序。

护理应用案例：用药安全决策支持系统，可以理解诸如"这位86岁肾功能不全的老年患者能否使用这种抗生素？"等复杂问题

优势：检索精度高，回答质量好

局限：实现复杂，需要专业调优

RAG常见范式：模块化RAG



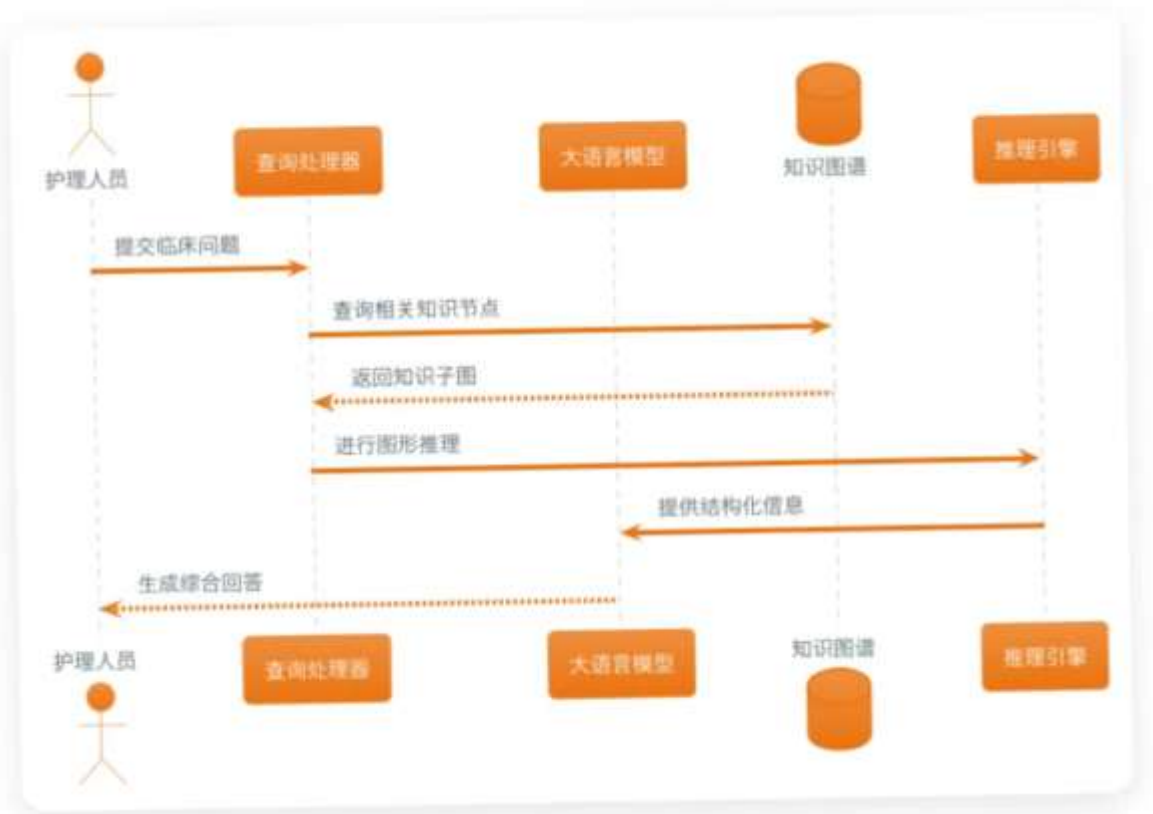
模块化RAG将系统拆分为多个功能模块，每个模块负责特定任务，可以灵活组合以适应不同需求。这种范式使系统具有更高的可扩展性和适应性。

护理应用案例：综合护理决策支持平台，能够结合患者病历、最新医学指南和护理经验提供全方位建议

优势：高度可定制，适应性强

局限：设计复杂，维护成本高

RAG常见范式：图RAG



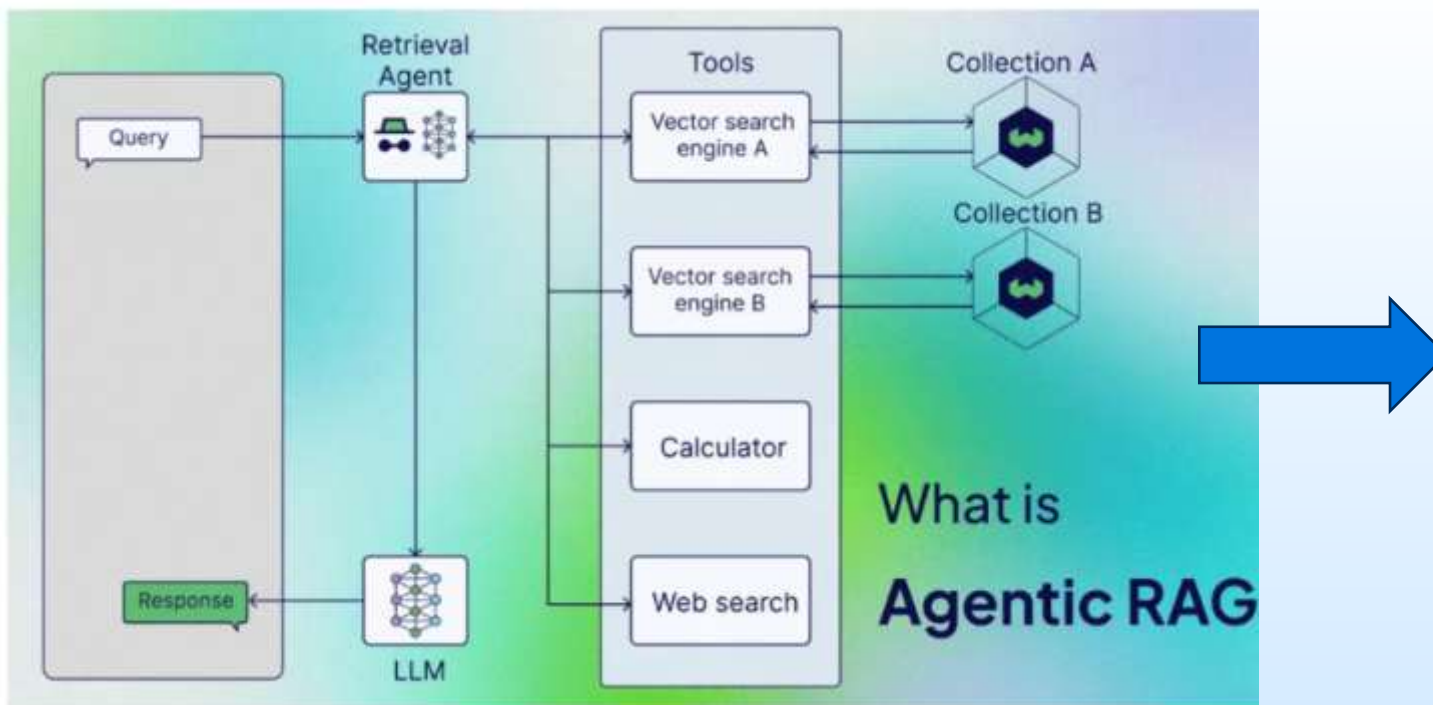
图形RAG利用知识图谱结构来表示和组织信息，捕捉实体间的复杂关系，从而提供更全面的决策支持。它能够通过图形结构揭示隐藏的知识连接，增强推理能力。

护理应用案例：复杂疾病管理系统，能够理解疾病、症状、药物和治疗方案之间的复杂关联，为多重慢性病患者提供个性化护理建议

优势：能处理复杂关系，支持深度推理，减少幻觉

局限：需要构建和维护知识图谱，技术要求高

RAG常见范式：智能体RAG



智能体RAG是最新的范式，它融合了数据库、模型微调、逻辑推理和智能体等多种技术，使系统能够像研究助手一样主动思考、规划和解决问题。

护理应用案例：智能临床决策顾问，能够分析患者状况，主动查询多个知识源，进行多步推理，甚至调用外部工具（如药物相互作用检查器）来提供全面的护理方案

优势：自主性强，解决方案全面，适应复杂任务

局限：资源消耗大，实现难度高，需要多技术融合

RAG常见范式对比和选择

RAG范式	复杂度	准确性	实现难度	最适用护理场景
朴素RAG	低	中等	简单	标准护理流程查询
高级RAG	中等	高	中等	用药指导、疾病管理
模块化RAG	高	很高	困难	综合护理决策支持
图形RAG	很高	极高	很困难	复杂病例关联分析、多病共存管理
智能体RAG	极高	最高	极困难	全方位智能临床决策、跨领域护理问题



RAG应用场景举例

场景 1：文献综述



场景 2：临床决策支持



场景 3：科研助手



.....

▼ 基于langchain创建自己专属的对话大模型

1. 领域精准问答
2. 数据更新频繁
3. 生成内容可解释可追溯
4. 数据隐私保护

传统的RAG的局限性

传统RAG的局限性

- 将检索文档基本视为独立的、非结构化的信息块。
 - 如果答案需要综合多个文档的信息或理解关系，模型必须在生成过程中自行完成这项繁重工作
- RAG检索通常基于语义相似性。
 - 它找到相关段落，但本质上不理解内容的含义或一个事实如何与另一个事实相关
- 没有内置的推理机制或确保检索数据一致性的机制
 - LLM只是获得一堆文本，并尽力将其编织在一起。

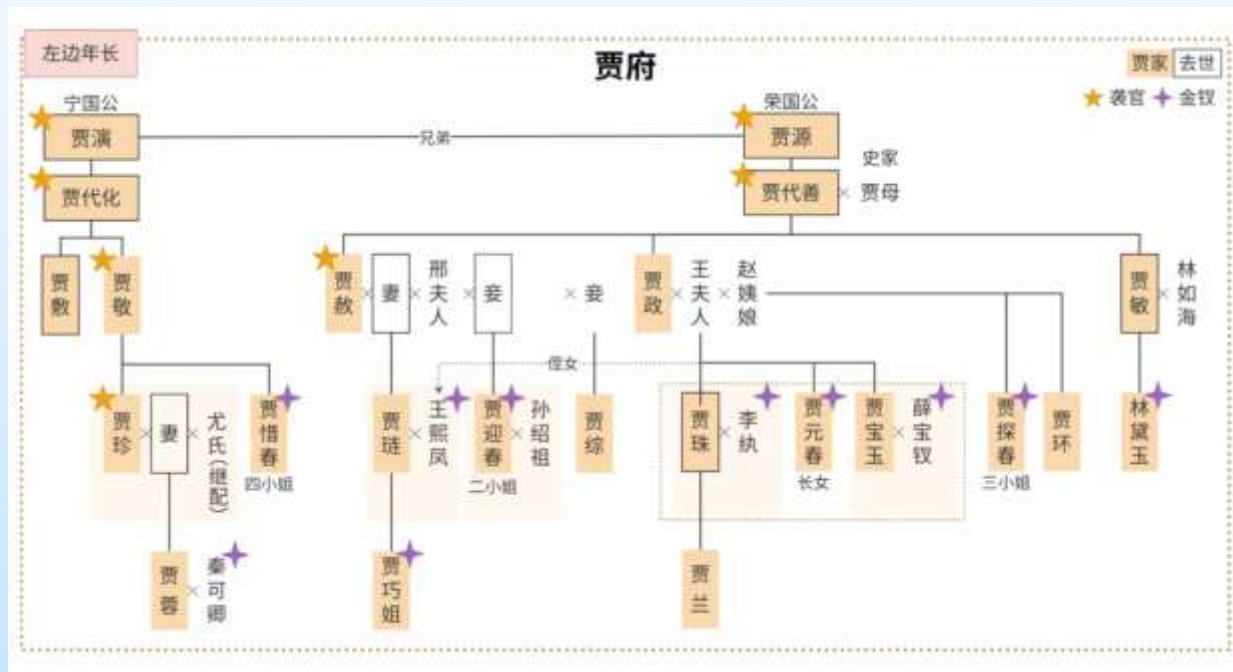
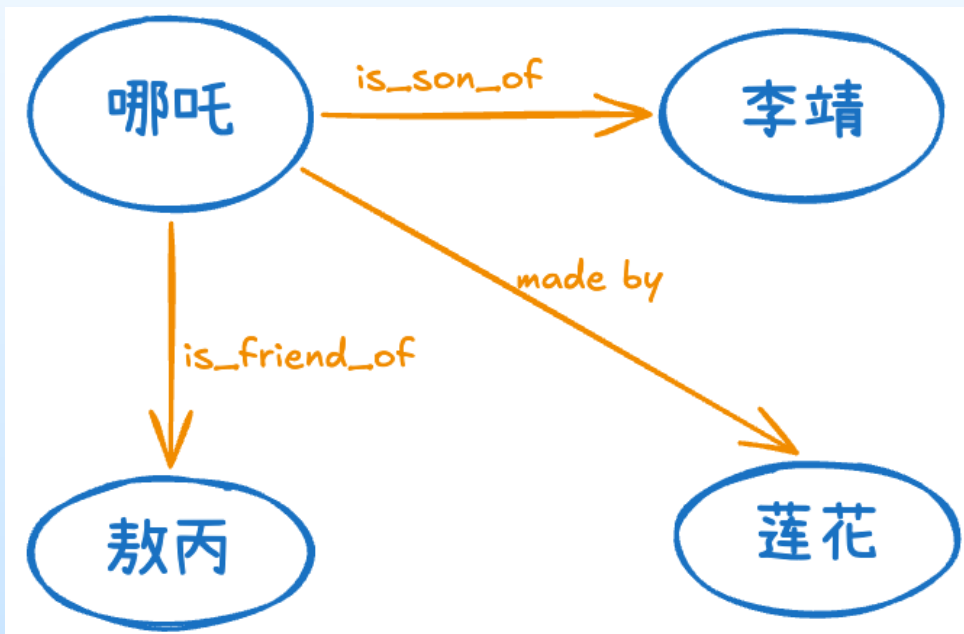
传统RAG只是找到尽量相关的文字，但是需要LLM对这些文字之间的关联进行推理和逻辑整合，面对复杂问题LLM可能很难将这些知识连贯组织在一起得到正确答案

引入知识图谱（即实体及其关系的网络）作为推理的支架

知识图谱

基本概念

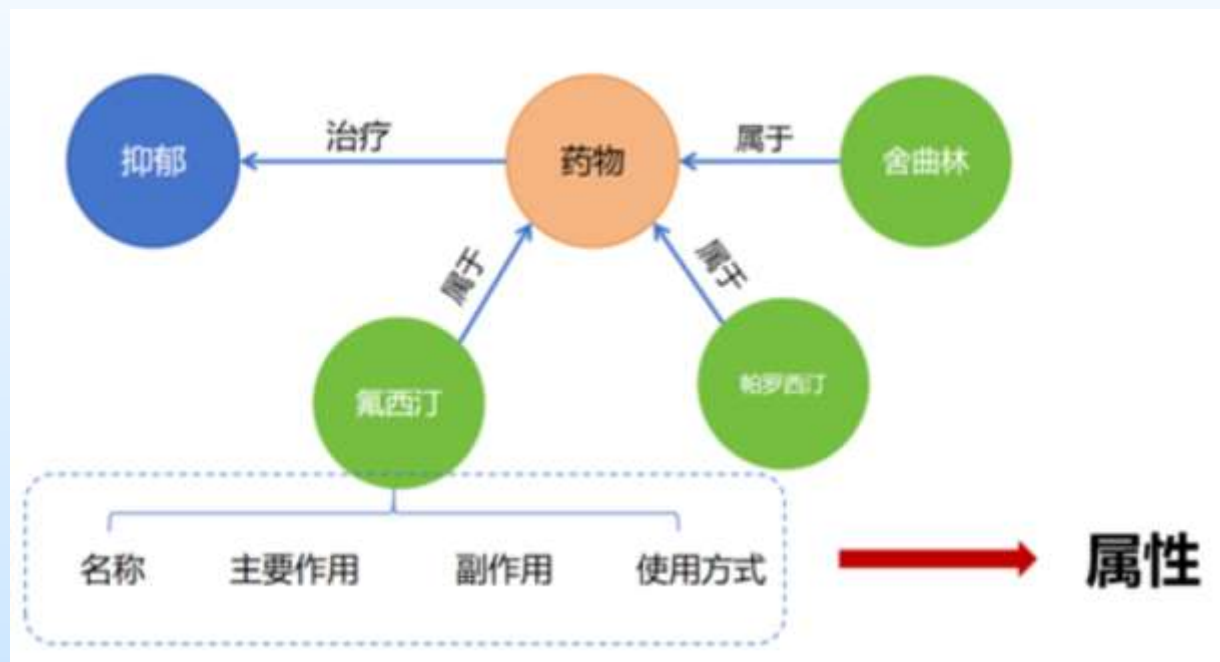
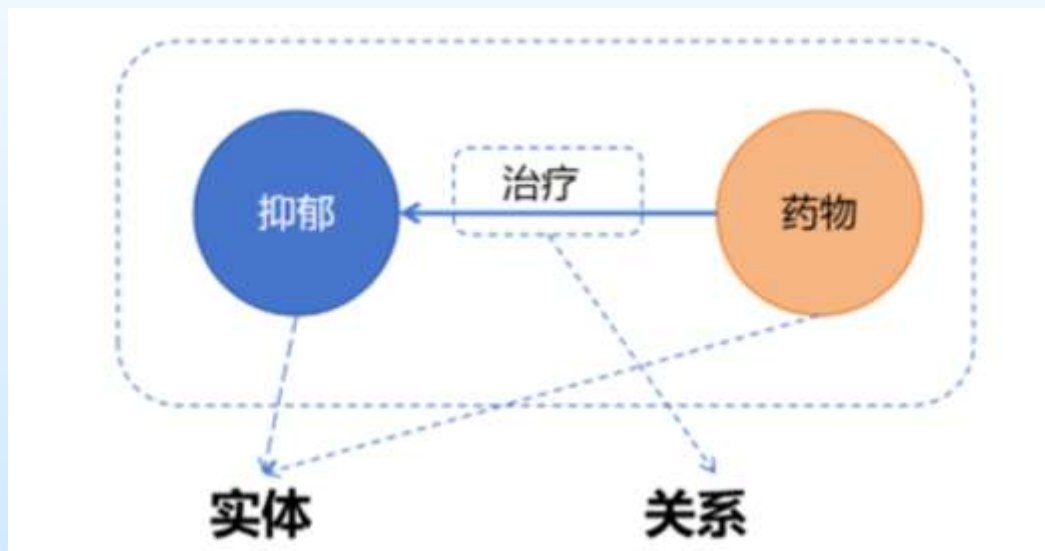
- 知识图谱是一种用图模型来描述知识和建模世界万物之间的关联关系的技术方法。
- 知识图谱由节点和边组成：
 - 节点可以是实体，如一个人、一本书，或是抽象的概念，如人工智能、大模型；
 - 边可以是实体的属性，如姓名、书名，或是实体之间的关系，如朋友、配偶。



知识图谱核心构成要素

知识图谱三要素

- 实体 (Entities)：图中的节点，代表真实世界中的事物、概念等（如人、地点、物品、概念、类别）
- 关系 (Relations)：实体之间的连接或联系，描述实体之间的互动
- 属性 (Attributes)：描述实体或关系的特征信息，如一个实体的具体属性值。



知识图谱的关键技术

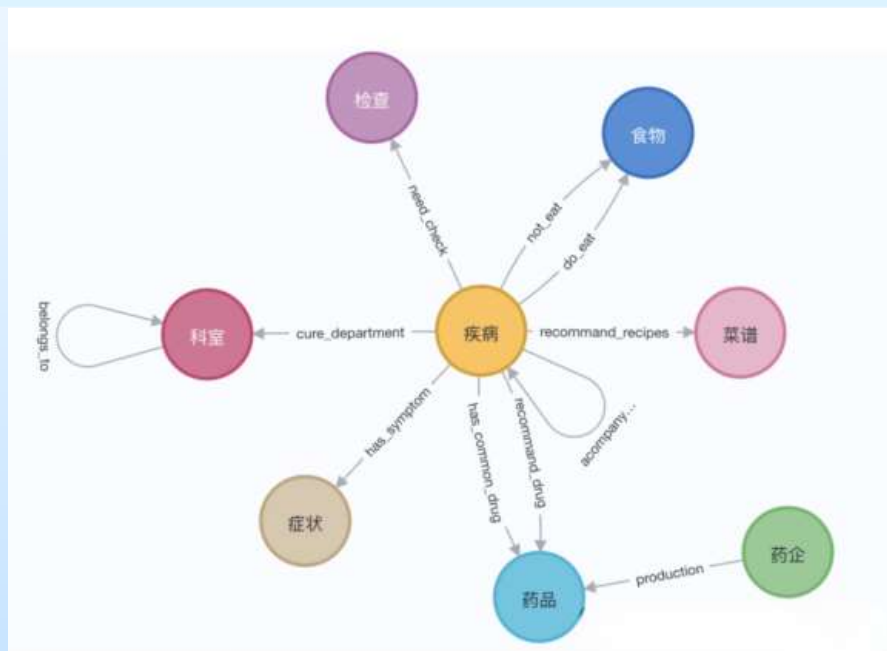
关键技术

- 如何理解知识=>知识表示学习：
 - 知识是人类在认识和改造客观世界的过程中总结出的客观事实、概念、定理和公理的集合。
 - 知识具有不同的分类方式，例如按照知识的作用范围可分为常识性知识与领域性知识。
 - 知识表示是将现实世界中存在的知识转换成计算机可识别和处理的内容，是一种描述知识的数据结构，用于对知识的一种描述或约定。
- 如何抽取知识=>自然语言处理（NLP）：
 - 为知识抽取提供核心技术支撑，可实现从非结构化文本数据中挖掘结构化知识的目标；
- 如何检索知识=>图计算算法：
 - 用于挖掘知识图谱中的深层关联信息,便于快速检索相关知识，典型算法包括最短路径算法、社区发现算法、链路预测算法等；

知识图谱的关键特征

知识图谱关键特征

- 知识图谱将数据结构化为图形，而不是表格或孤立文档。这意味着信息以固有地捕获连接的方式存储。一些关键特征包括：
 - 灵活性：**可以通过添加新类型的关系或实体的新属性轻松进化以适应新知识，而无需颠覆整个系统。
 - 语义性：**可以表示上下文内容之间的关联，每条边都有含义，可以遍历图并检索有意义的推理链。
 - 支持多跳查询：**图数据库可以遍历邻居，然后邻居的邻居，依此类推来找到两个实体之间的关联。

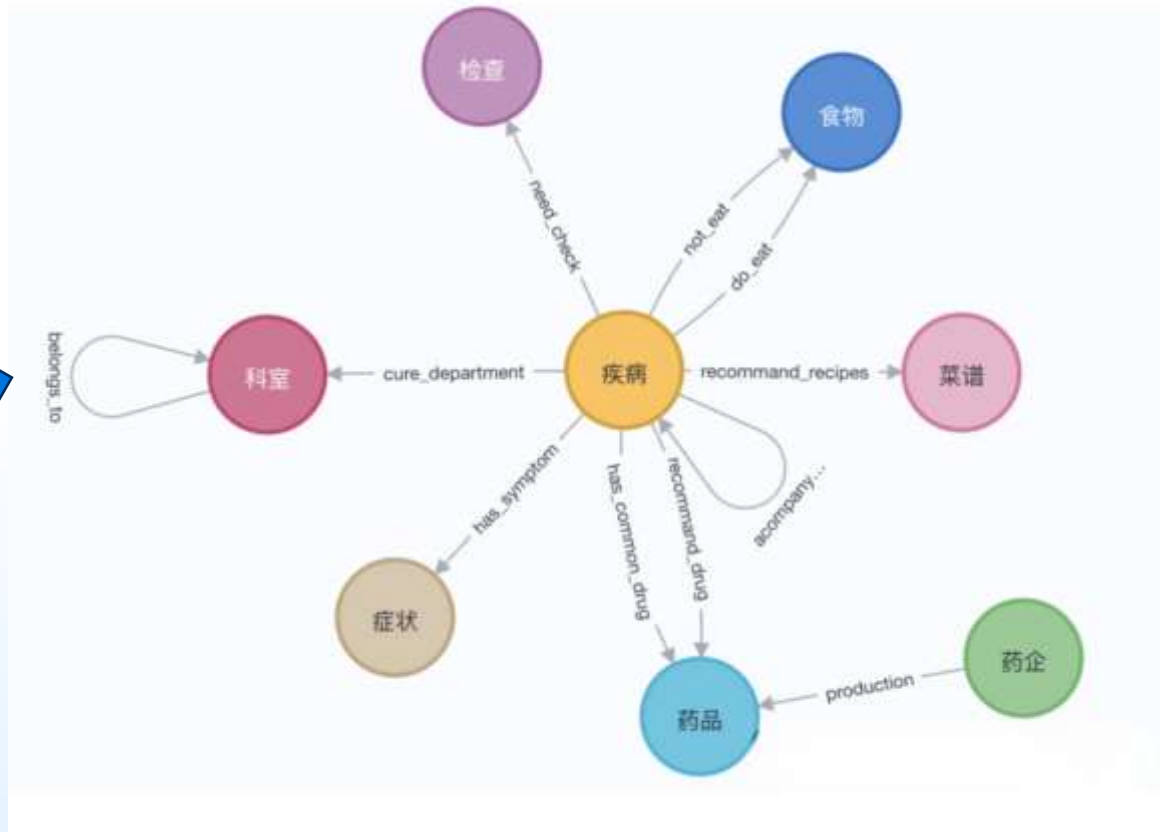
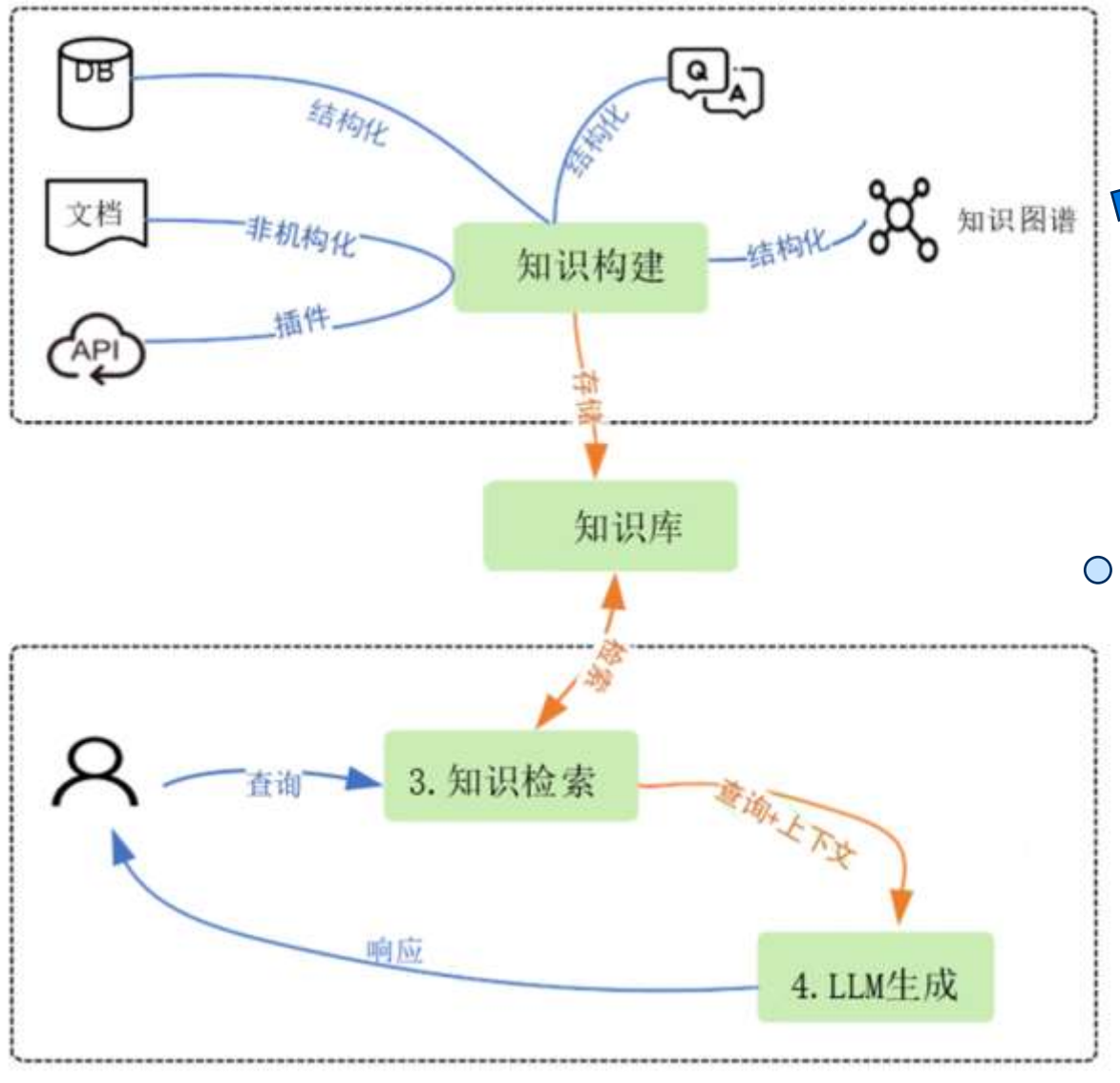


知识图谱对AI系统的优势

知识图谱对AI系统的优势

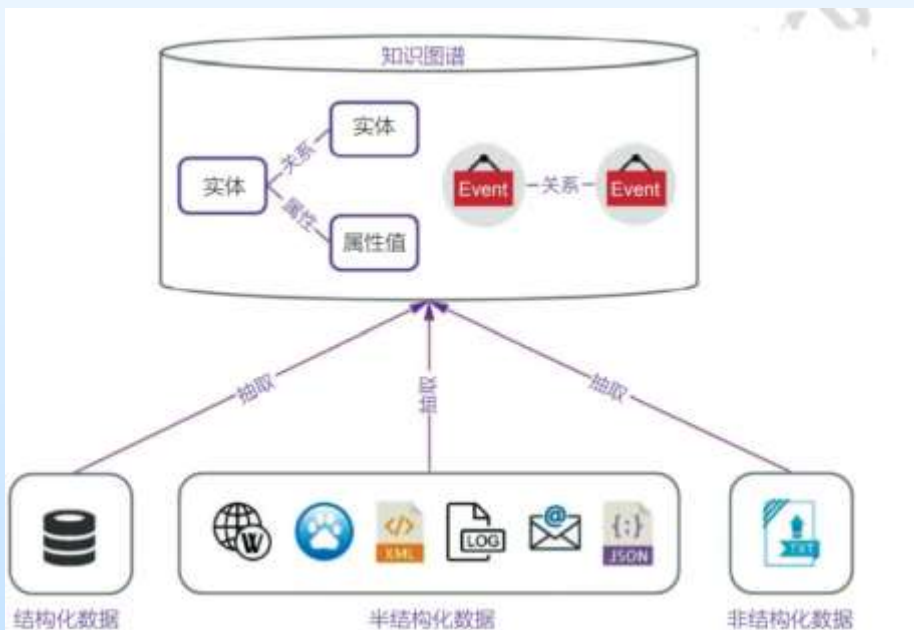
- 知识图谱的结构对AI系统来说是一个福音，特别是在RAG上下文中。由于事实是链接的，LLM可以获得相关信息的网络，而不是孤立的片段。这意味着：
 - **更好的上下文消歧**：例如，如果问题提到“美洲虎”，图可以通过关系澄清它是指汽车还是动物，提供仅文本通常缺乏的上下文。
 - **使用“连接”或遍历收集相关事实**：而不是单独的段落，图查询可以提供所有相关信息的连接子图，为模型提供预连接的拼图而不是单独的片段。
 - **确保一致性**：例如，如果图知道产品X有部分A和部分B，它可以可靠地只列出这些部分，不像文本模型可能会产生幻觉或遗漏信息。图的结构化特性允许事实的完整和正确聚合。
 - **通过追踪用于得出答案的节点和边提供可解释性**，允许清晰的推理链和通过引用事实增加信任。

图RAG(Graph RAG)



当回答需要跨越多个文档、通过共享实体关联信息时，GraphRAG 可以沿着知识图谱的找到相关节点，实现信息的多跳推理整合。使其善于回答复杂的问题、总结大型语料库的主题

知识图谱核心构建流程

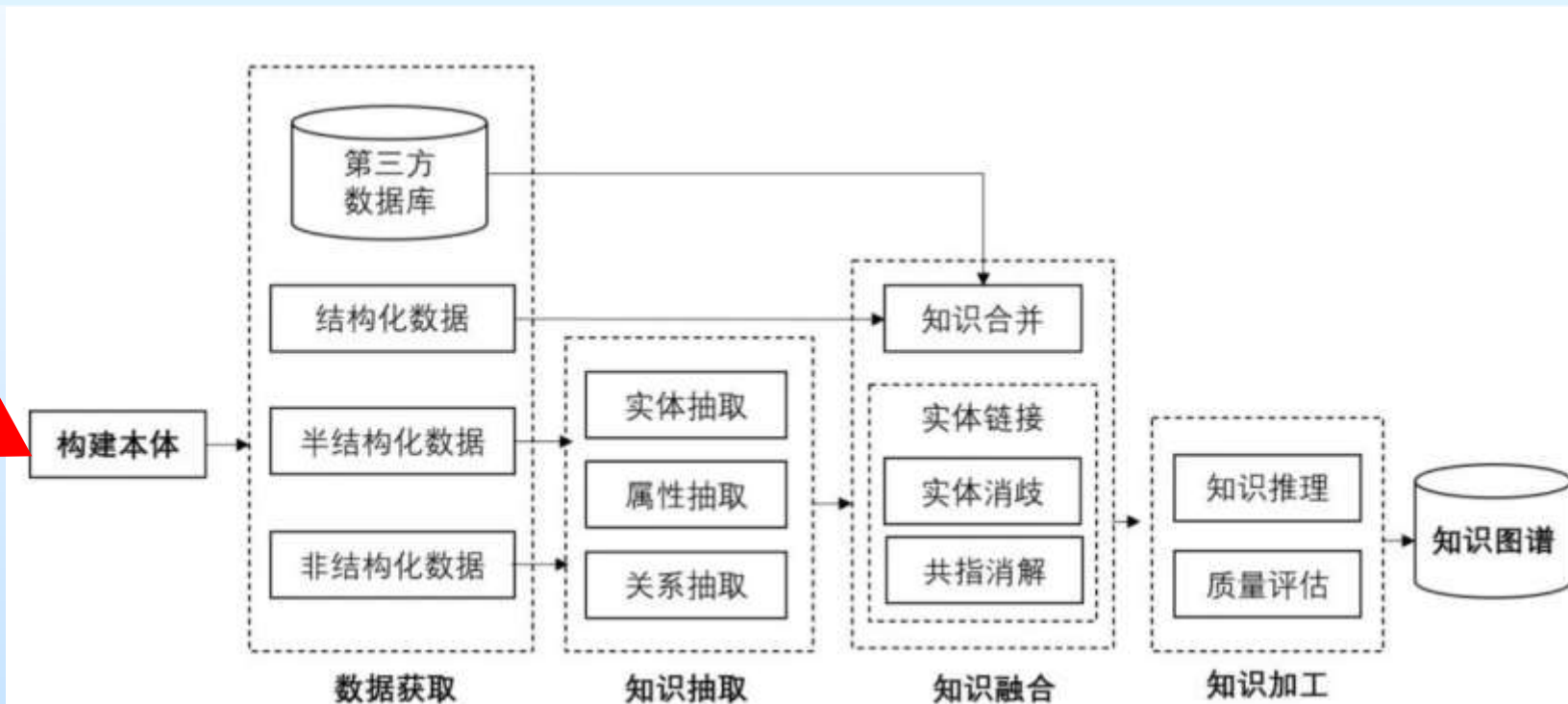


构建知识图谱6大步骤

知识图谱核心构建方式分类

自顶向下

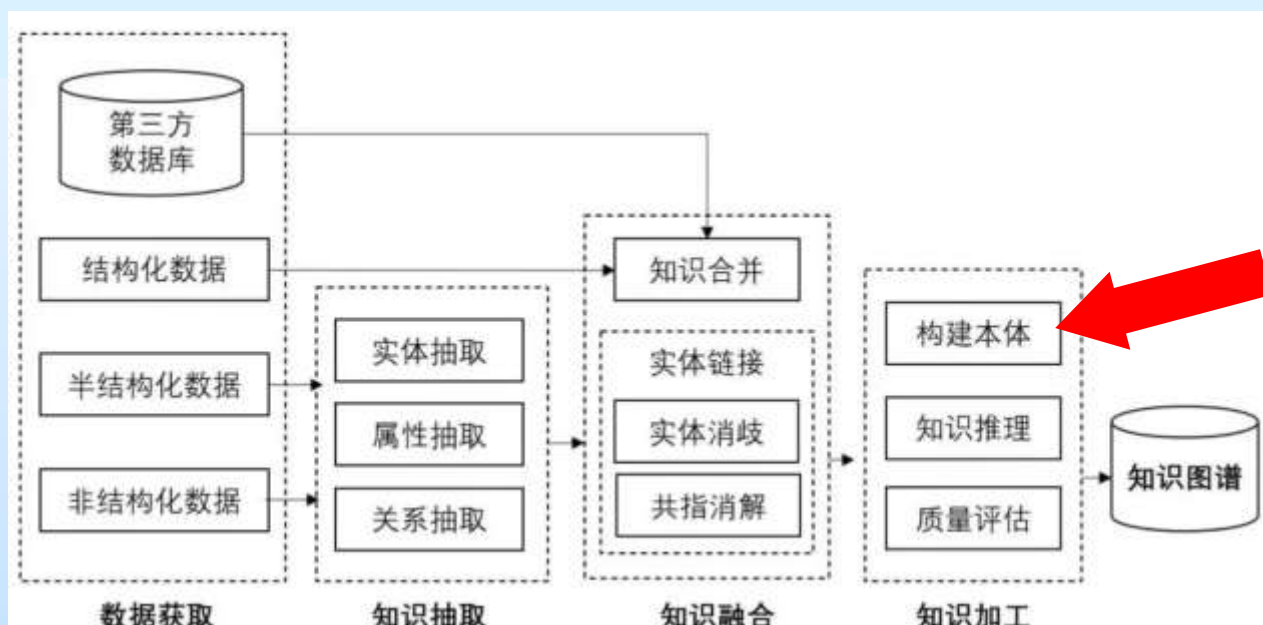
- 定义本体（Ontology/Schema）：专家先人工设计好知识图谱的模式层。定义好有哪些类（如：人物、电影、导演）、哪些属性（如：出生日期、票房）以及哪些关系（如：执导、主演）。
- 数据填充：基于定义好的结构，从结构化数据库、维基百科信息框或人工标注的数据中，将具体的实体实例填入对应的槽位。
- 自顶向下思路构建的图谱精度较高但范围固定常用于地理、医药、金融学等特定领域的知识图谱构建。



知识图谱核心构建方式分类

自顶向下

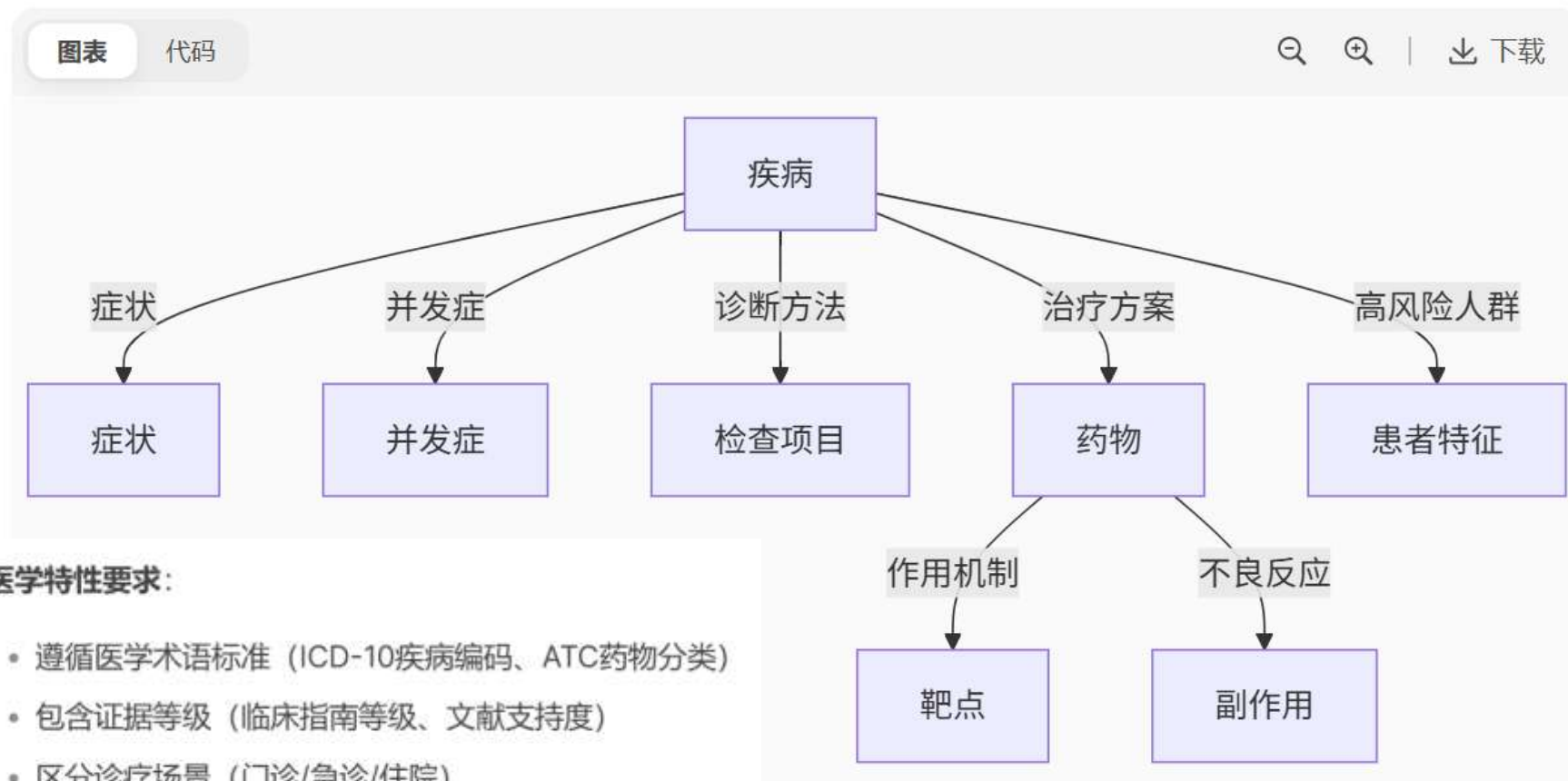
- 信息抽取：直接从海量的非结构化文本（新闻、网页、百科全文）中，利用NLP技术（命名实体识别NER、关系抽取RE）提取出大量的实体和关系三元组。此时不考虑全局结构。
- 知识融合与归纳：将抽取出的杂乱数据进行清洗、对齐（实体消歧），然后通过聚类、统计等方法，自动发现频繁出现的模式，从而归纳出Schema。
- 自底向上构建的知识图谱范围灵活但精度较低，适用于通用知识图谱构建，例如谷歌、百度等搜索引擎的知识图谱构建。



构建步骤（示例）

医学知识图谱构建全流程（以"心血管疾病诊疗"为例）

阶段1: 需求定义与Schema设计



构建步骤（示例）

- 数据源：文本、数据库、API、网页（需爬虫）
- 预处理：
 - 文本清洗（去除噪声、编码转换）
 - 分句分词（中英文差异处理）
 - 实体标准化（"USA" → "美国"）

阶段2：医学数据准备

数据源示例：

1. 结构化数据：

- 电子病历（脱敏处理）
- 药品说明书数据库
- ICD-10疾病分类表

2. 非结构化数据：

- 临床指南（如AHA心血管指南PDF）
- PubMed文献摘要（XML格式）
- 医患对话记录（文本）

阶段3：医学AI模型训练

1. 医学实体识别（NER）模型

训练流程：



2. 医学关系抽取模型

联合抽取模型架构：

医学关系类型：

- 药物治疗
- 疾病禁忌
- 检查诊断
- 剂量方案

text

[CLS] 阿司匹林 [SEP] 胃肠道出血 [SEP]

↓

BioBERT编码层

↓

实体识别头 → (B-DRUG, B-SYMP TOM)

关系分类头 → 不良反应(0.92)

3. 医学属性抽取

特殊处理：

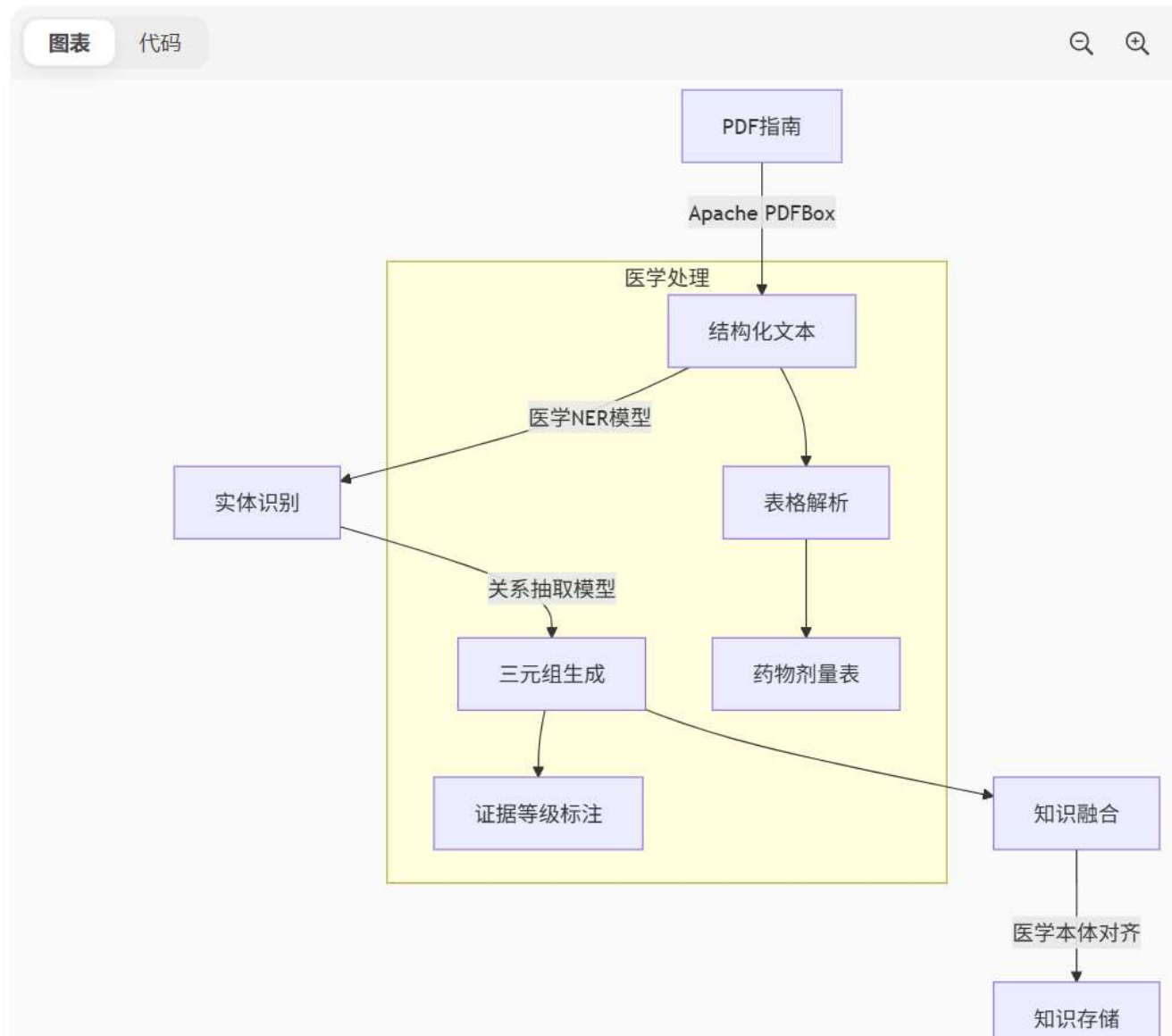
- 剂量表达式解析：100mg qd → {value:100, unit:mg, frequency:daily}
- 实验室指标动态范围：HDL-C > 40mg/dL
- 证据等级标注：Class I, LOE A

构建步骤（示例）

- (1) 命名实体识别 (NER)
 - **任务**: 识别文本中的实体 (人物、地点、组织等)
 - **模型**:
 - 传统: CRF、HMM
 - 深度学习: BiLSTM+CRF、BERT (微调)
 - **训练数据**: 标注文本 (如 `[爱因斯坦]PER` 出生于 `[德国]LOC`)
- (2) 关系抽取 (RE)
 - **任务**: 识别实体间关系 (如“爱因斯坦-出生地-德国”)
 - **方法**:
 - 监督学习: **BERT关系分类模型** (输入: 实体对上下文)
 - 弱监督: **远程监督** (利用知识库对齐文本)
 - 联合抽取: **GlobalPointer**、**SPAN模型** (避免管道误差)
- (3) 属性抽取
 - 提取实体属性 (如“爱因斯坦-出生日期-1879年3月14日”)
 - 方法: 规则模板、序列标注模型

构建步骤（示例）

阶段4：医学知识抽取流水线



医学三元组示例：

turtle

```
:心绞痛      :首选药物      :β受体阻滞剂 .  
:β受体阻滞剂 :禁忌症        :哮喘          .  
:阿司匹林    :最大剂量      :"325mg/d"^^xsd:string .  
:心电图      :诊断敏感性    :"45%"^^xsd:float .
```

阶段5：医学知识融合与验证

关键挑战与解决方案：

挑战	医学解决方案
术语冲突	UMLS (统一医学语言系统) 对齐
证据矛盾	临床指南优先级 > 文献 > 病历
时效性	添加时间标签 (指南版本2023)
地域差异	标注适用地区 (中国/美国)

构建步骤（示例）

阶段6：医学知识存储与应用

Neo4j医学图谱示例：

```
cypher

// 创建疾病节点
CREATE (:Disease {
  name:"冠心病",
  icd10:"I25.1",
  guideline:"ESC 2023"
})

// 创建药物节点
CREATE (:Drug {
  name:"阿托伐他汀",
  atc_code:"C10AA05",
  dose_range:"10-80mg/d"
})

// 创建治疗方案关系
MATCH (d:Disease {name:"冠心病"})
MATCH (dr:Drug {name:"阿托伐他汀"})
CREATE (d)-[:FIRST_LINE_TREATMENT {
  evidence_level:"Class I",
  recommendation:"所有患者应使用他汀类"
}]->(dr)
```

医学领域特殊处理

1. 隐私保护：

- 病历数据：差分隐私处理
- 实体脱敏：【患者ID】 → PAT001

2. 动态知识更新：

```
python

# 监控新指南发布
def monitor_guideline_updates():
    if new_guideline_released("ACC/AHA"):
        trigger_knowledge_refresh()
```

3. 多语言支持：

- 构建中英文医学概念映射表
- 跨语言实体链接（如“心肌梗死” ↔ “Myocardial Infarction”）

医学知识图谱验证指标

评估维度	医学标准
临床准确性	专家评审通过率 > ██████████
时效性	指南更新后72小时内同步
覆盖率	核心疾病诊疗点覆盖 > ██████████
应用价值	CDSS推荐采纳率 > ██████████

实际案例：某三甲医院心血管知识图谱实施后，平均处方合理率从 ██████% 提升至 ██████%，药物冲突预警准确率达 ██████%。

通过此医学定制化流程，可构建符合临床需求的高质量知识图谱，为智慧医疗提供核心支持。

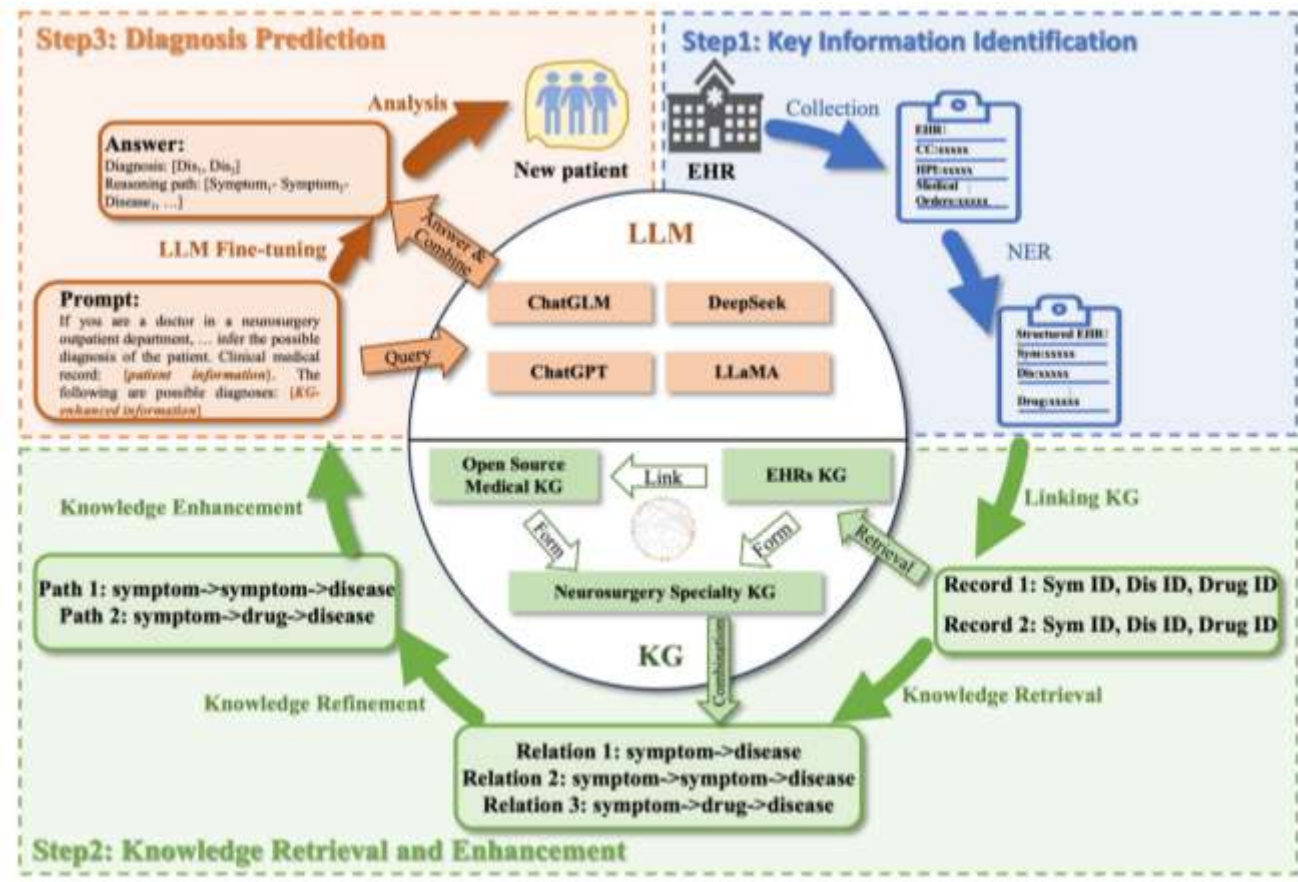
构建步骤（总结）

总结

构建知识图谱的核心是**将非结构化数据转化为结构化知识**，需结合NLP、图计算、机器学习技术。AI模型训练重点包括：

1. **数据标注**：高质量标注决定上限
2. **模型选型**：BERT类模型已成为NER/RE基础
3. **流水线优化**：避免误差传播（联合模型优于管道模型）
4. **持续迭代**：通过用户反馈更新图谱

知识图谱RAG应用案例介绍：NRAG



1. **LLM-KG融合框架：**通过问答形式整合医学知识图谱和大语言模型，实现可解释的诊断推理。
2. **神经外科问答数据集：**本研究构建了专门为神经外科疾病诊断设计的新型问答数据集，它包含强调可解释性的注释，使其更适合医学知识推理和临床决策场景。
3. **卓越的实验性能：**实证评估表明，NRAG在门诊部数据集上取得了优异性能（F1分数=0.8112），超越了通用开源大语言模型（如ChatGPT、DeepSeek）和专业医疗大语言模型。专家评估和案例分析进一步证实，NRAG的诊断建议为神经外科实践提供了有临床意义的指导。

知识图谱RAG应用案例介绍：NRAG-关键技术架构

第一步：关键信息识别

利用信息提取和实体链接技术，从输入数据（例如电子病历）中快速识别患者的症状表现。

- 对所有文本内容进行严格筛查，移除任何可识别的个人信息
- 通过预处理步骤构建高质量结构化数据集
- 整个工作流程遵守伦理指南

采用NLP技术（BERT-CRF、BERT-BiLSTM-CRF）从电子病历中提取症状、疾病、用药等实体；对提取的实体进行标准化处理，形成症状集合（S）和疾病集合（D），构建归一化的症状-疾病映射关系。

知识图谱RAG应用案例介绍：NRAG-关键技术架构

第二步：知识检索与论证

采用开源医学知识图谱，根据患者信息检索相关领域知识并组织成潜在的疾病相关关联路径，反馈到系统中。

- 以患者症状信息为源头，检索知识图谱的可靠知识
- 特别关注"症状-疾病"多跳路径
- 将检索到的核心路径作为输入的一部分，丰富模型的提示信息

这种方式使NRAG不仅能从知识图谱中检索可靠知识进行知识增强，还能优化大语言模型推理的准确性和依据

知识图谱基础

整合多源数据（百万级病历+UMLS、SympGAN等公共医学知识库），构建含6.8k实体、330k关系的神经外科领域KG；

检索路径设计

聚焦“症状-疾病”相关单跳/多跳路径，包括症状→疾病、症状→药物→疾病、症状→新症状→疾病等7类核心路径；

检索优化

采用Personalized PageRank（PPR）算法筛选关键节点，结合医学先验约束（限制1-2阶邻域），平衡检索效率与临床相关性；

路径筛选策略：

策略1（路径优先）：统计子路径频率，按重要性保留Top- α 路径；
策略2（节点优先）：当子路径稀疏时，统计疾病节点频率，保留Top- β 比例疾病作为候选诊断。

知识图谱RAG应用案例介绍：NRAG-关键技术架构

第三步：诊断预测

使用在神经外科问答数据集上微调的大语言模型进行疾病预测，同时为推理过程提供可解释的知识图谱路径。该框架实现了结构化医学知识与神经推理能力的系统整合，确保诊断的准确性和可解释性。

基础模型与微调

以ChatGLM3-6B为基础模型，采用P-Tuning v2进行参数高效微调；

提示词设计

构建临床场景化模板，格式为“角色定义+患者病历信息+KG增强信息”，引导模型生成诊断结果；

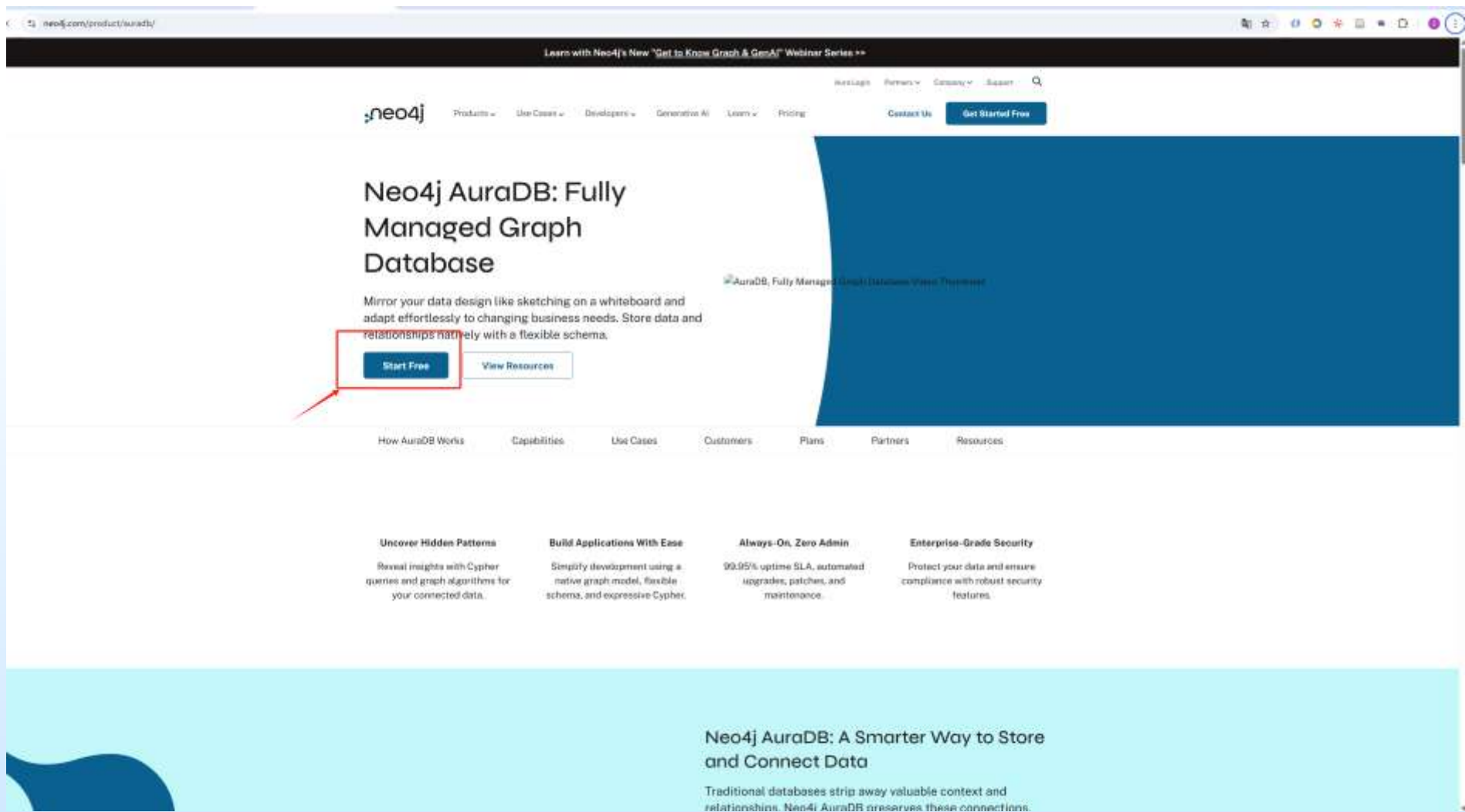
任务形式

将疾病诊断转化为多标签文本分类问题，模型输入为（提示词+归一化症状+KG增强信息），输出为预测诊断（D'）。

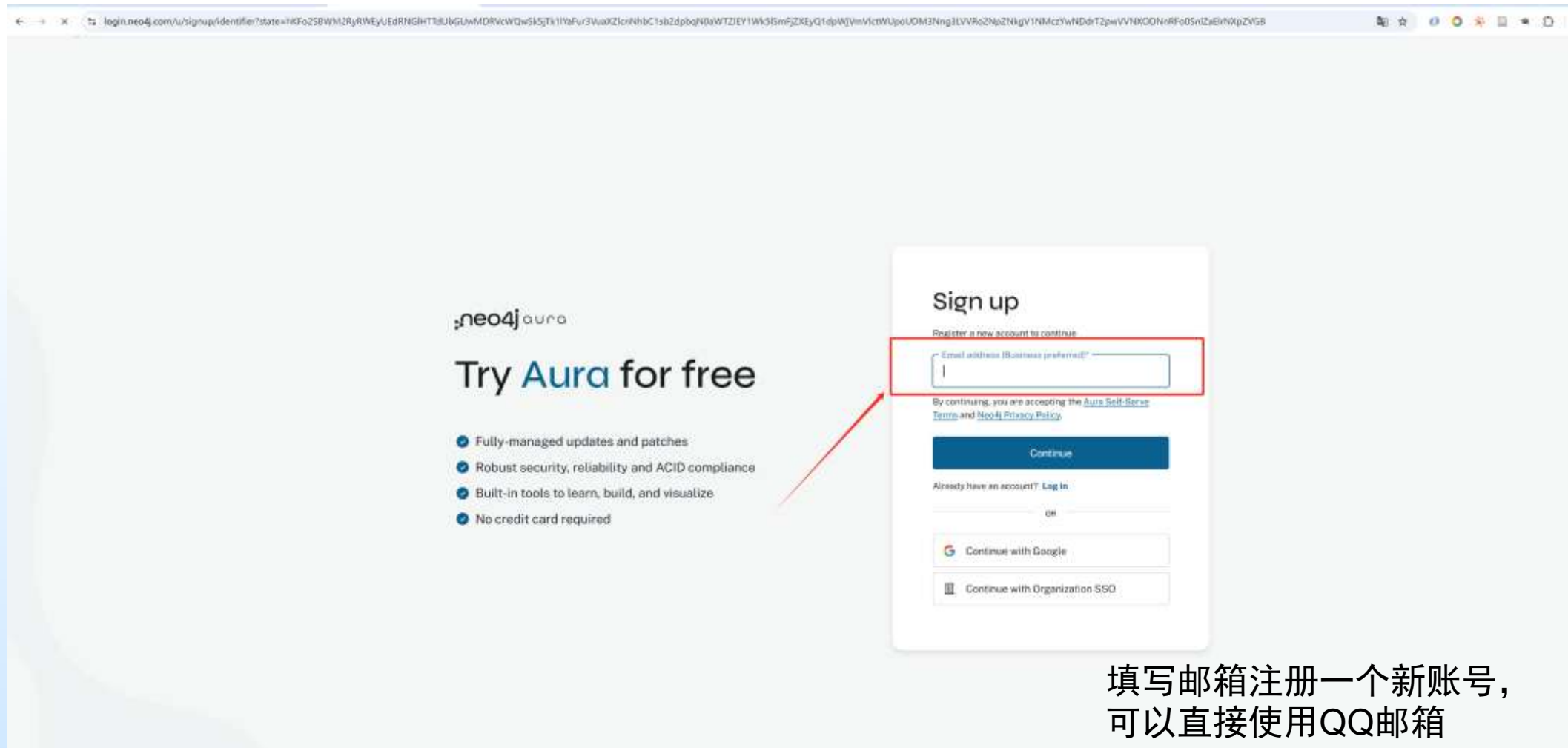
知识图谱构建：图数据库搭建

去官方注册一个账号，登录，然后创建一个数据库实例。（用于学习的话选择免费的就行。）

知识图谱构建：图数据库搭建



知识图谱构建：图数据库搭建



neo4j aura

Try Aura for free

- Fully-managed updates and patches
- Robust security, reliability and ACID compliance
- Built-in tools to learn, build, and visualize
- No credit card required

Sign up

Register a new account to continue


Email address (Business preferred)


By continuing, you are accepting the [Aura Self-Serve Terms](#) and [Neo4j Privacy Policy](#).

Continue

Already have an account? [Log In](#)

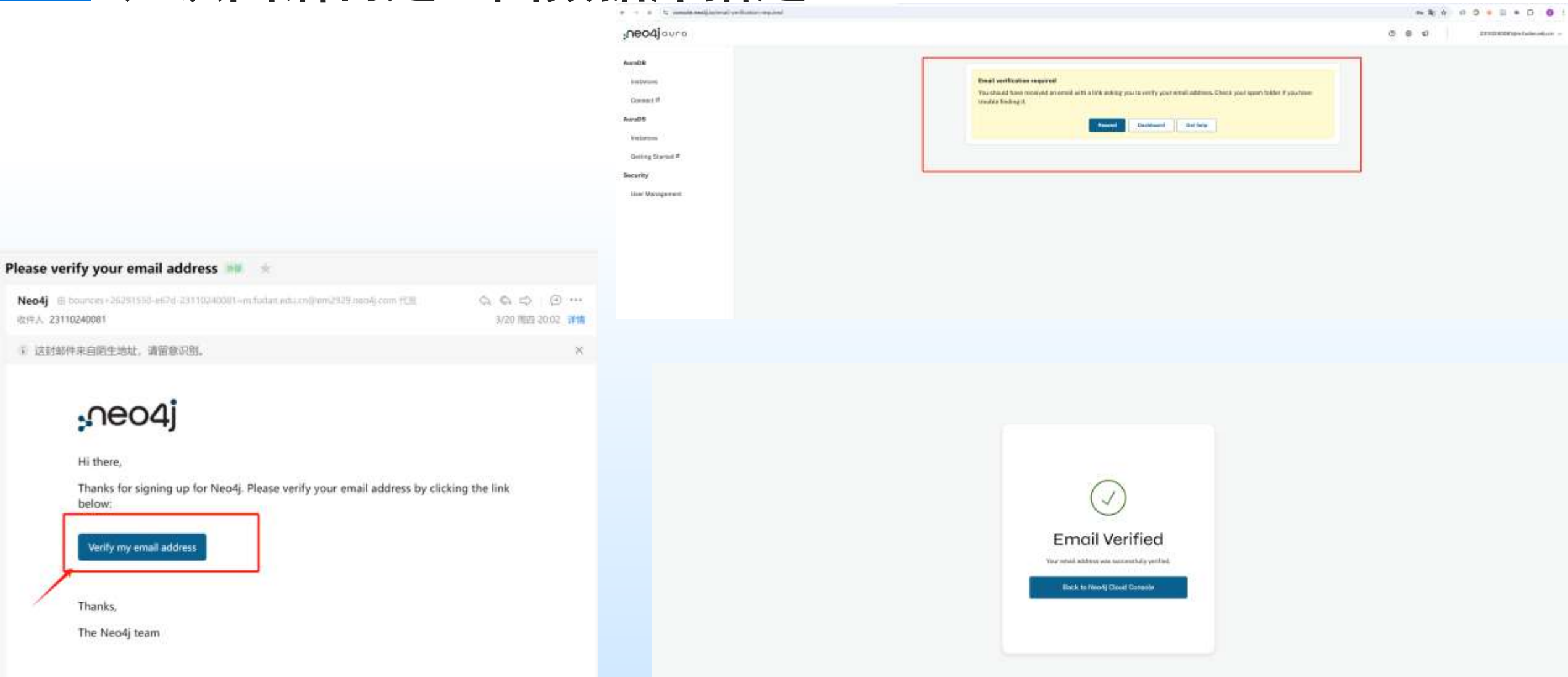
or

 Continue with Google

 Continue with Organization SSO

填写邮箱注册一个新账号，
可以直接使用QQ邮箱

知识图谱构建：图数据库搭建



填写完邮箱和密码后，打开自己邮箱，找到验证邮件，点击验证

知识图谱构建：图数据库搭建

console-preview.neo4j.io/onboarding

neo4j aura

Welcome

Thanks for signing up! We just need a little bit of information to help you best get started with Neo4j.

• • • • •

First (Given) Name

Last Name

Company / Institution

We will send occasional emails about products, services, and events. You can unsubscribe at any time.

Next

console-preview.neo4j.io/onboarding

neo4j aura

Why did you sign up for Neo4j Aura?

• • • • •

Working on a commercial project

Working on a personal project

console-preview.neo4j.io/onboarding

neo4j aura

What best describes your role?

• • • • •

Software Developer

Data Scientist

Data Analyst

Operations Manager

Architect / Tech lead

Student

Other

neo4j aura

What is your use case?

• • • • •

Generative AI

Knowledge Graph

Fraud Detection

Identity Access Management

Master Data Management

Network and IT Operations

Recommendations

Data Privacy, Risk, and Compliance

Supply Chain Analytics

Other

填写姓名等基本信息

知识图谱构建：图数据库搭建

The screenshot displays the Neo4j console interface. On the left, a sidebar contains the text "neo4j Aura" and "Get started with a free instance". A red box highlights a blue button labeled "Create instance" with the subtext "No credit card required". Below this button, there is a link: "Not looking to create a free instance? Select another instance type".

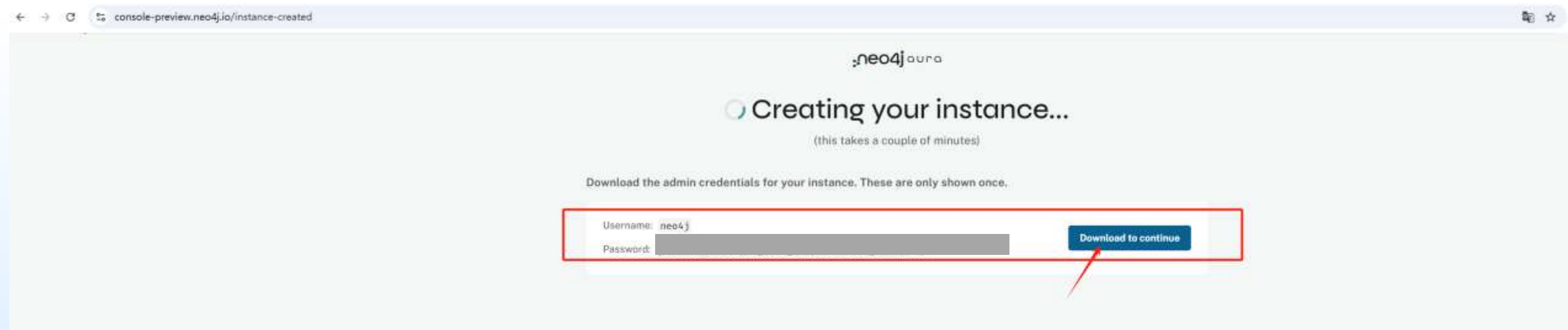
On the right, a graph visualization shows five nodes connected by directed edges:

- Fudan University** (purple circle)
- Zongan Huang** (yellow circle)
- Student** (green circle)
- Knowledg-e...** (yellow circle)
- Working on a pers...** (teal circle)

The relationships between the nodes are as follows:

- Zongan Huang** *WORKS AT* **Fudan University**
- Zongan Huang** *ROLE IS* **Student**
- Fudan University** *HAS USE CASE* **Knowledg-e...**
- Zongan Huang** *WILL USE MEDIA FOR* **Working on a pers...**

知识图谱构建：图数据库搭建



这里的文件一定要下载保存下来里面有连接数据库必要的信息，这里可能需要等待一段时间

```
# Wait 60 seconds before connecting using these details, or login to https://console.neo4j.io to validate the Aura Instance is available
NEO4J_URI=neo4j+s://dabd8fb1.databases.neo4j.io
NEO4J_USERNAME=neo4j
NEO4J_PASSWORD=[redacted]
AURA_INSTANCEID=dabd8fb1
AURA_INSTANCENAME=Free instance
```

知识图谱构建：图数据库搭建



完成远程图数据库的创建

谢谢大家

