QCon
全球软件开发大会

InfoQ 极客传媒

动态化与参数化RAG技术探索

艾清遥

清华大学计算机系副教授

目录

InfoQ 极客传媒

北京

QCon 全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

北京

AiCon 全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

上海

QCon 全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月 — 5月 — 6月 — 8月 — 10月 — 12月

上海

AiCon 全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

深圳

AiCon 全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

北京

AiCon 全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

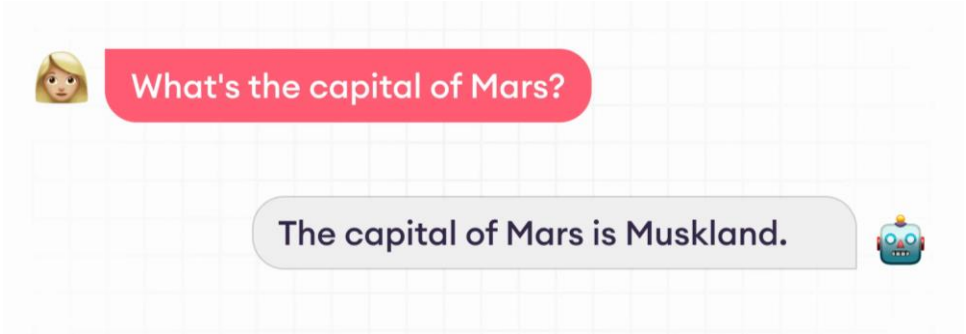
大语言模型的崛起

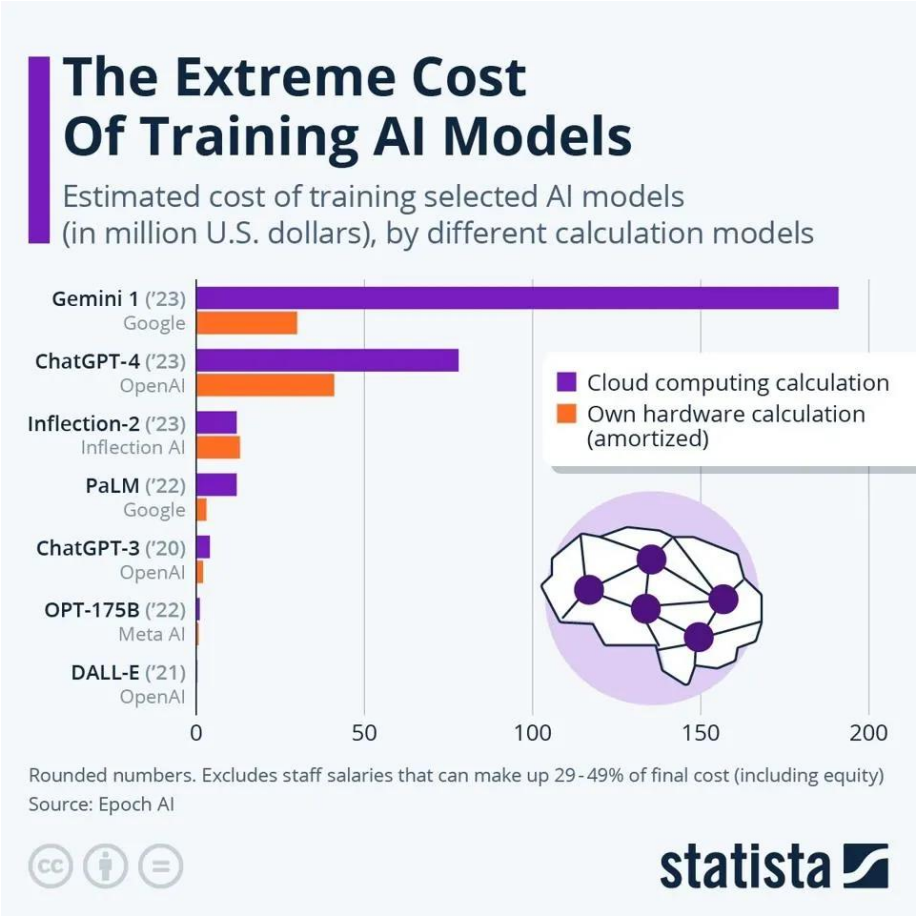
- 作为信息工具，大模型有诸多优点：
  - 自然的用户交互方式
  - 自然语言理解与推理能力
  - 强大的任务泛化能力

获取1亿用户所用时间

大语言模型的崛起

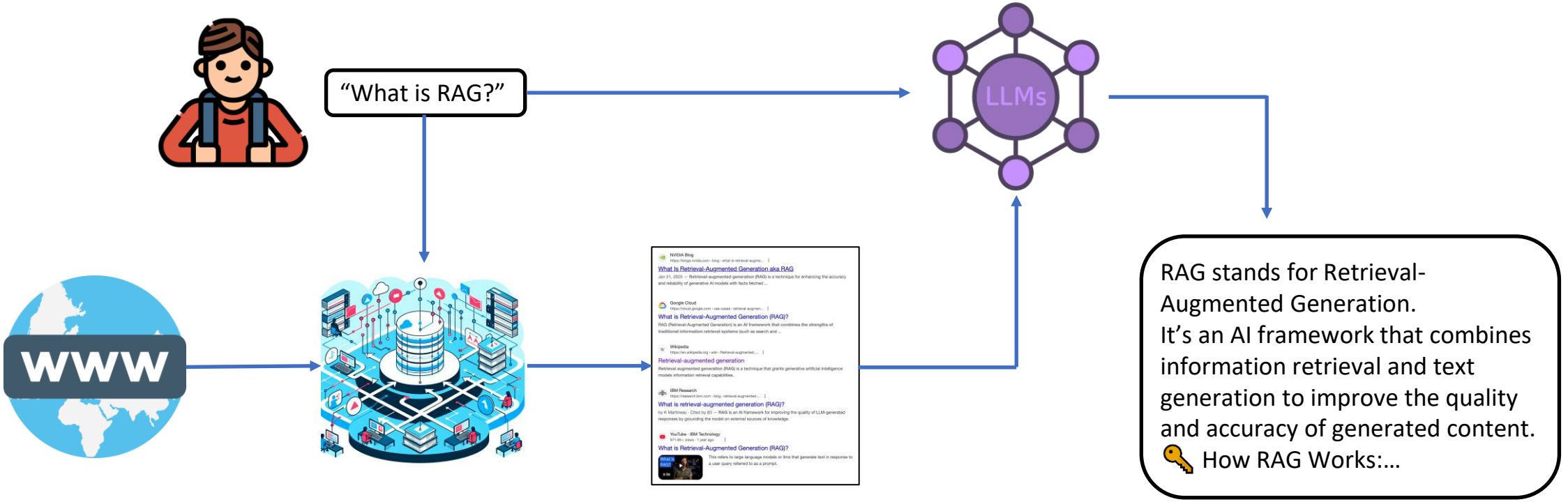
- 与传统检索系统相比，大模型也存在致命缺陷：
  - 严重的幻觉问题
  - 无法追溯的信息来源
  - 极高的成本和极低的灵活性





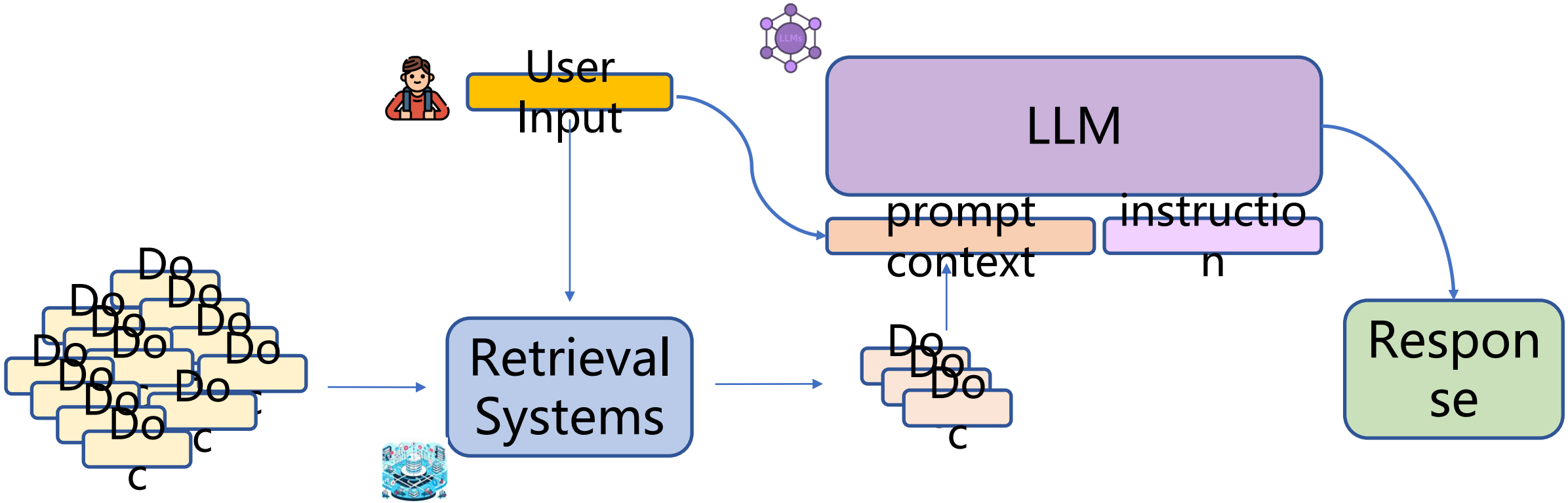
检索增强的生成范式 RAG

- 检索增强的生成范式（RAG）是解决大语言模型上述问题的最有效方式



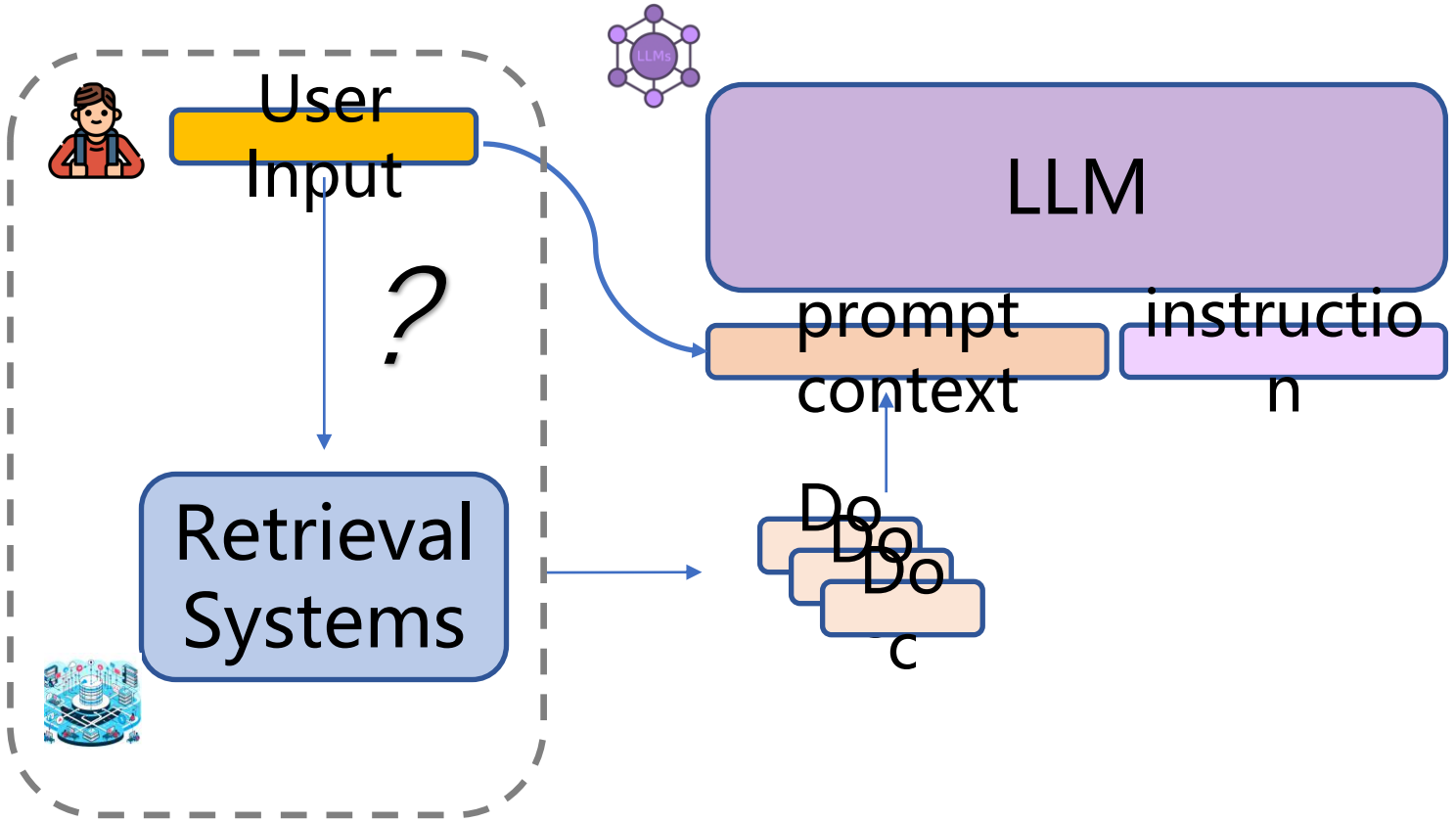
检索增强生成与知识注入

- 检索增强生成的本质是一种对生成式大模型进行外部知识注入的方式



检索增强生成与知识注入

- 大模型何时需要外部知识的注入?
  - 常见方法
    - 默认大模型时刻都需要外部知识
    - 让用户点选是否大模型需要外部知识
  - “Deep Research”
    - 人工定义工作流，并按工作流进行检索调用
    - 把检索功能当成工具，并训练大模型使用

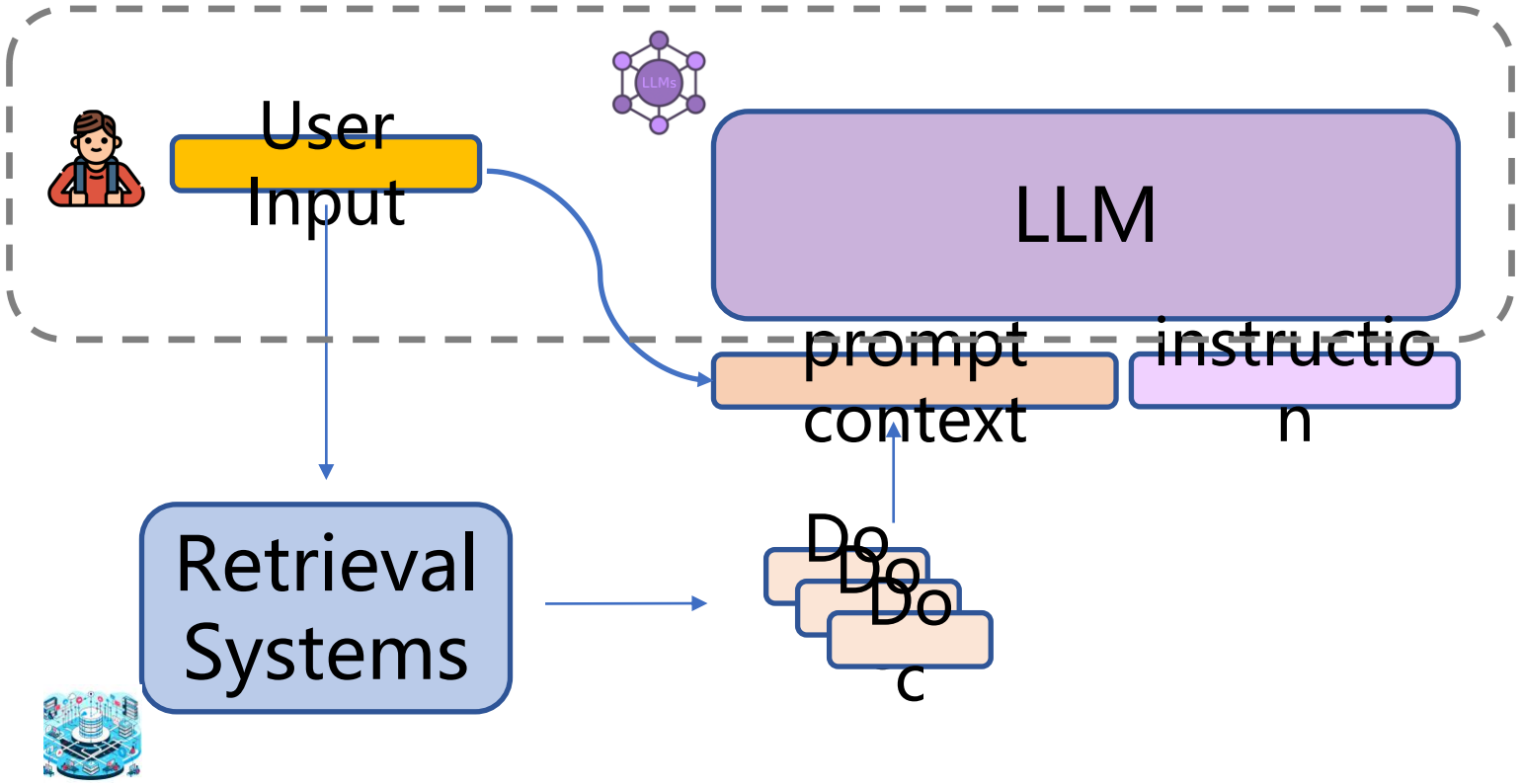


检索增强生成与知识注入

- 大模型如何生成合适的查询来检索外部知识?

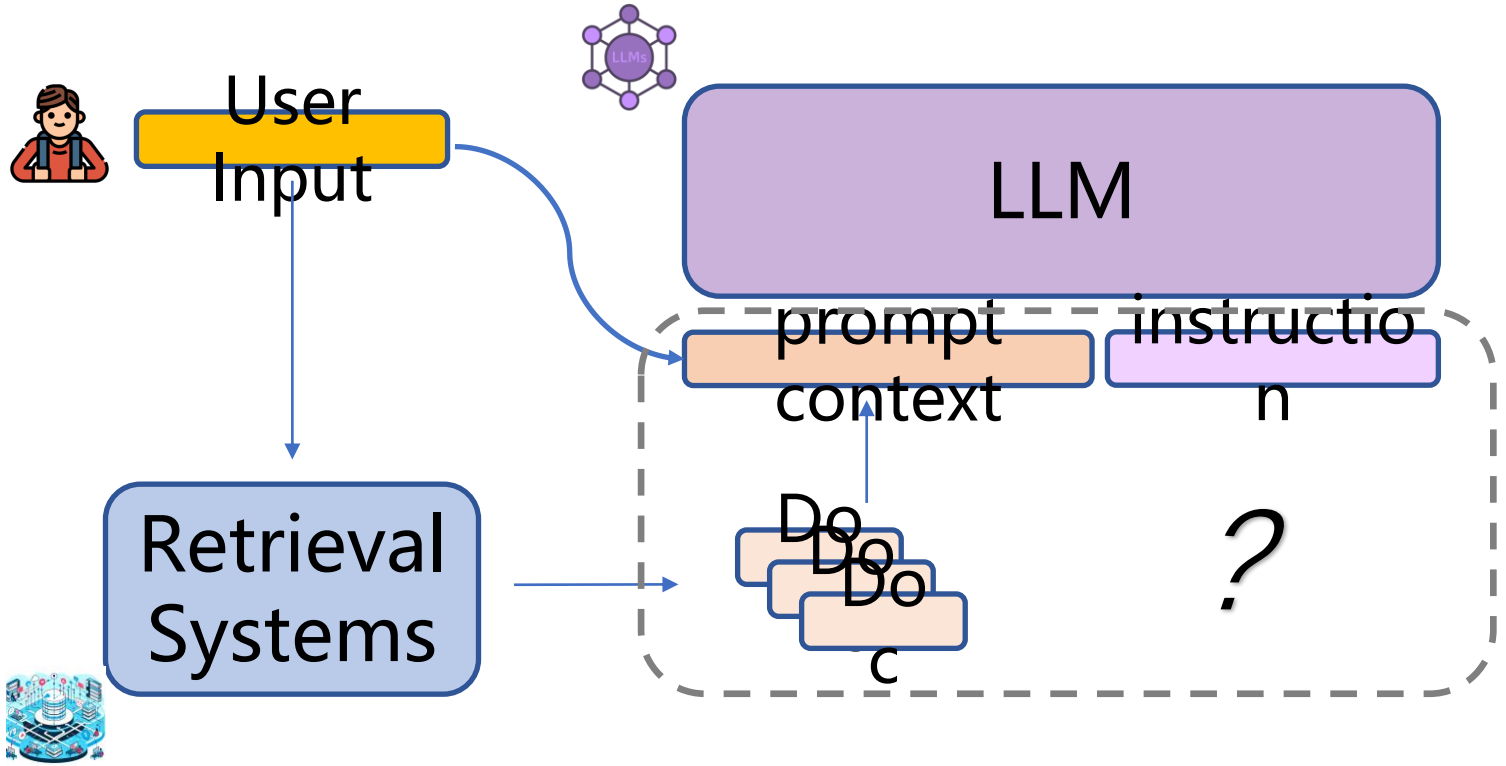
?

  - 常见方法
    - 直接把用户输入当成查询词
  - “Deep Research”
    - 人工定义查询词生成工作流
    - 提示或者训练大模型来提出查询词



检索增强生成与知识注入

- 如何将检索到的外部知识注入大模型?
  - 常见方法
    - 直接把外部文档放进提示词
  - “Deep Research”
    - 人工定义工作流进行多步拆解
    - 最终将外部知识文字放进提示词执行



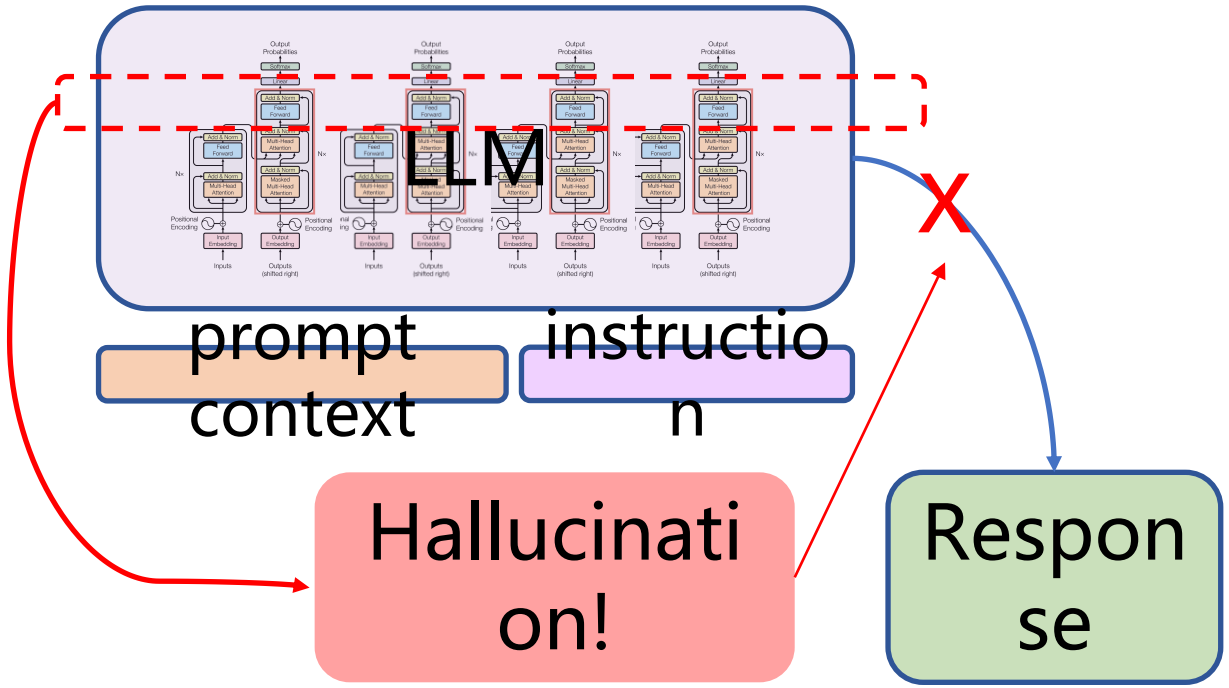
检索增强生成与知识注入

- 本质上，已有的检索增强技术均把大模型当成一个**静态黑盒**使用

问题：
是否可以深入探索大模型的**内部状态**和**信息流动**，
构造出更加**性能更好**且**效率更高**的
检索增强框架呢？

动态化与参数化的检索增强生成

- 深入探索大模型的内部状态，实现动态化、参数化的高效外部知识注入



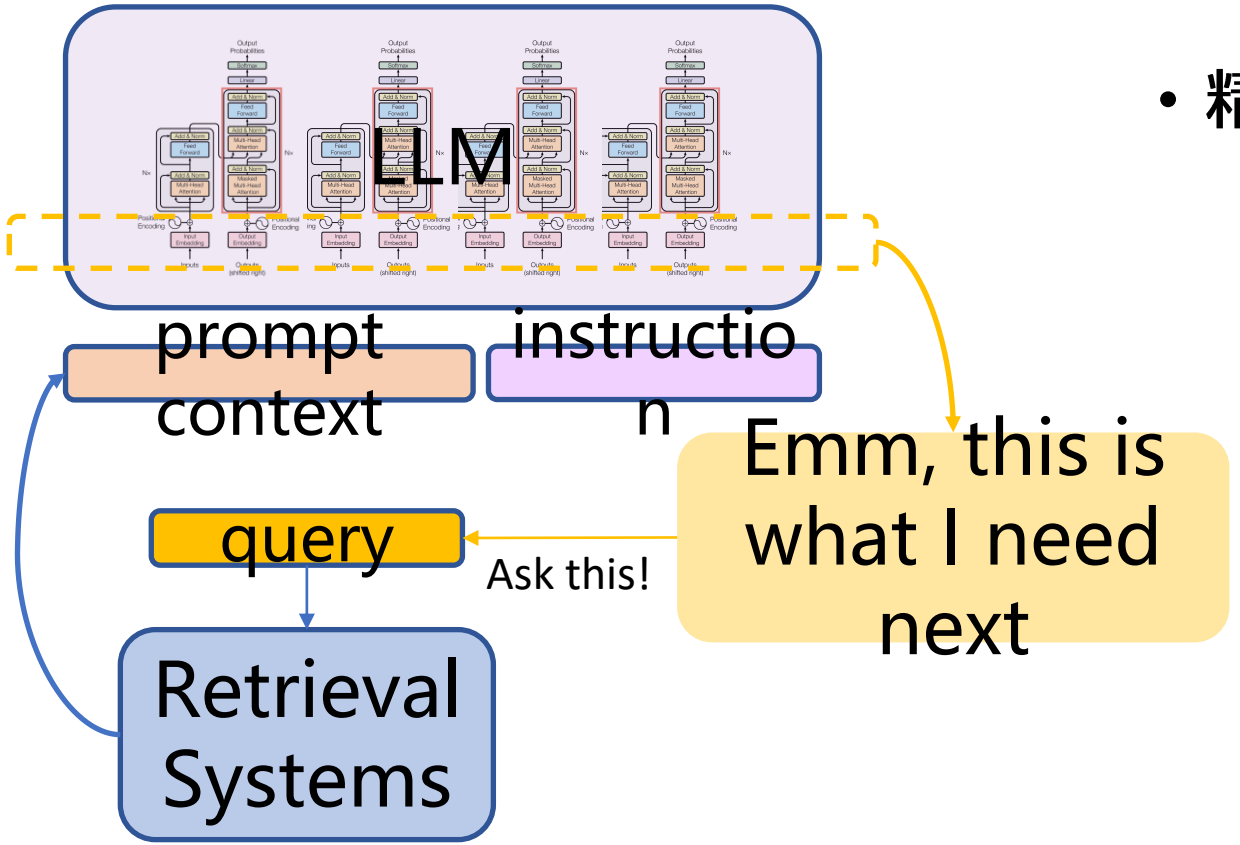
- 在且只在大模型需要时进行检索
  - 分析大模型的内部激活情况
  - 量化生成过程中的不确定性大小

    [Su et al. 2024] [Quevedo et al.] [Herrlein et al.] …
  - 实时监测可能幻觉的生成

    [Su et al. 2024] [Ji et al. 2024] [Song et al. 2024] …

动态化与参数化的检索增强生成

- 深入探索大模型的内部状态，实现动态化、参数化的高效外部知识注入



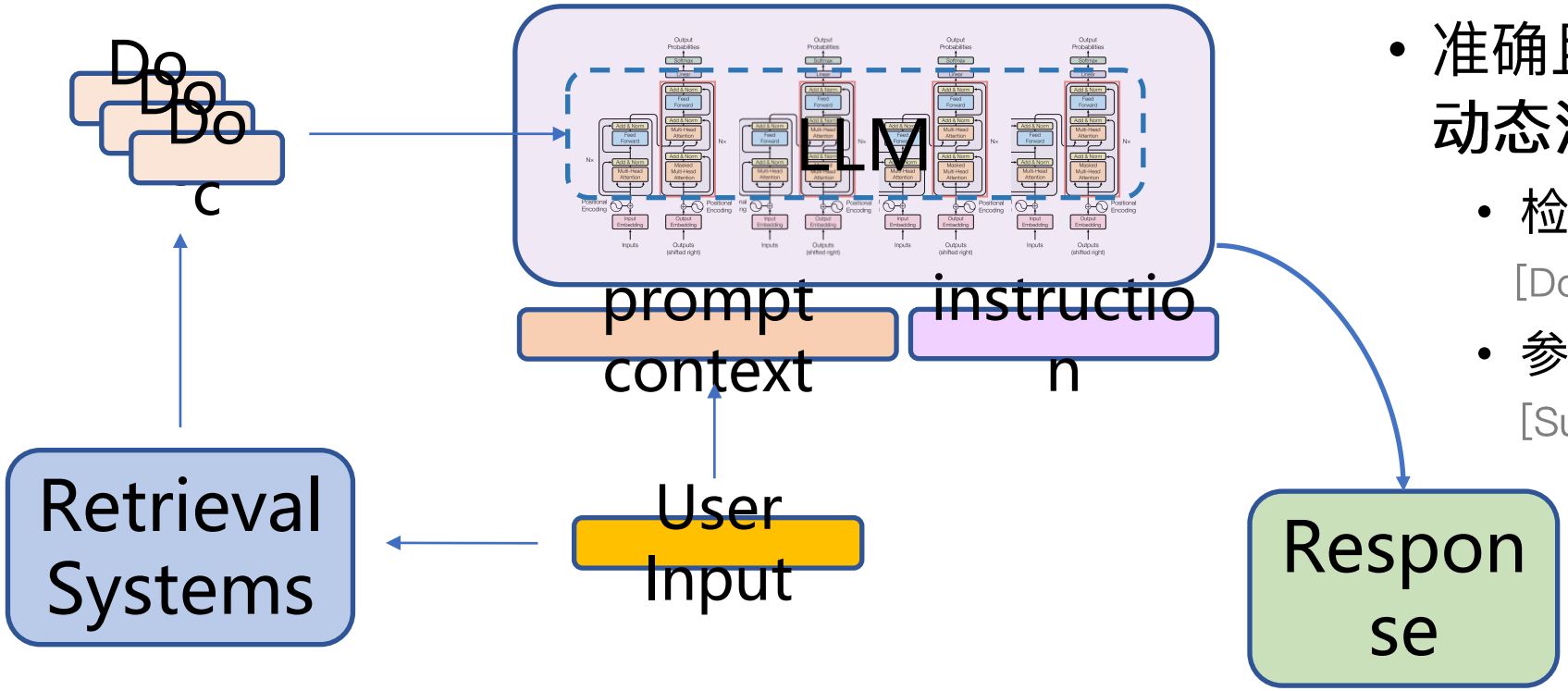
- **精准建模大模型需求的查询信息**
  - 实时捕捉大模型推理过程中的信息需求

    [Su et al. 2024] [Dong et al. 2025]
  - 基于大模型的内部状态构建查询词/向量

    [Su et al. 2024] [Dong et al. 2025] [Jiang et al. 2024]

动态化与参数化的检索增强生成

- 深入探索大模型的内部状态，实现动态化、参数化的高效外部知识注入



- 准确且高效的将外部知识动态注入生成过程
  - 检索与生成过程的解耦运行
    [Dong et al. 2025]
  - 参数化的知识模块构建
    [Su et al. 2025]

01

生成中的动态信息需求建模

检索增强的静态范式与动态范式

检索增强范式分类:

- 静态范式:
  - 在生成过程开始前，基于用户指令或者提前定义好的智能体工作流进行检索

    每次生成过程激活并只激活一次检索过程.
- 动态范式:
  - 在生成过程之中，根据大模型的需求动态调用检索过程

    每次生成过程可以不激活或激活多次检索过程

静态检索生成的局限

简单信息任务：→ 大模型的信息需求可以根据初始指令直接确定。

复杂信息任务：→ 大模型的信息需求可能在推理过程中发生变化。

指令:撰写一篇关于阿根廷在卡塔尔世界杯胜利的长篇评论。

LLM:阿根廷在2022年卡塔尔世界杯的胜利是一场凝聚了戏剧性、激情与……的辉煌。XXX曾说过：世界上只有一种真正的英雄主义，那就是看清世界的本来面目，并热爱它。

初始信息需求：
- 阿根廷队
- 卡塔尔世界杯

推理过程中的信息需求:
“世界上只有一种真正的英雄主义..."
→ 这句话出自于谁?

动态检索生成

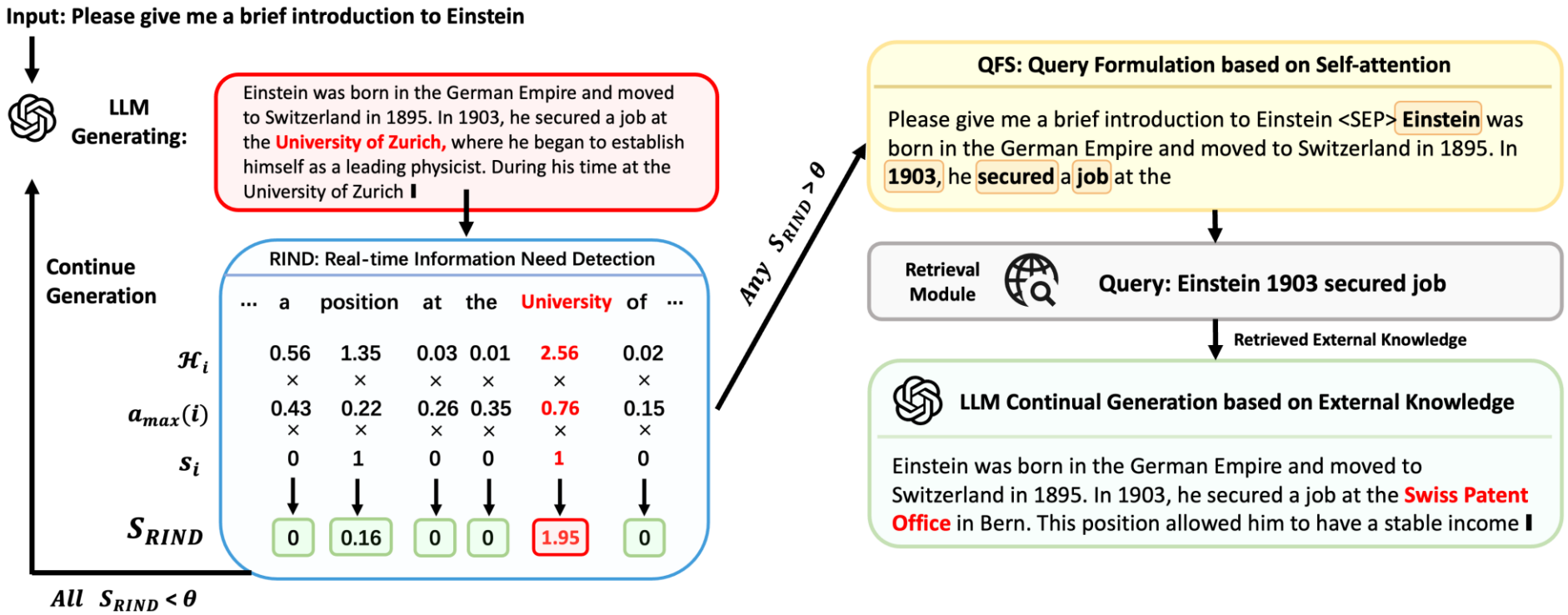
核心问题：

如何在推理过程中发现并解决大型语言模型的信息需求?

- 在生成过程中何时**触发检索**?
- 一旦触发检索，如何判断目标信息并**生成对应查询**?

DRAGIN: 基于大模型信息需求的动态检索增强生成框架

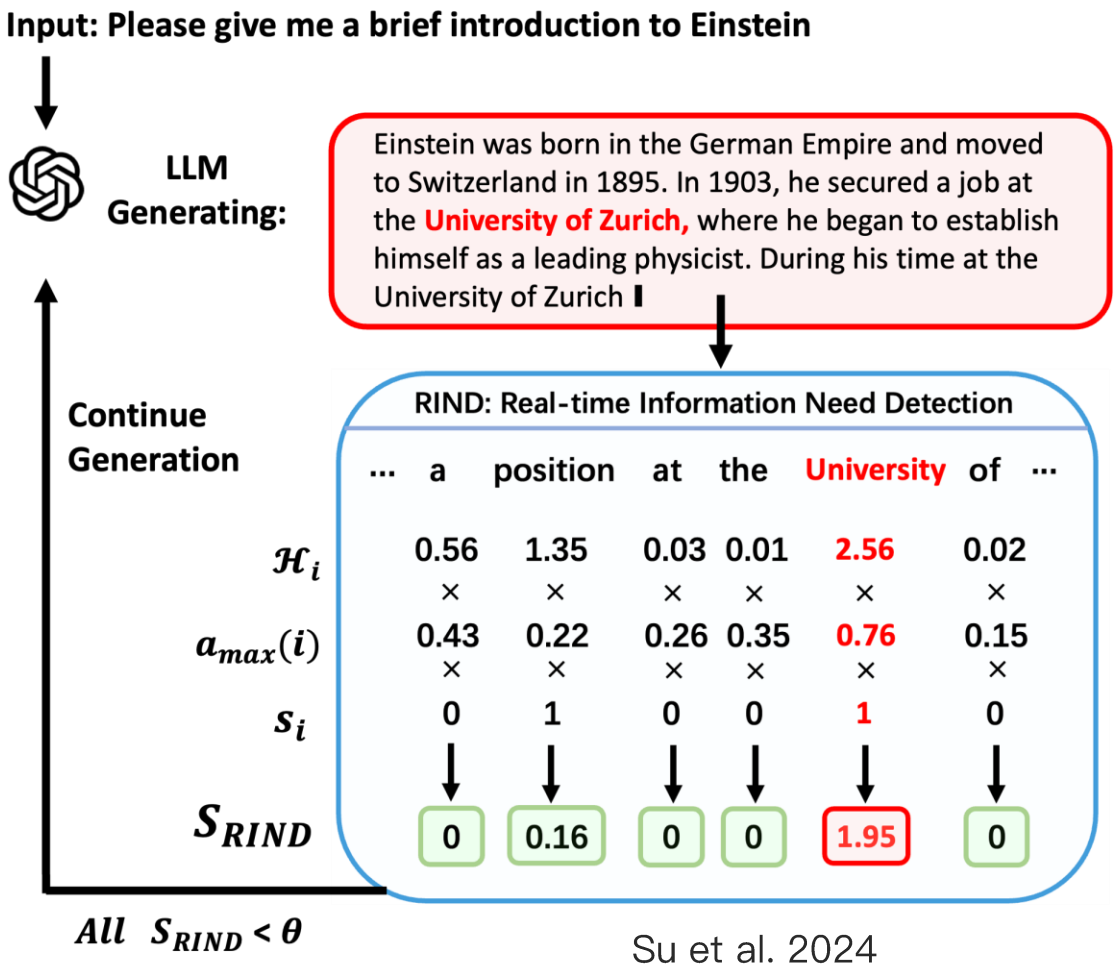
- 在大语言模型生成的过程中，实时根据其**内部状态**（如logits、token熵和注意力分布等）对模型的**信息需求**进行建模。



[Su et al. 2024]

DRAGIN: 基于内部状态监测的检索时机预测

v: 词表中一个token
i: 位置
t_i: 在位置 i 的token
d_k: 向量维度
S: 停用词表



Su et al. 2024

不确定性

$$\mathcal{H}_i = -\sum_{v \in \mathcal{V}} p_i(v) \log p_i(v),$$

重要性

$$A_{i,j} = \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right),$$

$$a_{\max}(i) = \max_{j>i} A_{i,j}$$

语义价值

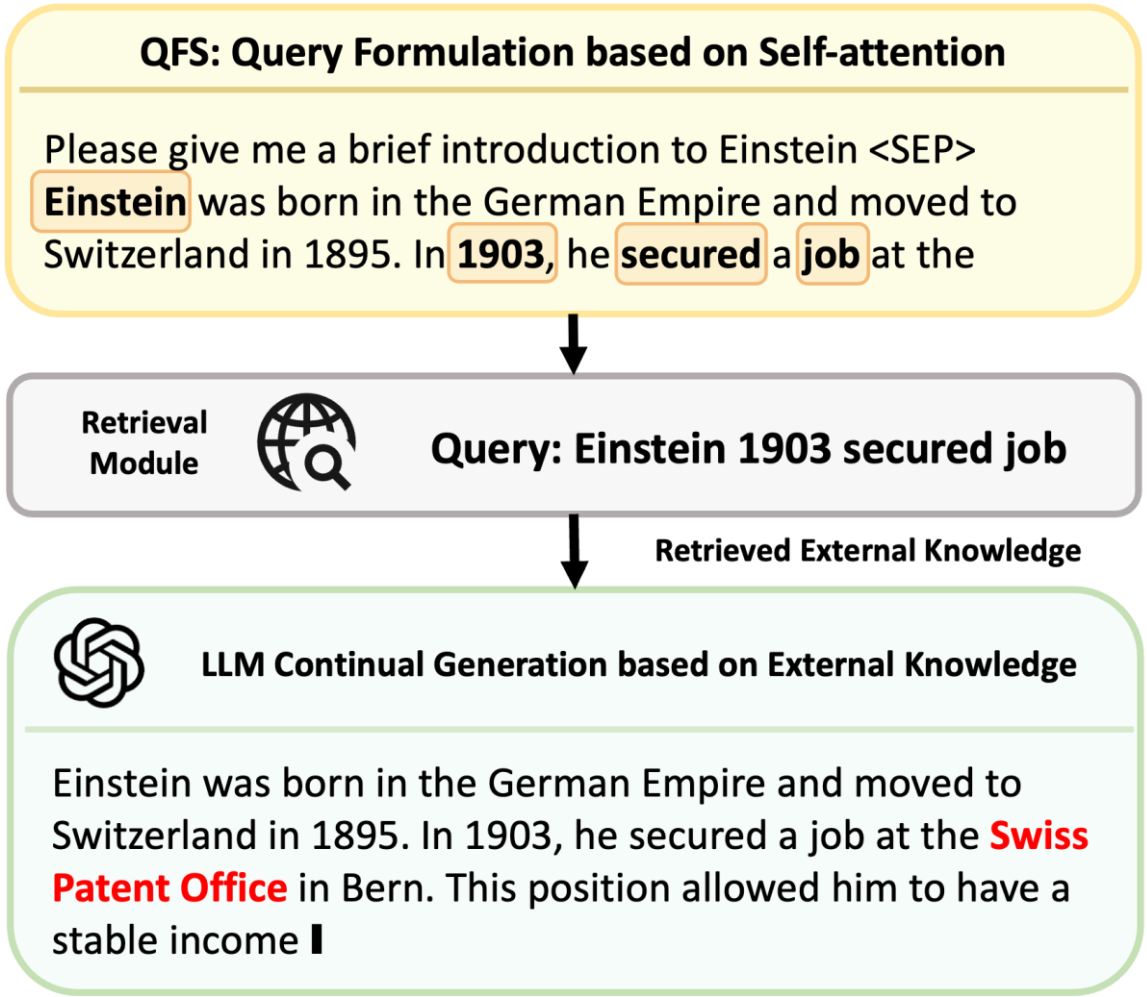
$$s_i = \begin{cases} 0, & \text{if } t_i \in S \\ 1, & \text{otherwise} \end{cases},$$

混合计算

$$\mathcal{S}_{RIND}(t_i) = \mathcal{H}_i \cdot a_{\max}(i) \cdot s_i$$

当 $S_{RIND}(t_i) > \theta$阈值时进行检索调用

DRAGIN: 基于注意力分布的动态检索查询生成



步骤一：
提取每个token在网络最后一层的**注意力得分**。

步骤二：
根据注意力得分对token进行**排序**，并选择前 n 个标记。

步骤三：
使用前 n 个token**构建查询**。

步骤四：
将检索到的知识添加到提示词上下文中，并让大型语言模型**从截断点继续生成内容**。

[Su et al. 2024]

DRAGIN: 实验结果

LLM	RAG Method	2WikiMultihopQA EM	2WikiMultihopQA F1	HotpotQA EM	HotpotQA F1	StrategyQA EM	IIRC EM	IIRC F1
Llama2-13b-chat	wo-RAG	0.187	0.2721					
	SR-RAG	0.245	0.3364					
	FL-RAG	0.217	0.3054					
	FS-RAG	0.270	0.3610					
	FLARE	0.224	0.3076	0.180	0.2756	0.655	0.138	0.1667
	DRAGIN (Ours)	**0.304**	**0.3931**	**0.314**	**0.4238**	**0.689**	0.185	0.2221
Llama2-7b-chat	wo-RAG	0.146	0.2232	0.184	0.2745	**0.659**	0.139	0.1731
	SR-RAG	0.169	0.2549	0.164	0.2499	0.645	0.187	0.2258
	FL-RAG	0.112	0.1922	0.146	0.2107	0.635	0.172	0.2023
	FS-RAG	0.189	0.2652	0.214	0.3035	0.629	0.178	0.2157
	FLARE	0.143	0.2134	0.149	0.2208	0.627	0.136	0.1644
	DRAGIN (Ours)	**0.220**	**0.2926**	**0.232**	**0.3344**	0.641	**0.192**	**0.2336**
Vicuna-13b-v1.5	wo-RAG	0.146	0.2232	0.228	0.3256	0.682	0.175	0.2149
	SR-RAG	0.170	0.2564	0.254	0.3531	0.686	0.217	0.2564
	FL-RAG	0.135	0.2133	0.187	0.3039	0.645	0.0985	0.1285
	FS-RAG	0.188	0.2625	0.185	0.3216	0.622	0.1027	0.1344
	FLARE	0.157	0.2257	0.092	0.1808	0.599	0.1174	0.1469
	DRAGIN (Ours)	**0.252**	**0.3516**	**0.288**	**0.4164**	**0.687**	**0.2233**	**0.2652**

性能显著高于其他使用和不使用RAG的模型框架

DRAGIN: 效率分析

		2WMQA	HQA	SQA	IIRC
		#Num	#Num	#Num	#Num
L13B	FL-RAG	3.770	3.194	3.626	3.426
	FS-RAG	3.131	4.583	4.885	4.305
	FLARE	1.592	3.378	0.625	5.521
	DRAGIN	2.631	3.505	4.786	2.829
L7B	FL-RAG	3.342	3.809	3.757	2.839
	FS-RAG	3.833	4.152	4.546	4.210
	FLARE	0.941	1.064	1.271	1.095
	DRAGIN	2.836	3.013	4.629	2.927
V13B	FL-RAG	4.199	3.564	3.591	3.189
	FS-RAG	3.720	5.701	6.820	6.032
	FLARE	1.093	1.078	1.118	0.335
	DRAGIN	2.542	3.184	3.744	3.120

DRAGIN检索模块**触发频率**：

- 低于RALM和IR–CoT
  - 基于固定规则确定检索调用时机
- 高于FLARE
  - 仅基于token不确定性的触发

[Su et al. 2024]

DRAGIN：部分消融实验

检索时机判断

- 更好的检索时机带来更高性能

		EM	F1	Prec.
L13B	FLARE	0.128	0.1599	0.1677
	FL-RAG	0.155	0.1875	0.1986
	FS-RAG	0.171	0.2061	0.2185
	DRAGIN	**0.187**	**0.2242**	**0.2319**
V13B	FLARE	0.097	0.1277	0.1324
	FL-RAG	0.099	0.1285	0.1324
	FS-RAG	0.103	0.1344	0.1358
	DRAGIN	**0.196**	**0.2367**	**0.2476**

用当前已生成的最后一个句子当成查询

检索查询生成

- 更好的查询生成带来更高性能

		EM	F1	Prec.
L13B	FLARE	0.262	0.3674	0.3792
	Full Context	0.252	0.3584	0.3711
	FS-RAG	0.255	0.3574	0.3685
	FL-RAG	0.241	0.3394	0.3495
	DRAGIN	**0.314**	**0.4238**	**0.4401**
V13B	FLARE	0.225	0.3366	0.3420
	Full Context	0.221	0.3402	0.3457
	FS-RAG	0.216	0.3432	0.3507
	FL-RAG	0.214	0.3268	0.3264
	DRAGIN	**0.288**	**0.4164**	**0.4226**

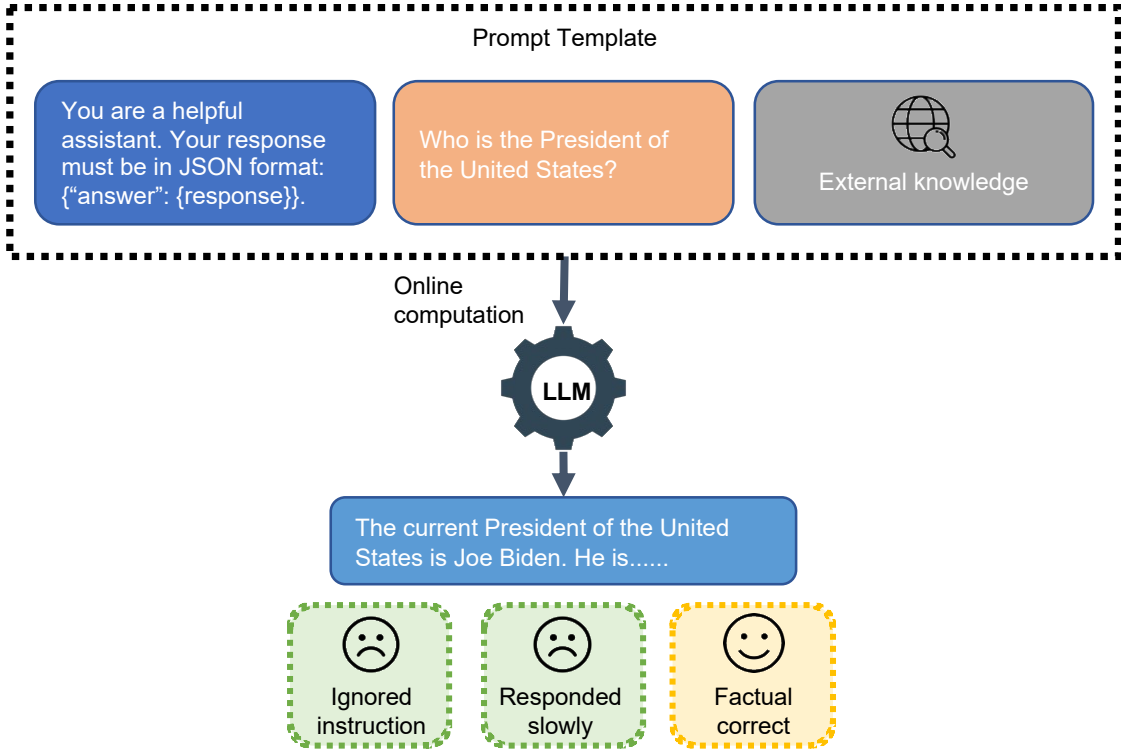
用DRAGIN的方法判断检索时机

02

检索与生成的动态信息解耦

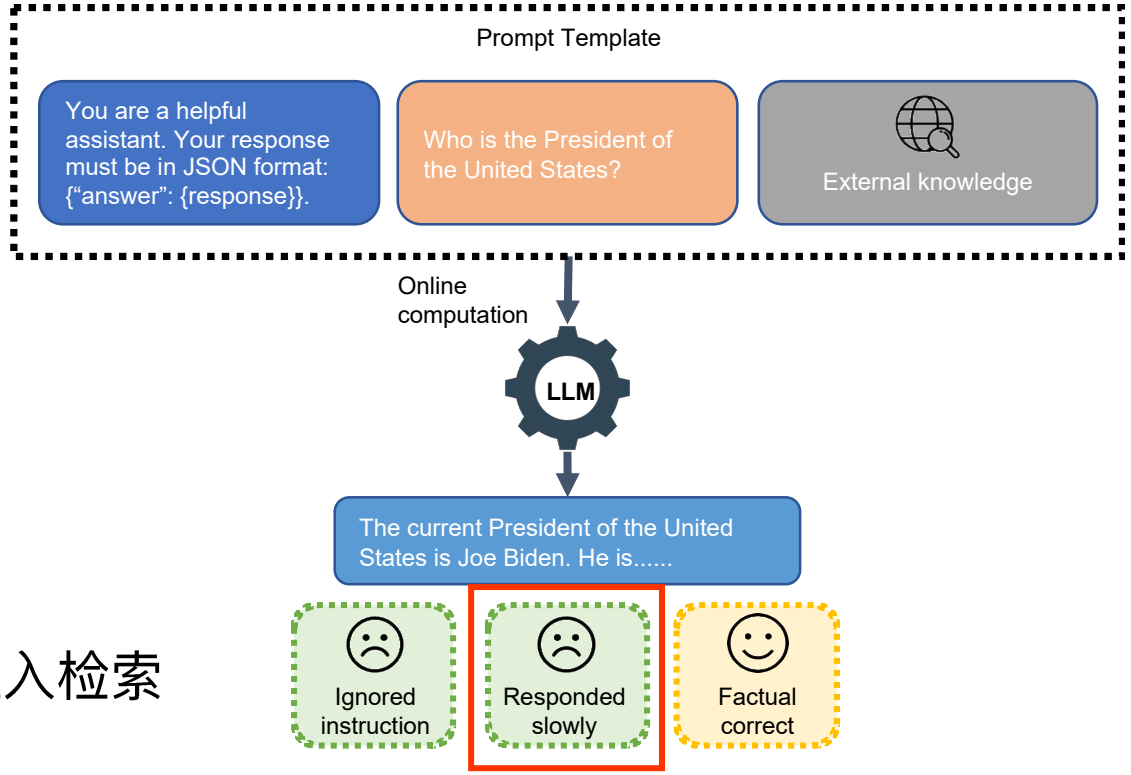
静态检索增强框架中的知识注入

- 传统RAG：基于上下文的知识注入
  - 提前定义指令模板
  - 使用用户问题检索文档
  - 将问题与文档放入提示词中
  - 将提示词输入大模型来激活生成过程



静态检索增强框架中的知识注入

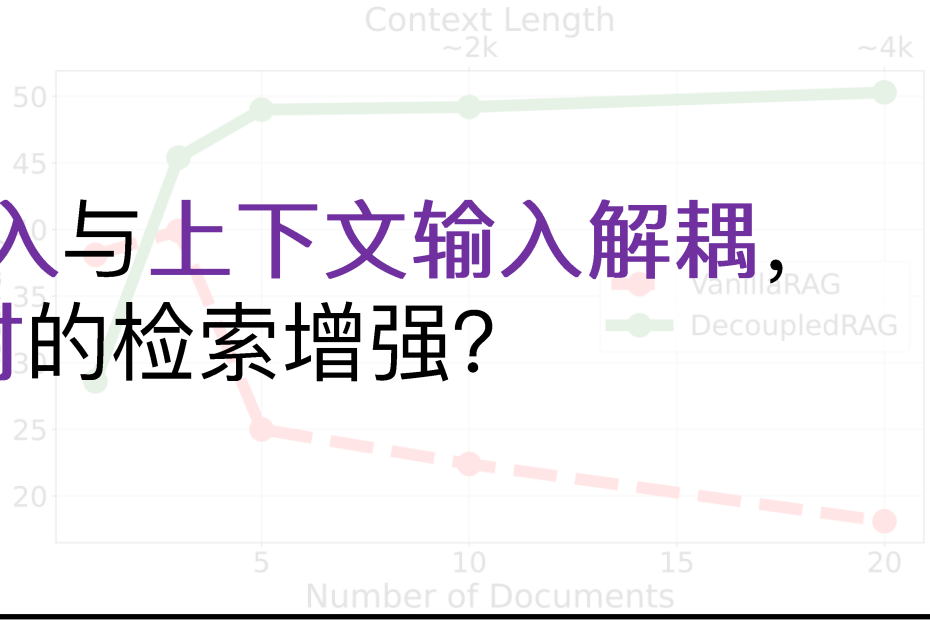
- 效率方面：
  - 存在上下文长度瓶颈
    - 文档可能较长
    - 推理成本增加（FLOPs ⬆）
    - 响应时间延长（时间 ⬆）
  - 频繁打断生成与推理
    - 文档必须输入到提示中
    - 每次检索都必须停止生成，并在注入检索内容后重头开始
    - 消耗大量重复的token



静态检索增强框架中的知识注入

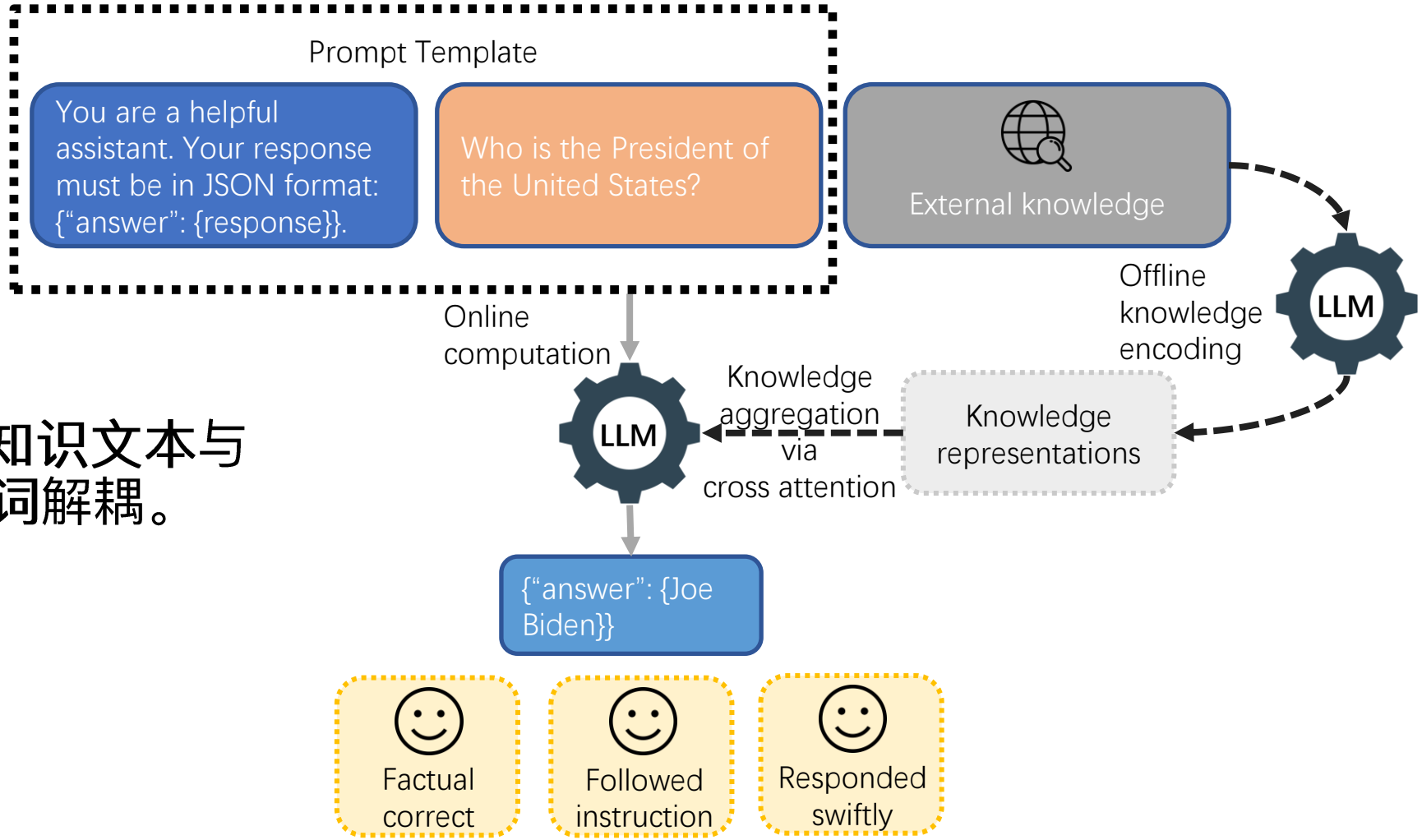
- 性能方面
  - 信息损失
    - “Lost in the middle”
    - 输入顺序敏感
  - 模型退化
    - 干扰大型语言模型的内部推理
    - 影响指令遵循能力

是否可以将检索知识注入与上下文输入解耦，
实现动态且实时的检索增强?

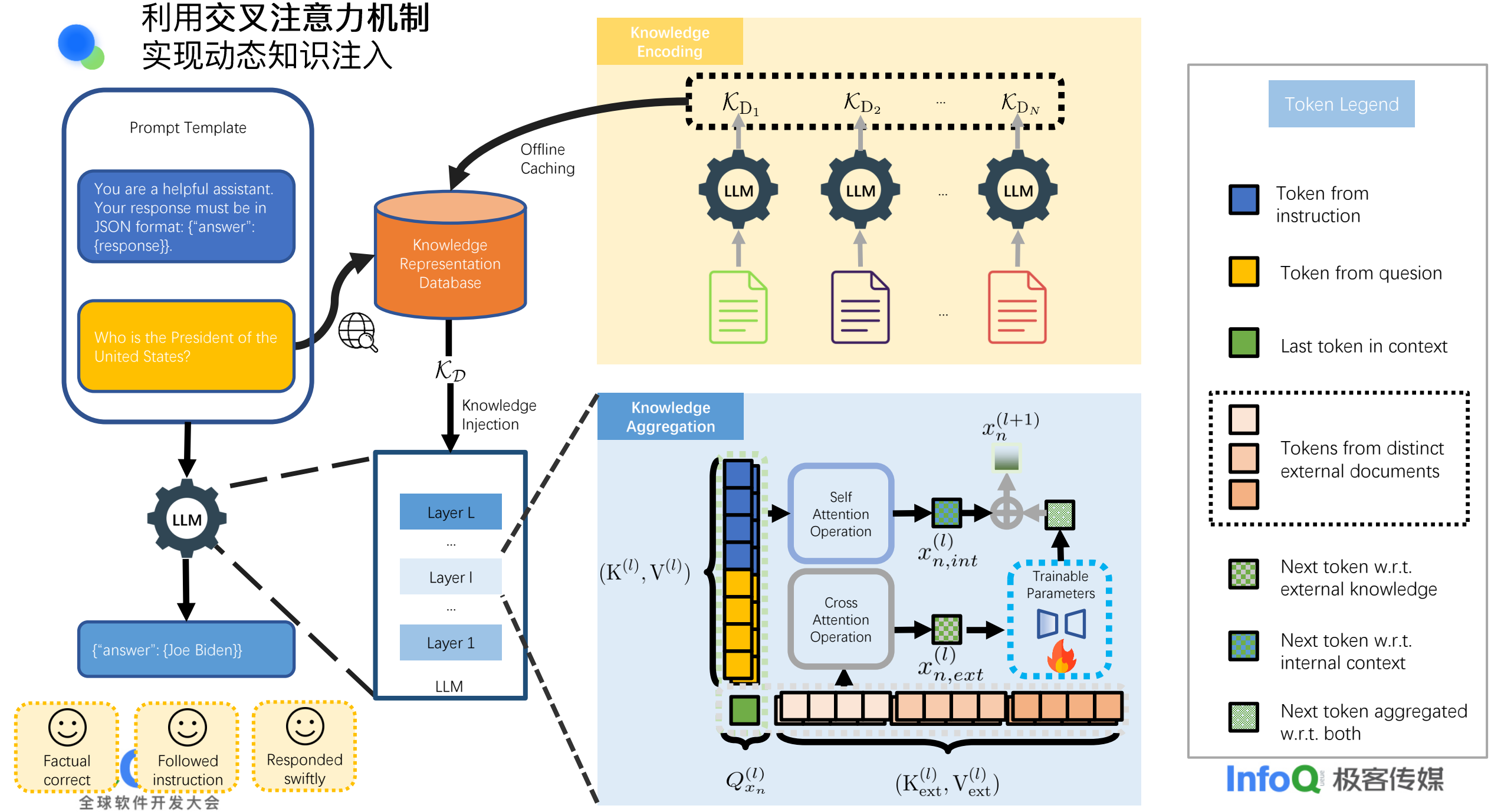


检索与生成的过程解耦

- 我们的目标
  - 将检索的知识文本与输入提示词解耦。

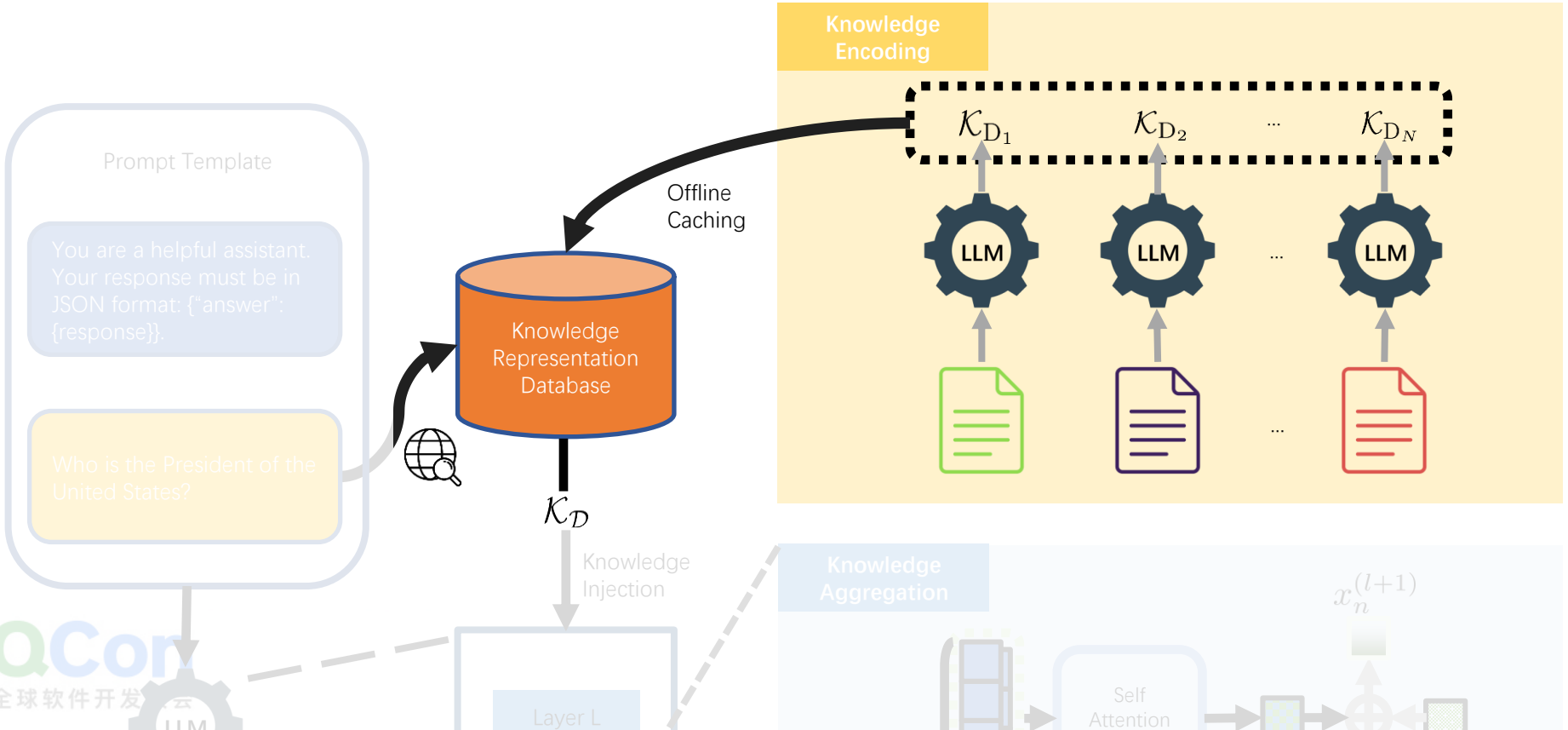


利用交叉注意力机制
实现动态知识注入



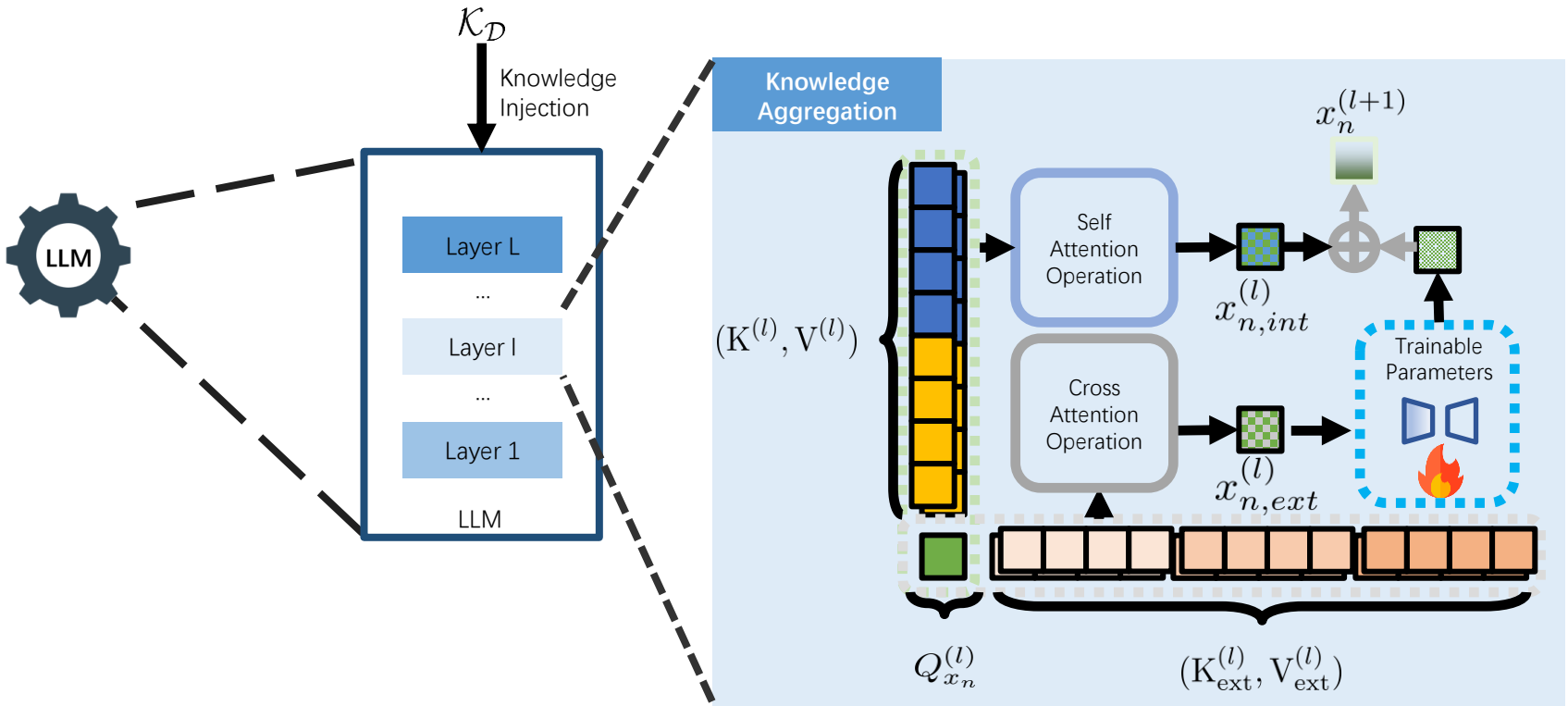
DecoupledRAG：检索与生成的动态信息解耦

- 知识编码
  - 将外部知识预先编码为键值对表示形式（KV Representation）



DecoupledRAG：检索与生成的动态信息解耦

- 通过交叉注意力进行知识聚合
  - 利用交叉注意力注入相关知识的键值表示。



DecoupledRAG：检索与生成的动态信息解耦

	上下文注入 (传统静态RAG)	VS	我们的方法 (DecoupledRAG)
效率	在线 文档编码		离线文档编码
扩展性	串行文档处理		并行文档处理
鲁棒性	**输入顺序敏感** 混淆文档与指令输入		**输入顺序不敏感** 独立 文档与指令输入流

DecoupledRAG：实验设置

- 基线模型
  - RAG FT [Jin et al. 2025]:
    - 根据数据集微调RAG模型.
- 测试模型
  - Llama3-8B-Instruct
  - Llama2-7B-Chat

- 实验数据集
  - 多条问答
    - 2WikiMultihopQA, ComplexWebQuestions
  - 完形填空
    - Zero-Shot RE, T-REx
  - 基于知识的对话生成
    - Wizard of Wikipedia (WoW)
- 实现细节:
  - 文档库: Wikipedia
  - 文档分块: 256 词
  - 检索模型: RetroMAE

DecoupledRAG：实验结果

- 多任务性能比较
  - 当注入更多知识时，DecoupledRAG的表现显著优于RAG FT (Fine-Tuning)

# Docs	Method	Slot Filling				Multi-hop Question Answering				Dialogue
		Zero-Shot RE		T-REx		2WikiMultihopQA		ComplexWebQuestions		WoW
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	F1
		Llama-3-8B-Instruct								
+1 doc	RAG FT	38.1	49.9	61.2	63.2	17.2	12.1	23.4	31.9	22.3
	DecoupledRAG	28.6	37.1	61.1	63.9	15.2	11.1	9.8	16.8	24.4
+3 doc	RAG FT	39.9	51.3	69.3						22.0
	DecoupledRAG	45.4	53.1	73.7						25.5
+5 doc	RAG FT	25.0	29.9	35.7						18.9
	DecoupledRAG	49.0	56.3	76.3						25.9
+10 doc	RAG FT	22.4	28.0	30.2						19.3
	DecoupledRAG	49.2	56.9	78.4	80.3	32.7	37.6	37.6	45.1	25.3
+20 doc	RAG FT	18.1	24.8	25.3	27.4	3.8	9.6	15.7	24.0	19.3
	DecoupledRAG	**50.3**	**57.5**	**80.2**	**81.9**	**33.4**	**38.1**	**39.6**	**47.1**	**26.1**

文档越多，提升越大

DecoupledRAG：效率分析

$$O\left(\overbrace{\overbrace{(N\cdot|\mathrm{D}|)^2}^{\text{Knowledge encoding}}+\overbrace{|\mathrm{A}|\cdot(N\cdot|\mathrm{D}|)}^{\text{Answer generation}}}^{\text{Online inference cost}}\right).$$

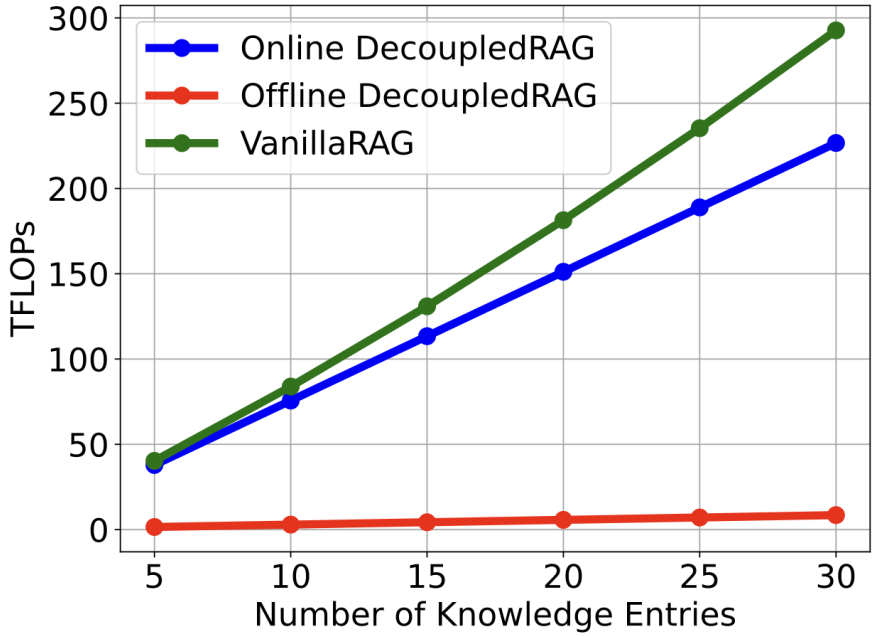
- 传统静态RAG
  - 在线编码过程
  - 复杂度随文档数平方增长

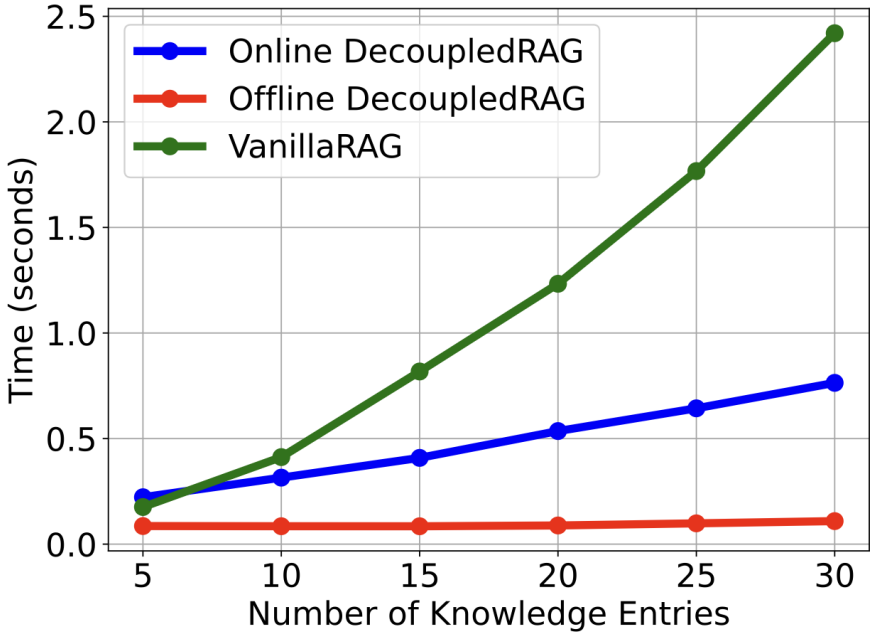
$$O\left(\overbrace{\underbrace{N\cdot|\mathrm{D}|^2}_{\text{Offline inference cost}}+\underbrace{|\mathrm{A}|\cdot(N\cdot|\mathrm{D}|)}_{\text{Online inference cost}}}^{\text{Knowledge encoding}}\right).$$

- Decoupled RAG
  - 离线编码过程
  - 复杂度随文档数**线性增加**

N: 文档数量
|D|: 文档长度
|A|: 输出回答的长度

DecoupledRAG：效率分析





VanillaRAG:
在线编码所有文档知识

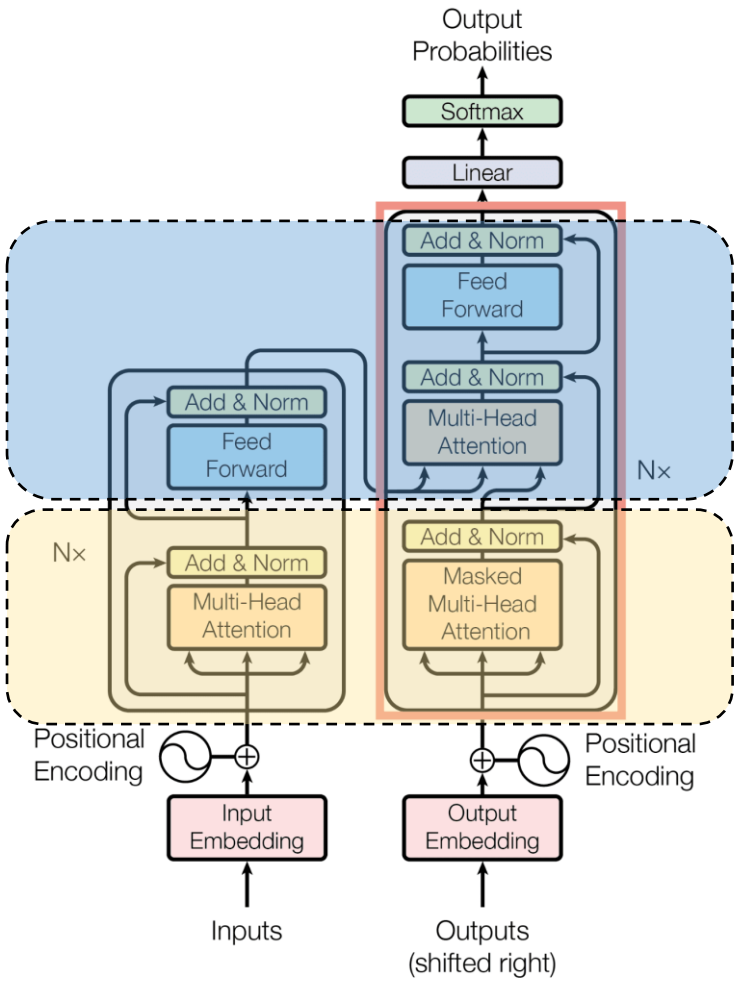
Offline DecoupledRAG:
离线编码文档知识

Online DecoupledRAG:
在线编码文档知识

03

参数化知识注入的检索增强

大模型的知识架构



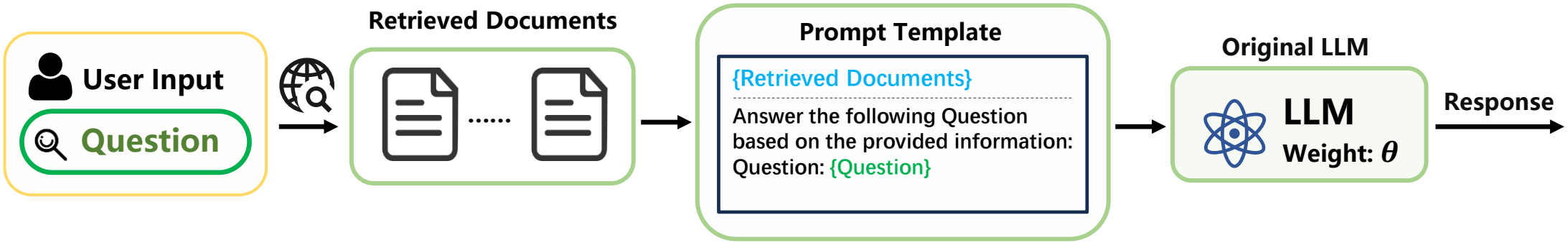
- 前馈神经网络 (FFN)
  - 内部知识存储
  - 推理能力的基础

  [Nanda et al. 2023, Yu and Ananiadou 2024]

- 注意力网络
  - 外部知识编码
  - 动态上下文建模

基于注意力网络的知识注入

- 基于上下文提示词的知识注入
  - 通过提示词输入将检索到的文档注入大模型的注意力网络
  - 几乎所有现有的检索增强生成（RAG）方法都依赖于上下文知识的注入。

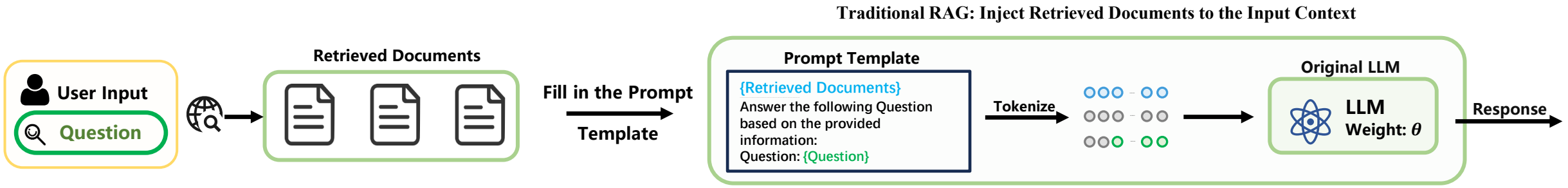


基于注意力网络的知识注入

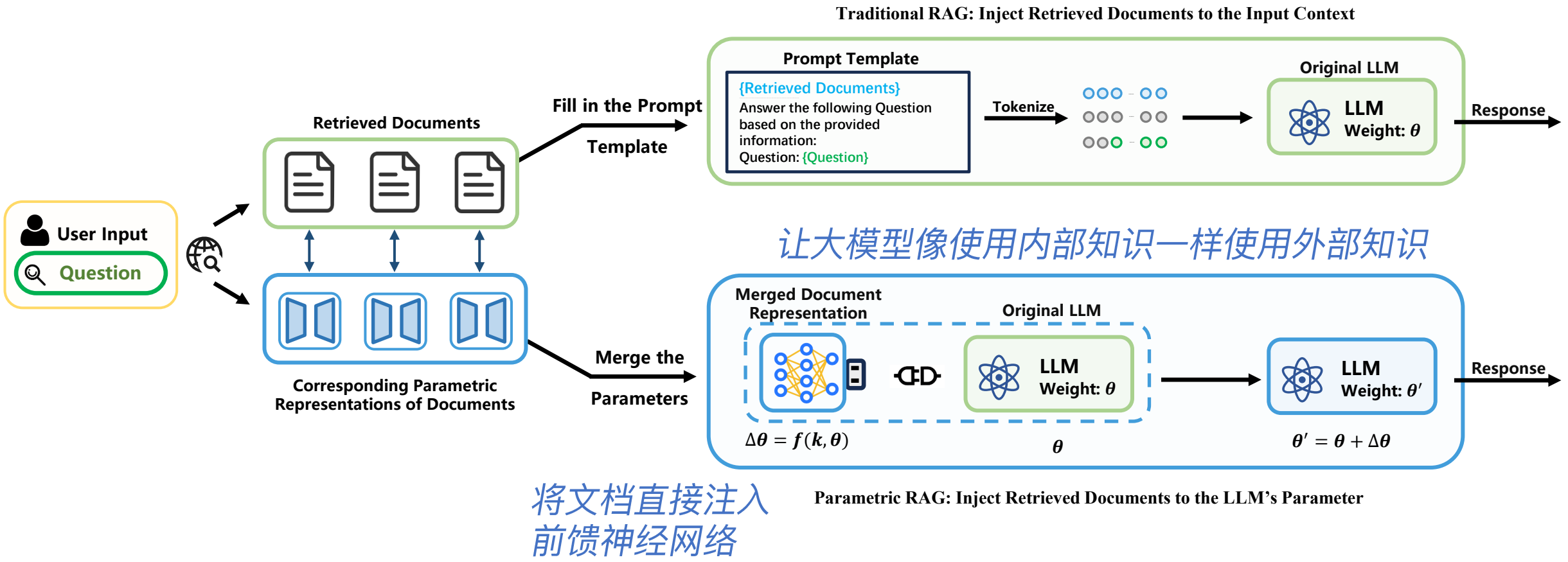
除了提示词输入，是否可以直接将外部知识注入大模型的内部参数？

- 基于上下文提示词知识注入的缺点
  - 模型上下文长度限制
  - 额外的token计算开销
  - 结构性缺陷：
    - 大模型或许永远无法像使用内部知识那样利用外部知识

参数化检索增强生成 Parametric RAG

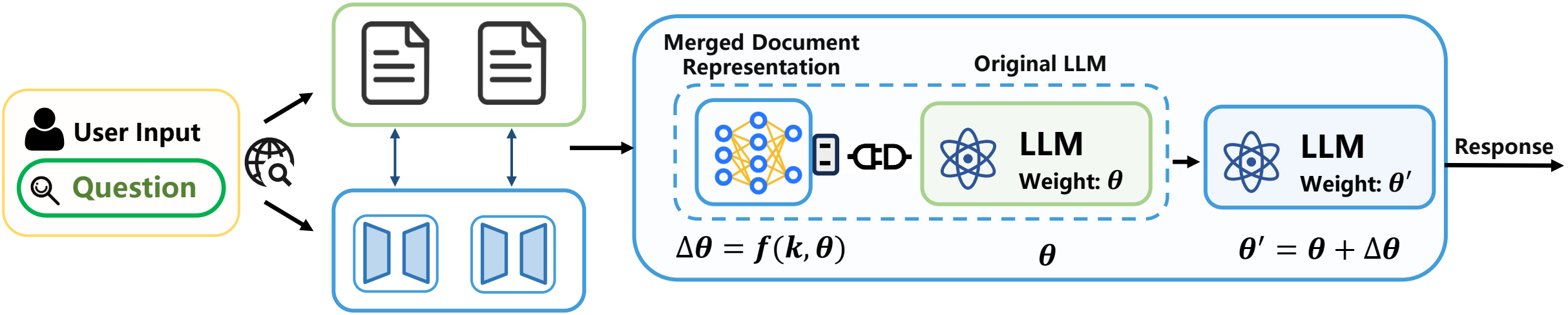


参数化检索增强生成 Parametric RAG



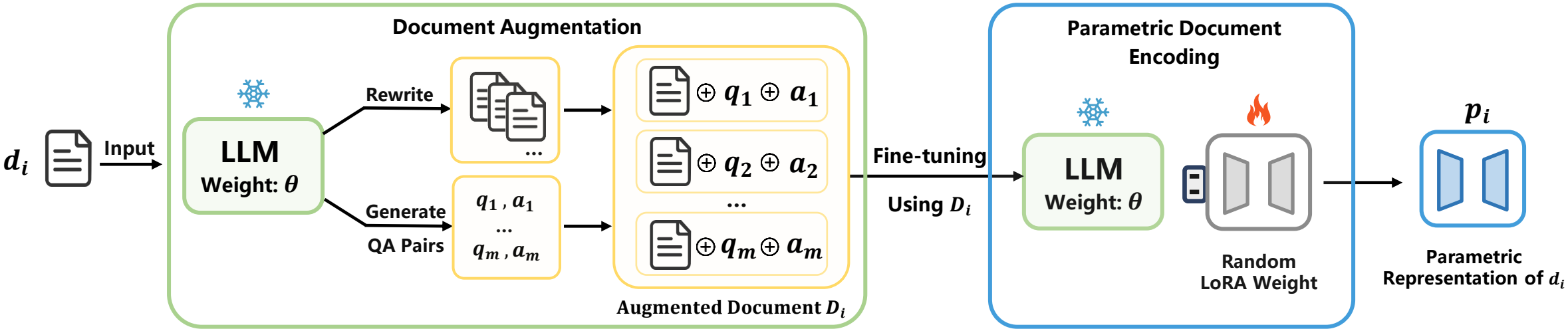
Parametric RAG：核心框架

- 离线文档编码
  - 构建文档的参数化表示
- 线上推理
  - 检索相关文档，合并文档参数，生成最终回复



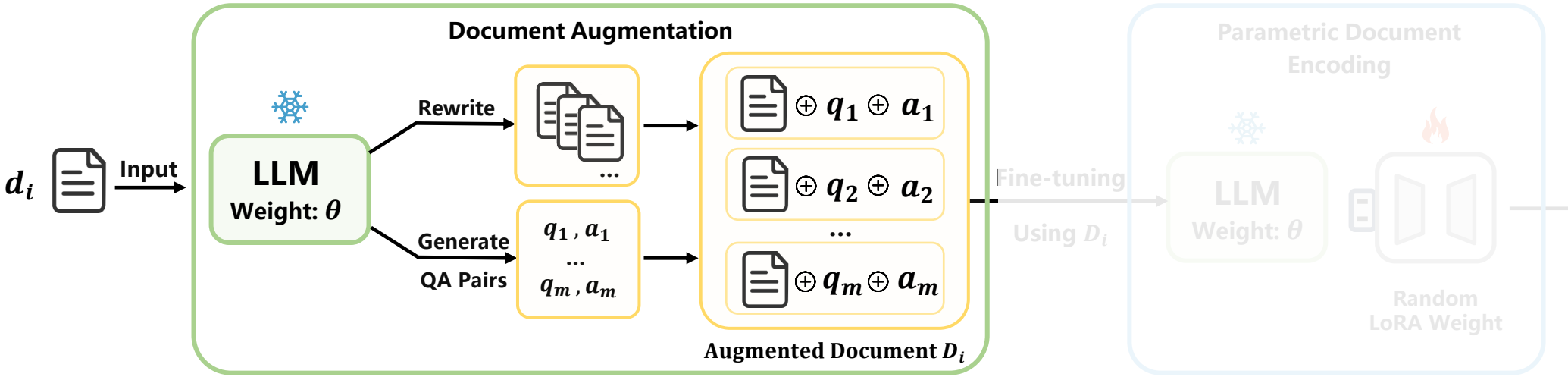
Parametric RAG：离线文档编码

- 两步流程：
  - 文档数据增广
    - 构造文档平行语料和问答样例
  - 文档参数编码
    - 基于增广数据学习文档的LoRA表示



Parametric RAG：离线文档编码

- 文档数据增广
  - 文档重写:
    - 保证事实信息不变的情况下生成平行语料.
  - 问答样例生成:
    - 依据原始文档，利用大模型生成潜在问题和答案样例.

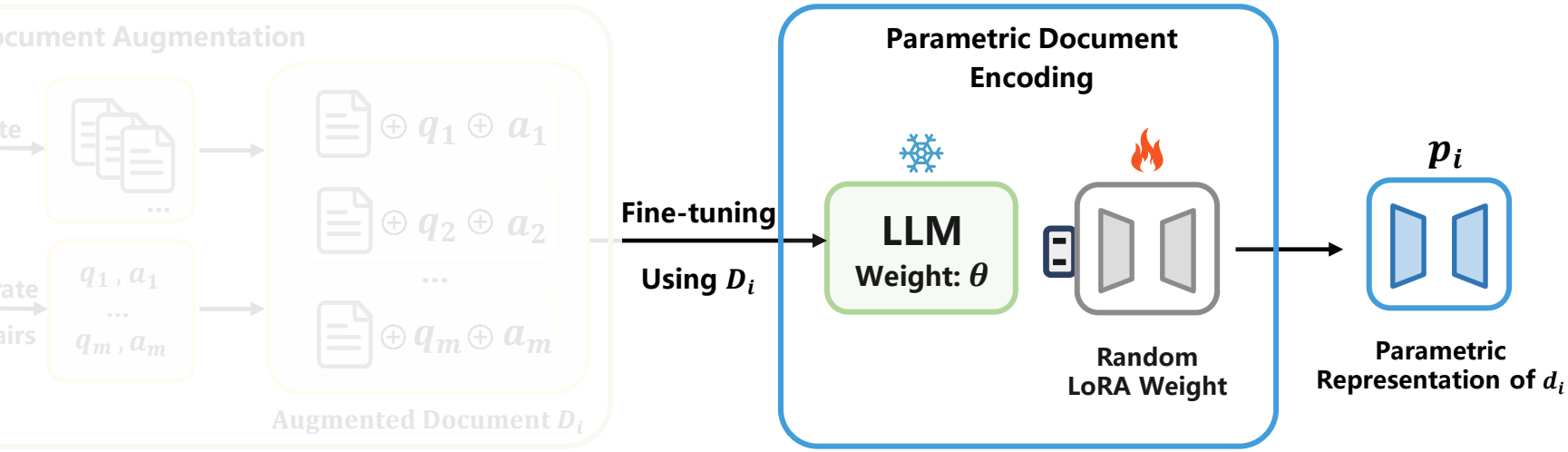


Parametric RAG：离线文档编码

- 文档参数编码
  - 利用可插拔的高效微调模块构建（如LORA ）。

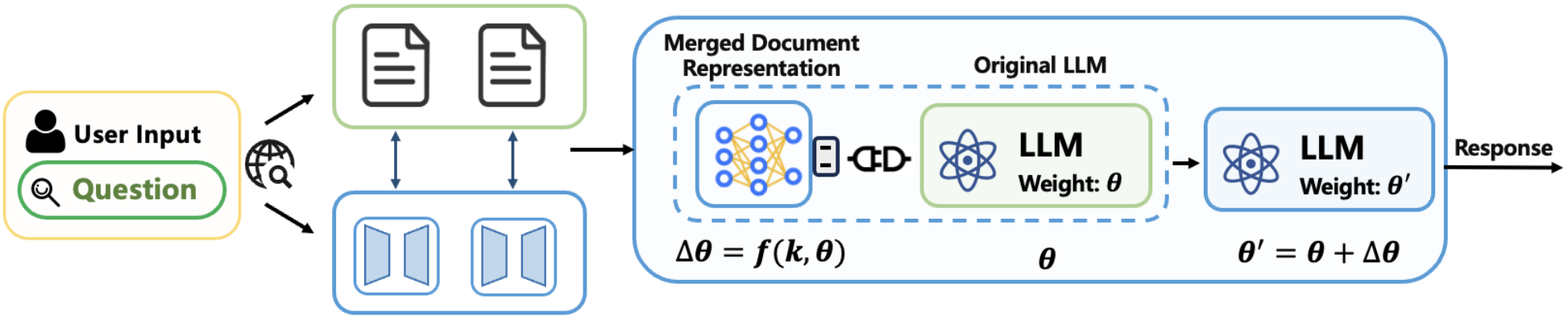
$$W' = W + \Delta W = W + A B^{\top}$$

  - 基于next token prediction loss训练1个epoch.



Parametric RAG：线上推理

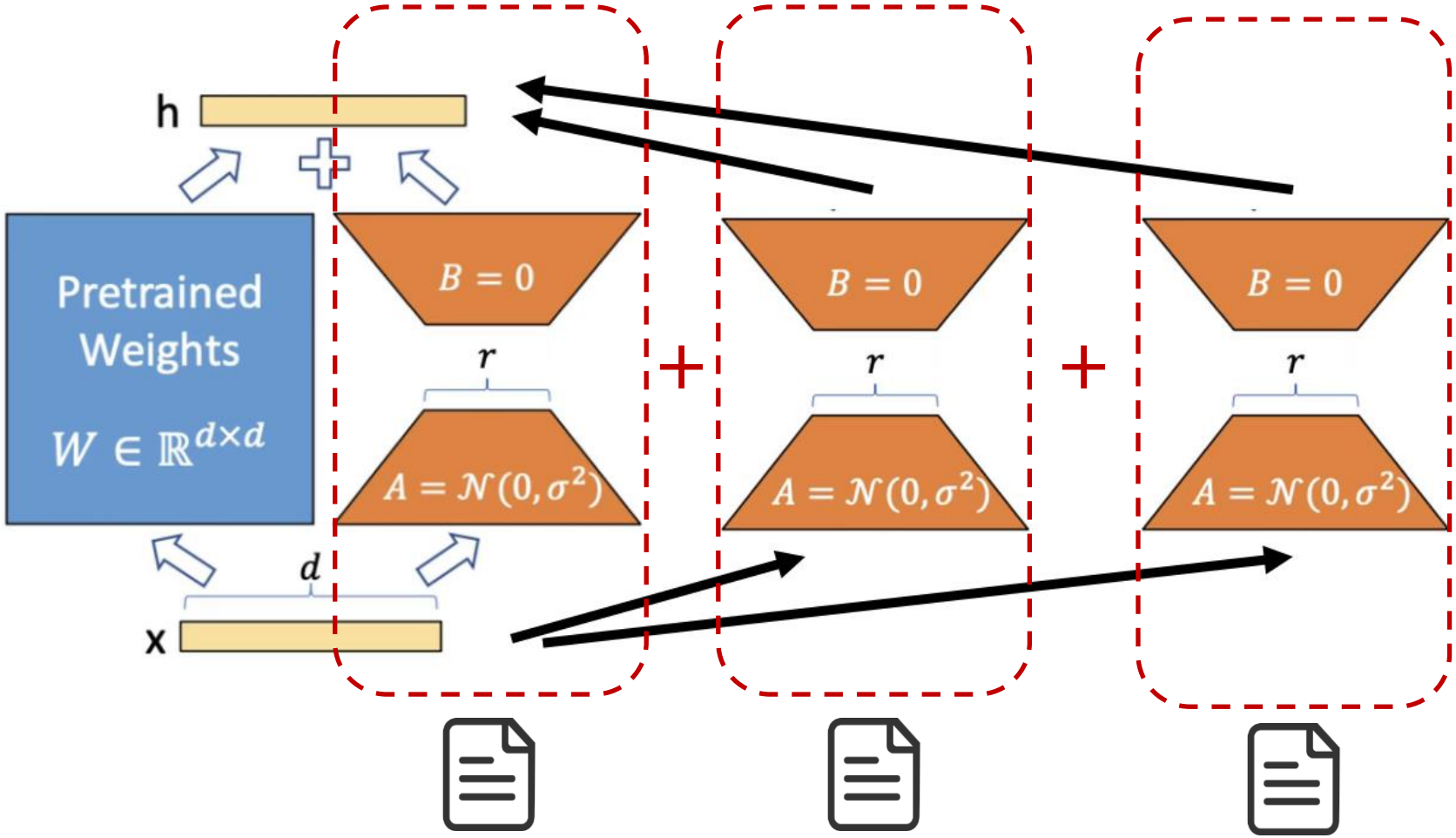
- 三步过程
  - **检索**: 检索top k文档，并获取对应文档的参数化表示
  - **更新**: 合并文档参数，并用其更新大模型参数
  - **生成**: 基于更新后的大模型生成答案



Parametric RAG：线上推理

- 多文档信息聚合

$$\Delta W_{\text{merge}} = \alpha \cdot \sum_{j=1}^{k} A_j B_j^{\top}$$



Parametric RAG：实验设置

- 基线模型
  - 标准 RAG
  - DARAG:标准 RAG + 数据增强模块
  - FLARE [Jiang et al. 2023]
  - DRAGIN [Su et al. 2024]
- 实验数据集
  - 2WikiMultihopQA
  - Hotpot QA
  - Complex Web Questions
  - Pop QA
- 测试模型
  - LLaMA-3-8B-Instruct
  - LLaMA-3.1-1.5B-Instruct
  - Qwen-2.5-1.5B-Instruct

- 实现细节:
  - 文库: Wikipedia
  - 检索模型: BM25, top-3文档

Parametric RAG：实验结果

		2WikiMultihopQA					HotpotQA			PopQA	CWQ
		Compare	Bridge	Inf.	Compose	Total	Bridge	Compare	Total		
LLaMA-1B	Standard RAG	0.4298†*	0.3032†*	0.2263	0.1064	0.2520*	0.2110	0.4083	0.2671	0.1839*	0.3726
	DA-RAG	0.3594†*	0.2587†*	0.2266	0.0869†*	0.2531*	0.1716*	0.3713†*	0.2221	0.2012*	0.3691
	FLARE	0.4013†*	0.2589†*	0.1960	0.0823†*	0.2234*	0.1630*	0.3784†*	0.1785*	0.1301†*	0.3173*
	DRAGIN	0.4556	0.3357*	0.1919	0.0901	0.2692*	0.1431*	0.4015	0.1830*	0.1056†*	0.3900
	P-RAG (Ours)	0.4920	0.3994	0.2185	0.1334	0.2764	0.1602*	**0.4493**	0.1999*	0.2205*	0.3482*
	Combine Both	**0.5046**	**0.4595**	**0.2399**	**0.1357**	**0.3237**	**0.2282**	0.4217	**0.2689**	**0.2961**	**0.4101**
Qwen-1.5B	Standard RAG	0.3875†*	0.3884†*	0.1187†*	0.0568†*	0.2431†*	0.1619*	0.3713†*	0.2073*	0.0999†*	0.2823*
	DA-RAG	0.3418†*	0.4015	0.1269†*	0.0514†*	0.2156†*	0.1182†*	0.3041†*	0.1683*	0.1197†*	0.2718†*
	FLARE	0.1896†*	0.1282†*	0.0852†*	0.0437†*	0.1004†*	0.0750†*	0.1229†*	0.0698†*	0.0641†*	0.1647†*
	DRAGIN	0.2771†*	0.1826†*	0.1025†*	0.0680†*	0.1538†*	0.0801†*	0.1851†*	0.0973†*	0.0548†*	0.1788†*
	P-RAG (Ours)	**0.4529**	**0.4494**	**0.2072**	**0.1372**	**0.3025**	0.1720	0.4623	0.2165*	0.1885	0.3280
	Combine Both	0.4053	0.4420	0.1705	0.1154	0.2627	**0.2383**	**0.5037**	**0.2942**	**0.2261**	**0.3495**
LLaMA-8B	Standard RAG	0.5843†*	0.4794†*	0.1833†*	0.0991†*	0.3372†*	0.1823†*	0.3493†*	0.2277†*	0.1613†*	0.3545†*
	DA-RAG	0.4921†*	0.3344†*	0.1523†*	0.0670†*	0.2396†*	0.1587†*	0.2860†*	0.1996†*	0.2255*	0.3481†*
	FLARE	0.4293†*	0.3769†*	0.3086	0.1627*	0.3492*	0.2493†*	0.4324†*	0.2771†*	0.2393*	0.3084†*
	DRAGIN	0.5185†*	0.4480†*	0.2664	0.1833	0.3544*	0.2618*	0.6116*	0.2924*	0.1772†*	0.3101†*
	P-RAG (Ours)	0.6353	0.5437	0.2471*	0.1992	0.3932	0.3115*	0.6557	0.3563*	0.2413*	0.4541
	Combine Both	**0.6432**	**0.5556**	**0.3160**	**0.2339**	**0.4258**	**0.4025**	**0.6918**	**0.4559**	**0.3059**	**0.4728**

Parametric RAG：实验结果

		2WikiMultihopQA					HotpotQA			PopQA	CWQ
		Compare	Bridge	Inf.	Compose	Total	Bridge	Compare	Total		
LLaMA-1B	Standard RAG	0.4298†*	0.3032†*	0.2263	0.1064	0.2520*	0.2110	0.4083	0.2671	0.1839*	0.3726
	DA-RAG	0.3594†*	0.2587†*	0.2266	0.0869†*	0.2531*	0.1716*	0.3713†*	0.2221	0.2012*	0.3691
	FLARE	0.4013†*	0.2589†*	0.1960	0.0823†*	0.2234*	0.1630*	0.3784†*	0.1785*	0.1301†*	0.3173*
	DRAGIN	0.4556	0.3357*	0.1919	0.0901	0.2692*	0.1431*	0.4015	0.1830*	0.1056†*	0.3900
	P-RAG (Ours)	0.4920	0.3994	0.2185	0.1334	0.2764	0.1602*	**0.4493**	0.1999*	0.2205*	0.3482*
	Combine Both	**0.5046**	**0.4595**	**0.2399**	**0.1357**	**0.3237**	**0.2282**	0.4217	**0.2689**	**0.2961**	**0.4101**
Qwen-1.5B	Standard RAG	0.3875†*	0.3884†*	0.1187							.2823*
	DA-RAG	0.3418†*	0.4015	0.1269							.2718†*
	FLARE	0.1896†*	0.1282†*	0.0852							.1647†*
	DRAGIN	0.2771†*	0.1826†*	0.1025							.1788†*
	P-RAG (Ours)	**0.4529**	**0.4494**	**0.207**							.3280
	Combine Both	0.4053	0.4420	0.1705							**.3495**
LLaMA-8B	Standard RAG	0.5843†*	0.4794†*	0.1833†*	0.0991†*	0.3372†*	0.1823†*	0.3493†*	0.2277†*	0.1613†*	0.3545†*
	DA-RAG	0.4921†*	0.3344†*	0.1523†*	0.0670†*	0.2396†*	0.1587†*	0.2860†*	0.1996†*	0.2255*	0.3481†*
	FLARE	0.4293†*	0.3769†*	0.3086	0.1627*	0.3492*	0.2493†*	0.4324†*	0.2771†*	0.2393*	0.3084†*
	DRAGIN	0.5185†*	0.4480†*	0.2664	0.1833	0.3544*	0.2618*	0.6116*	0.2924*	0.1772†*	0.3101†*
	P-RAG (Ours)	0.6353	0.5437	0.2471*	0.1992	0.3932	0.3115*	0.6557	0.3563*	0.2413*	0.4541
	Combine Both	**0.6432**	**0.5556**	**0.3160**	**0.2339**	**0.4258**	**0.4025**	**0.6918**	**0.4559**	**0.3059**	**0.4728**

优势来源于**参数化编码**，而不是简单的数据增强

Parametric RAG：实验结果

		2WikiMultihopQA					HotpotQA			PopQA	CWQ
		Compare	Bridge	Inf.	Compose	Total	Bridge	Compare	Total		
LLaMA-1B	Standard RAG	0.4298†*	0.3032†*	0.2263	0.1064	0.2520*	0.2110	0.4083	0.2671	0.1839*	0.3726
	DA-RAG	0.3594†*	0.2587†*	0.2266	0.0869†*	0.2531*	0.1716*	0.3713†*	0.2221	0.2012*	0.3691
	FLARE	0.4013†*	0.2589†*	0.1960	0.0823†*	0.2234*	0.1630*	0.3784†*	0.1785*	0.1301†*	0.3173*
	DRAGIN	0.4556	0.3357*	0.1919	0.0901	0.2692*	0.1431*	0.4015	0.1830*	0.1056†*	0.3900
	P-RAG (Ours)	0.4920	0.3994	0.2185	0.1334	0.2764	0.1602*	**0.4493**	0.1999*	0.2205*	0.3482*
					.357	**0.3237**	**0.2282**	0.4217	**0.2689**	**0.2961**	**0.4101**
					.568†*	0.2431†*	0.1619*	0.3713†*	0.2073*	0.0999†*	0.2823*
					.514†*	0.2156†*	0.1182†*	0.3041†*	0.1683*	0.1197†*	0.2718†*
					.437†*	0.1004†*	0.0750†*	0.1229†*	0.0698†*	0.0641†*	0.1647†*
					.680†*	0.1538†*	0.0801†*	0.1851†*	0.0973†*	0.0548†*	0.1788†*
					.372	**0.3025**	0.1720	0.4623	0.2165*	0.1885	0.3280
	Combine Both	0.4053	0.4420	0.1705	0.1154	0.2627	**0.2383**	**0.5037**	**0.2942**	**0.2261**	**0.3495**
LLaMA-8B	Standard RAG	0.5843†*	0.4794†*	0.1833†*	0.0991†*	0.3372†*	0.1823†*	0.3493†*	0.2277†*	0.1613†*	0.3545†*
	DA-RAG	0.4921†*	0.3344†*	0.1523†*	0.0670†*	0.2396†*	0.1587†*	0.2860†*	0.1996†*	0.2255*	0.3481†*
	FLARE	0.4293†*	0.3769†*	0.3086	0.1627*	0.3492*	0.2493†*	0.4324†*	0.2771†*	0.2393*	0.3084†*
	DRAGIN	0.5185†*	0.4480†*	0.2664	0.1833	0.3544*	0.2618*	0.6116*	0.2924*	0.1772†*	0.3101†*
	P-RAG (Ours)	0.6353	0.5437	0.2471*	0.1992	0.3932	0.3115*	0.6557	0.3563*	0.2413*	0.4541
	Combine Both	**0.6432**	**0.5556**	**0.3160**	**0.2339**	**0.4258**	**0.4025**	**0.6918**	**0.4559**	**0.3059**	**0.4728**

参数化方式与上下文方式能力互补，
结合后性能更佳

Parametric RAG：效率分析

- 理论复杂度分析
  - **离线编码成本:**
    - 数据增强与LoRA模块训练复杂度:

$$O(|d|^2 h + |d|h)$$

    - 约等效于解码 $12|d|$ token的成本.
  - **在线编码成本:**
    - Parametric RAG推理复杂度:

$$O(|q|^2 h + |q|h^2)$$

    - 没有上下文输入成本，效率显著高于传统RAG方案

|D|: 文档长度
|Q|: 用户输入长度
h: 模型隐藏层大小

Parametric RAG：效率分析

- 实验复杂度分析
  - PRAG有效降低推理时延: 相比传统RAG提速29%—36%
  - 融合传统方法和参数化方法后:
    - 效率与传统RAG相当
    - 任务性能显著提升

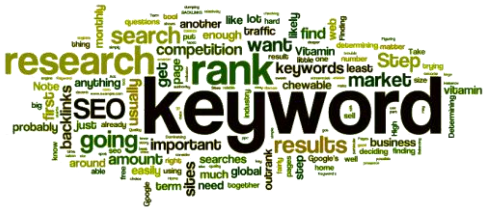
	2WQA		CWQ	
	Time(s)	Speed Up	Time(s)	Speed Up
P-RAG	$2.34_{+0.32}$	1.29x	$2.07_{+0.32}$	1.36x
Combine Both	$3.08_{+0.32}$	0.98x	$2.84_{+0.32}$	0.99x
Standard RAG	3.03	1.00x	2.82	1.00x
FLARE	10.14	0.25x	11.31	0.25x
DRAGIN	14.60	0.21x	16.21	0.17x

Inference time comparison of LLaMA3-8B across different baselines, measuring online latency per question.

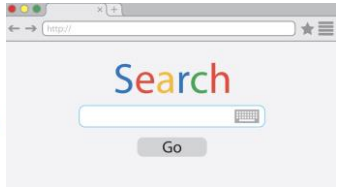
04

未来展望

未来展望



信息管理时代



互联网时代

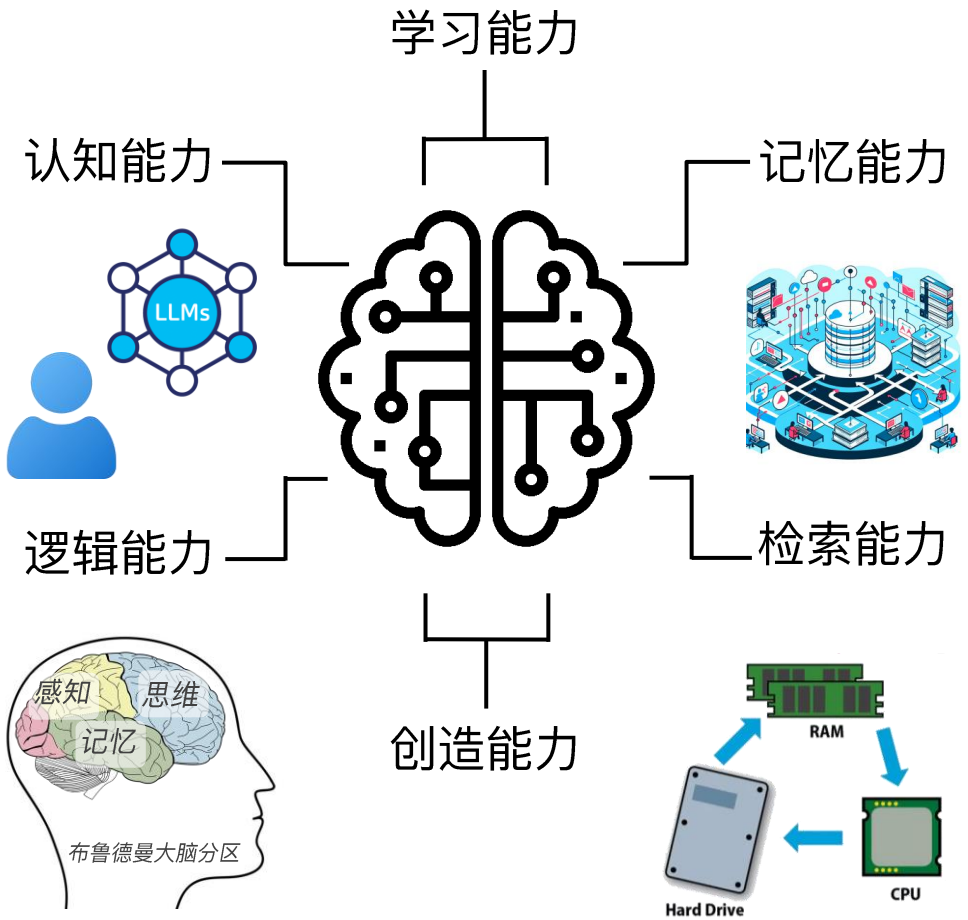
?

（生成式）人工智能时代

时代塑造着信息检索，
信息检索也塑造着时代

1960 1990 2020

未来展望



思维同步

打通知识结构交流壁垒

- 检索模型与生成模型的参数共享
- 参数化信息检索增强生成
- 基于检索需求的生成模型输出对齐
- …

构建高效可持续学习框架

- 基于参数与非参优化的知识编辑技术
- 考虑信息数据与知识结构特性的学习路由
- 用户反馈的收集、过滤、利用机制
- …

提升动态分析与规划能力

- 实时模型信息需求识别
- 分阶段的信息任务拆解
- 基于实时检索结果的思维链重构
- …

行为协同

通用智能体系结构设计

- 多模块协同架构设计
- 检索与生成性能联合优化
- 复杂用户意图拆解与模块对齐
- 动态性能与效率平衡
- …

自动化定制与应用适配

- 自动化定制应用评价体系
- 基于任务描述的工作流规划
- 利用用户反馈的提示词改写
- 基于非参优化的系统持续迭代
- …

InfoQ 极客传媒

北京

QCon 全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

北京

AiCon 全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

上海

QCon 全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月 — 5月 — 6月 — 8月 — 10月 — 12月

上海

AiCon 全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

深圳

AiCon 全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

北京

AiCon 全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

QCon
全球软件开发大会

InfoQ 极客传媒

THANKS

大模型正在重新定义信息检索的发展
而信息检索也在重新塑造大模型的未来

Website: www.thuir.cn
Twitter: @thuir
Linkedin: THUIR