QCon
全球软件开发大会

InfoQ 极客传媒

突破参数知识边界：自增强优化的检索增强大模型技术

庞 亮

中国科学院计算技术研究所 副研究员

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

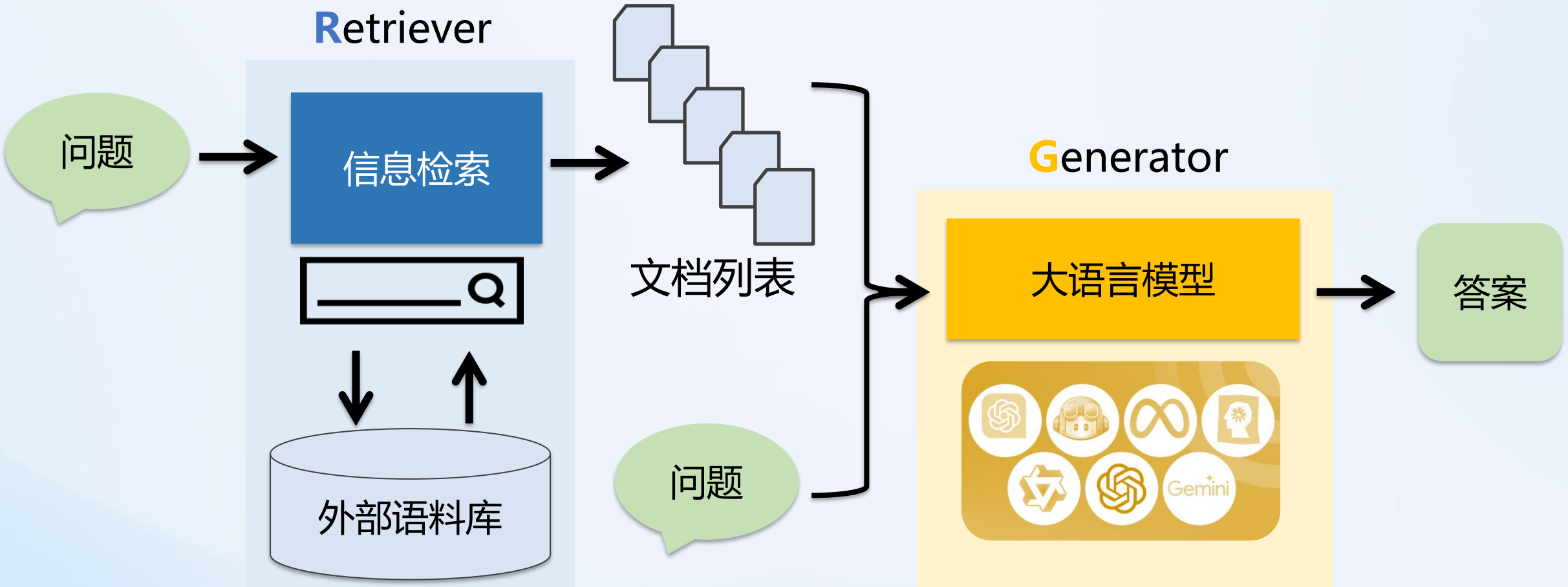
参会咨询

查看会议

01

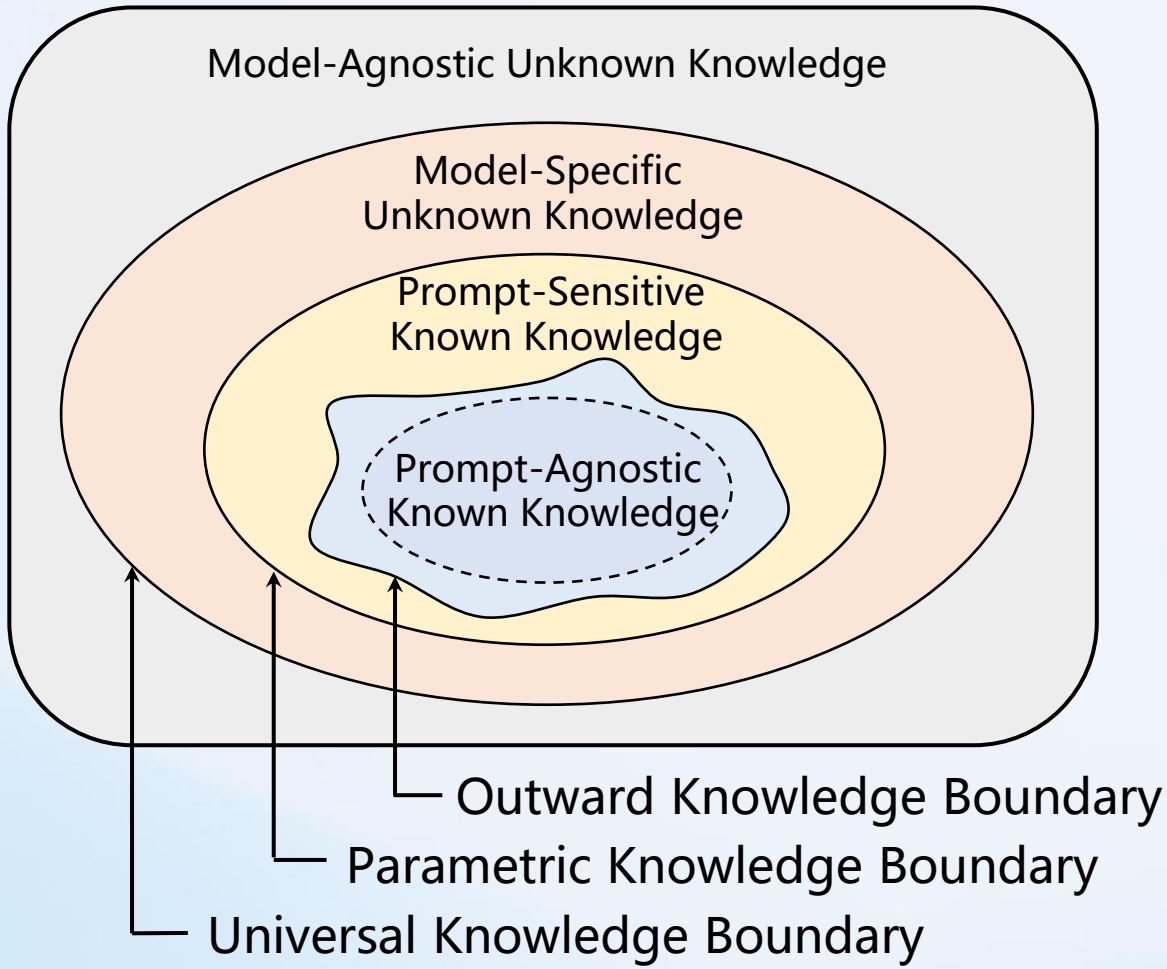
RAG 的技术背景与核心挑战

What：检索增强大模型



经典的检索增强生成的流程

知识边界的定义



三种知识边界

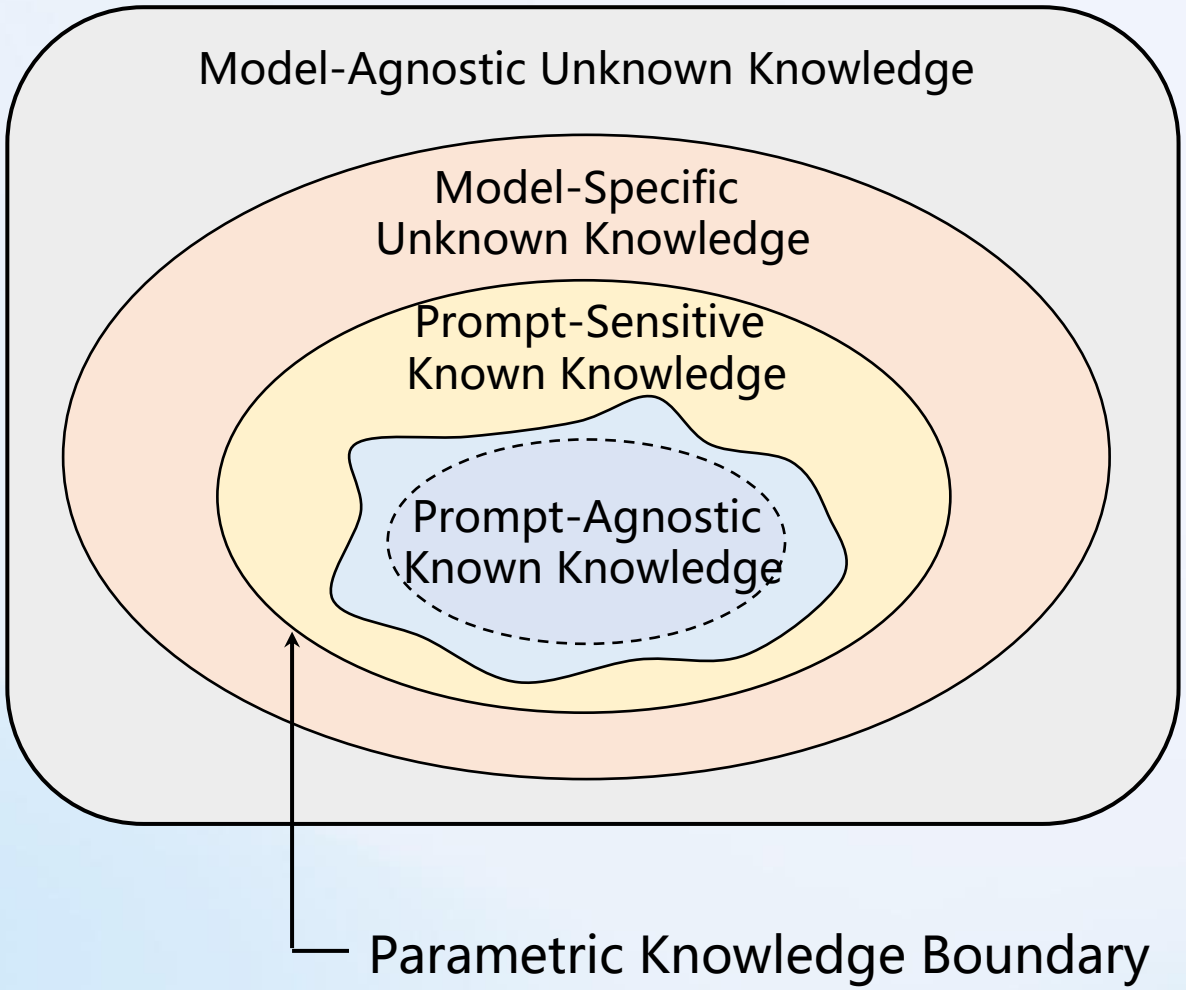
- 外部知识边界（Outward Knowledge Boundary）
- 参数化知识边界（Parametric Knowledge Boundary）
- 通用知识边界（Universal Knowledge Boundary）

四种类型的知识

- 提示无关的已知知识 Prompt-Agnostic Known Knowledge (PAK)
- 提示敏感的已知知识 Prompt-Sensitive Known Knowledge (PSK)
- 模型特定的未知知识 Model-Specific Unknown Knowledge (MSU)
- 模型无关的未知知识 Model-Agnostic Unknown Knowledge (MAU)

Knowledge Boundary of Large Language Models: A Survey, ACL 2025
Unveiling knowledge boundary of large language models for trustworthy information access, SIGIR 2025

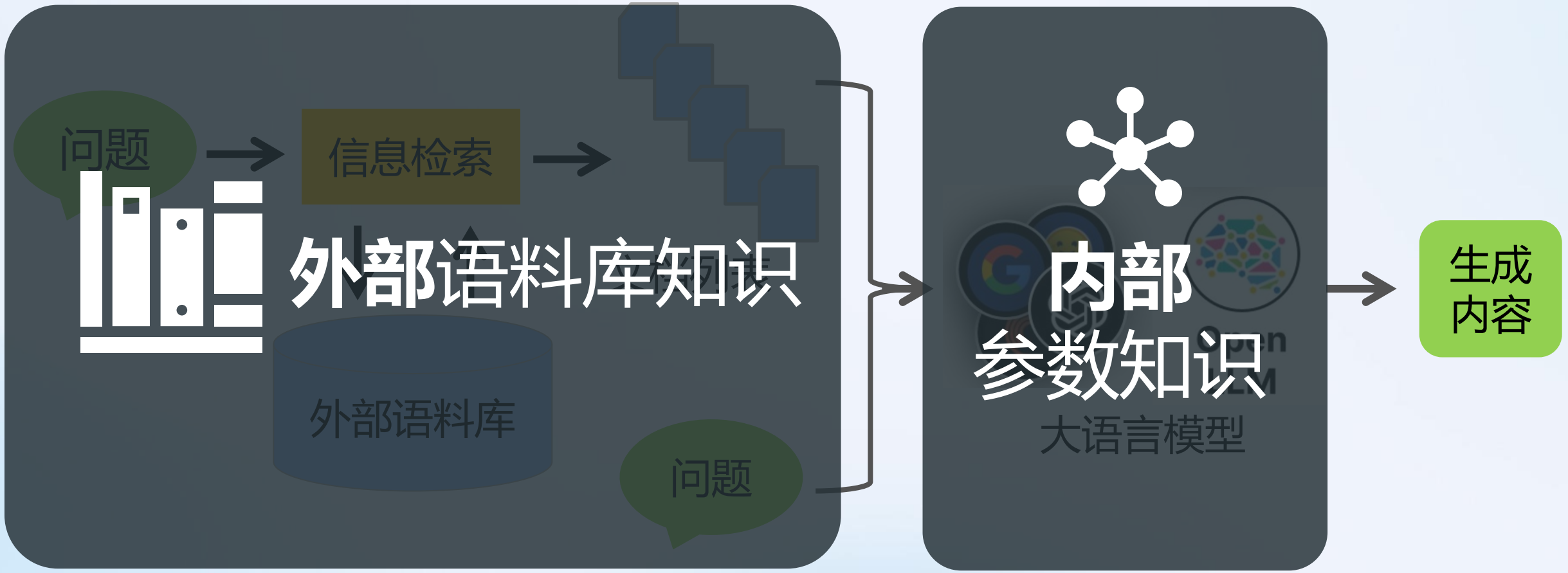
参数化知识边界（Parametric Knowledge Boundary）



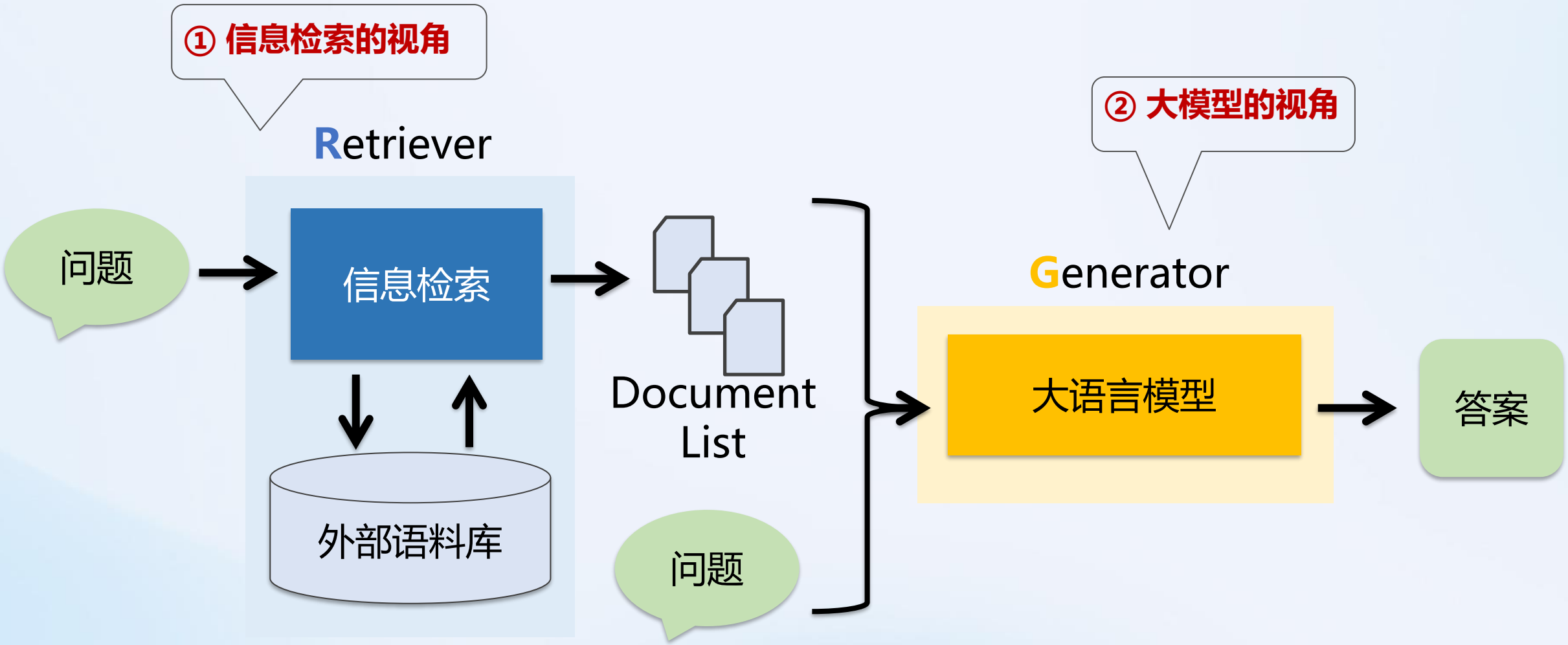
- 参数化知识边界，定义了某个特定 LLM 的抽象知识边界.
- 位于该边界内的知识是指：其已被模型参数所掌握，并且能够通过Q_k中至少一个表达式进行验证的知识.

Why：知识融合的角度

利用信息检索连接外部知识库和大语言模型的参数内知识



How：检索增强大模型的研究路径



02

在信息检索视角下构建用户反馈优化的 RAG 实例

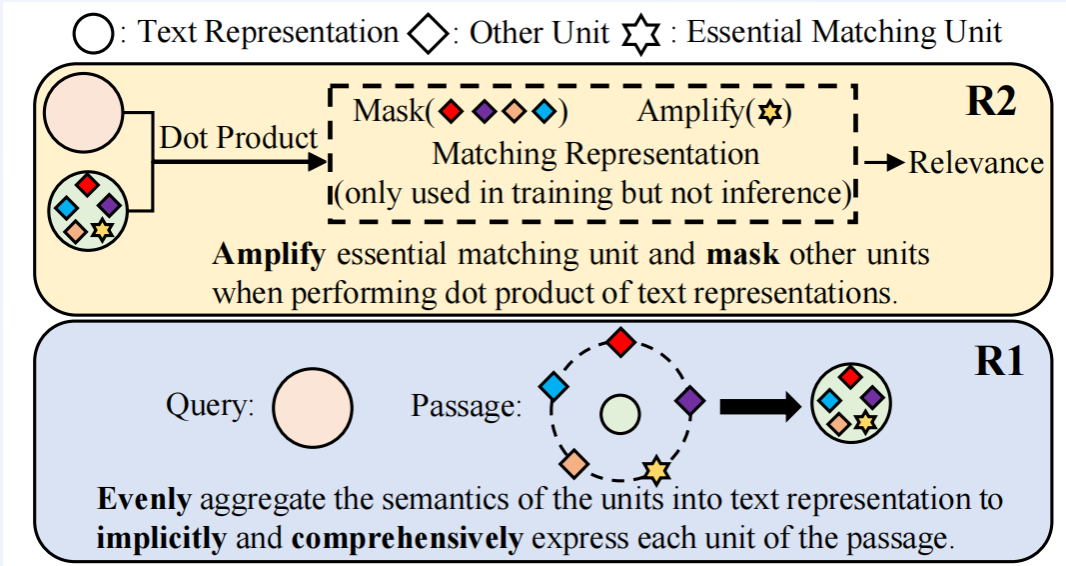
How 1：检索视角下的检索增强大模型

传统的信息检索模型针对人类用户优化，那什么是适合大语言模型的信息检索模型?

需求①：任务泛化性

应用任务纷繁复杂

Constraint in Text Rep. for Dense Retrieval

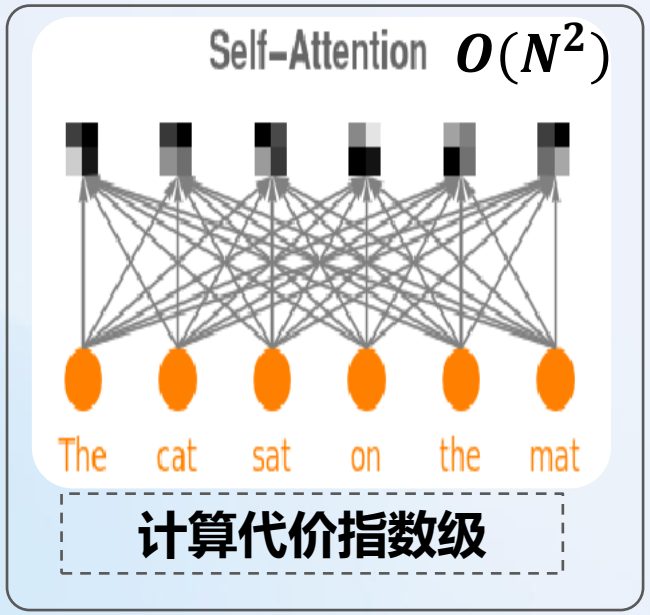


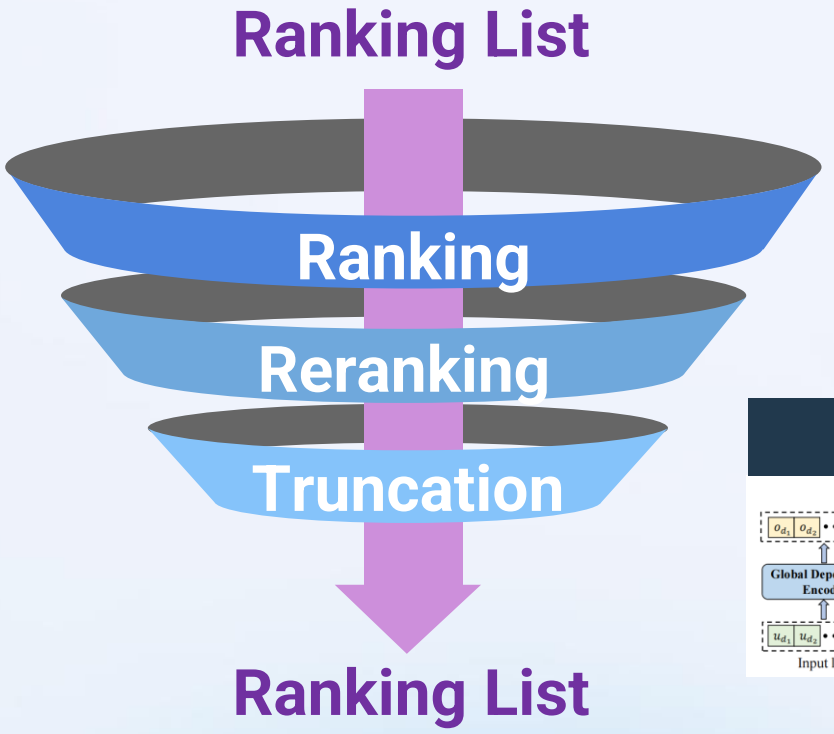
- **Constraint 1: Semantic Unit Balance**
- **Constraint 2: Essential Matching Unit Extractability**

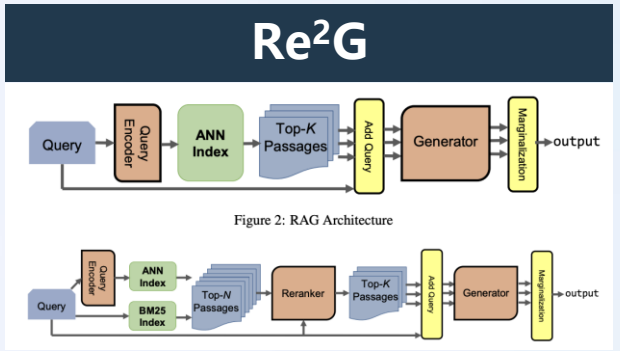
How 1：检索视角下的检索增强大模型

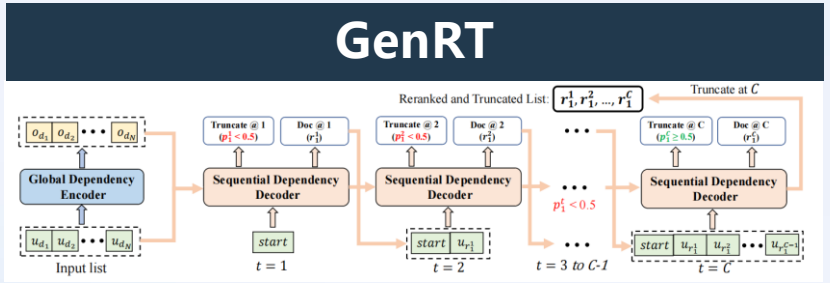
传统的信息检索模型针对人类用户优化，那什么是适合大语言模型的信息检索模型?

需求②： 信息精准性

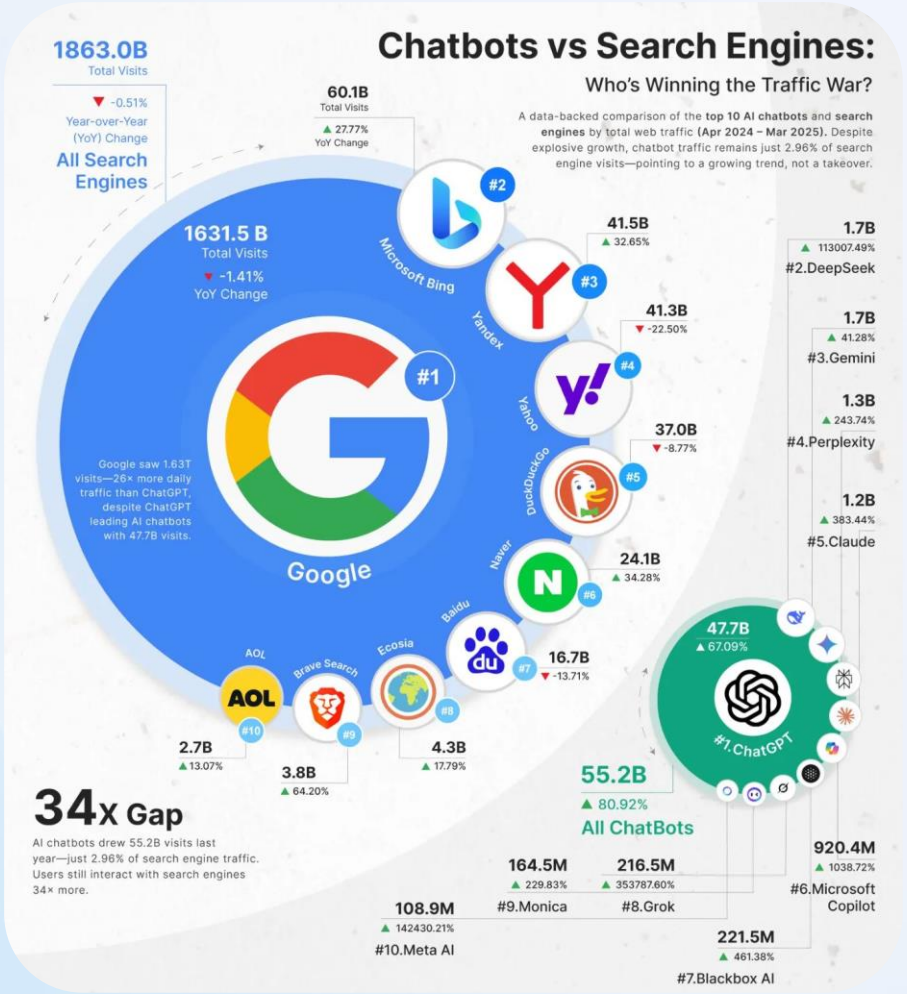


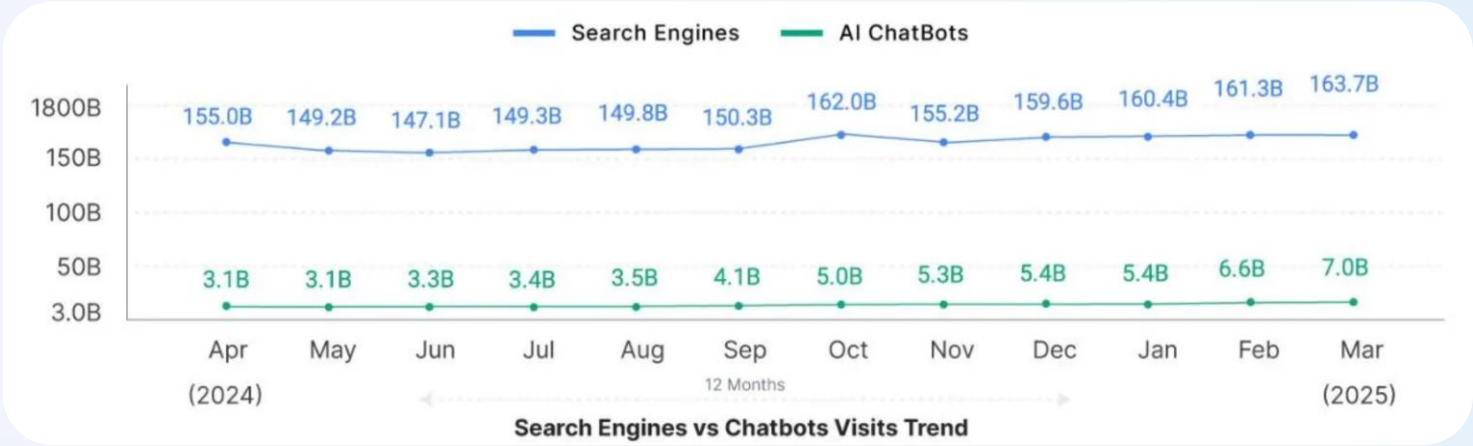






但流量入口的战争仍旧劣势





2024 年，人工智能聊天机器人仅占据 2.96% 的搜索流量。

为什么生成式人工智能搜索引擎尚未达到传统搜索引擎的规模与成就?

[1] Image source: OneLittleWeb

检索增强生成的数据飞轮没有转动

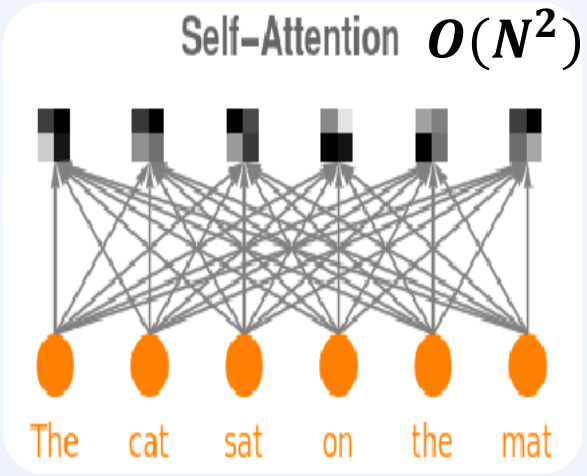
模型大

需求①：任务泛化性

应用任务纷繁复杂

算力大

需求②： 信息精准性



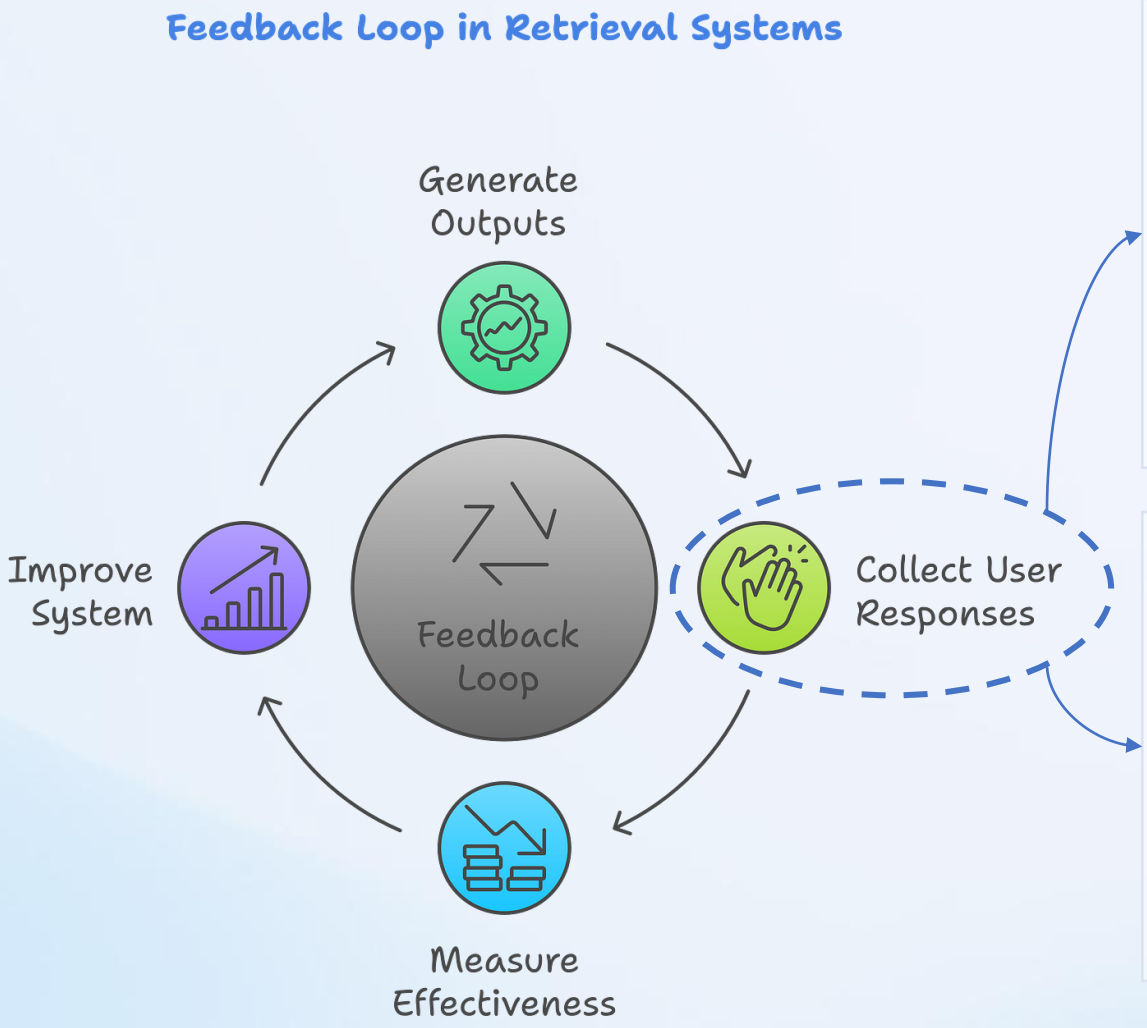
计算代价指数级

数据大?

需求③：持续可优化

模型反馈信号缺乏

传统信息检索的数据飞轮（Data Flywheel）



搜索引擎（Search Engine）

反馈信号：停留、点击、转化

微软/Bing 需要一个强大的**搜索飞轮**，更多查询→更好的模型和更好的信号（点击量、满意度、下游转化率）→更好的结果→更多查询

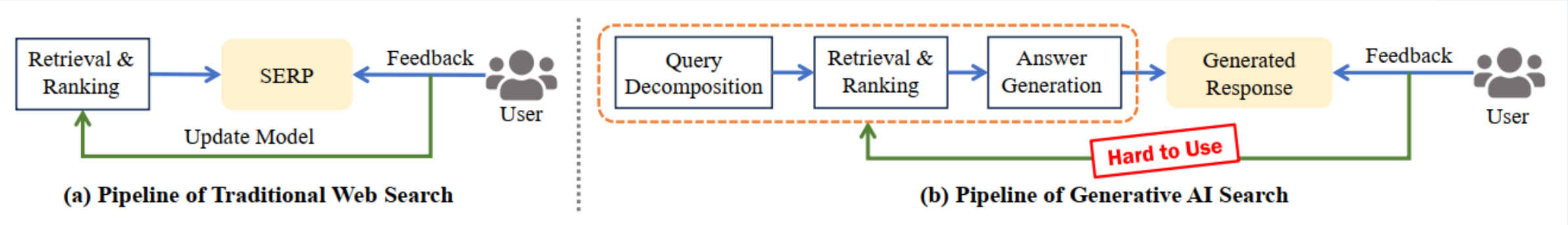
推荐系统（Recommender System）

反馈信号：停留、点击、转化、完播、点赞

TikTok的算法会**采集1000多个信号**，快速的信号积累（短视频、大量互动）可带来非常快速的模型更新和强大的**个性化循环**，从而快速提升参与度

检索增强生成的数据飞轮卡在哪里?

如何为生成式 AI 搜索重建用户反馈生态系统?



优势:

直接答案: 减少手动页面跳转操作

端到端解决方案: 一步完成复杂任务处理

劣势:

稀疏的用户反馈: 数据飞轮会受影响

更难的错误归因: 无法明确是哪个流水线环节出现了问题

NExT-Search: Rebuilding User Feedback Ecosystem for Generative AI Search, SIGIR 2025, Positional Paper

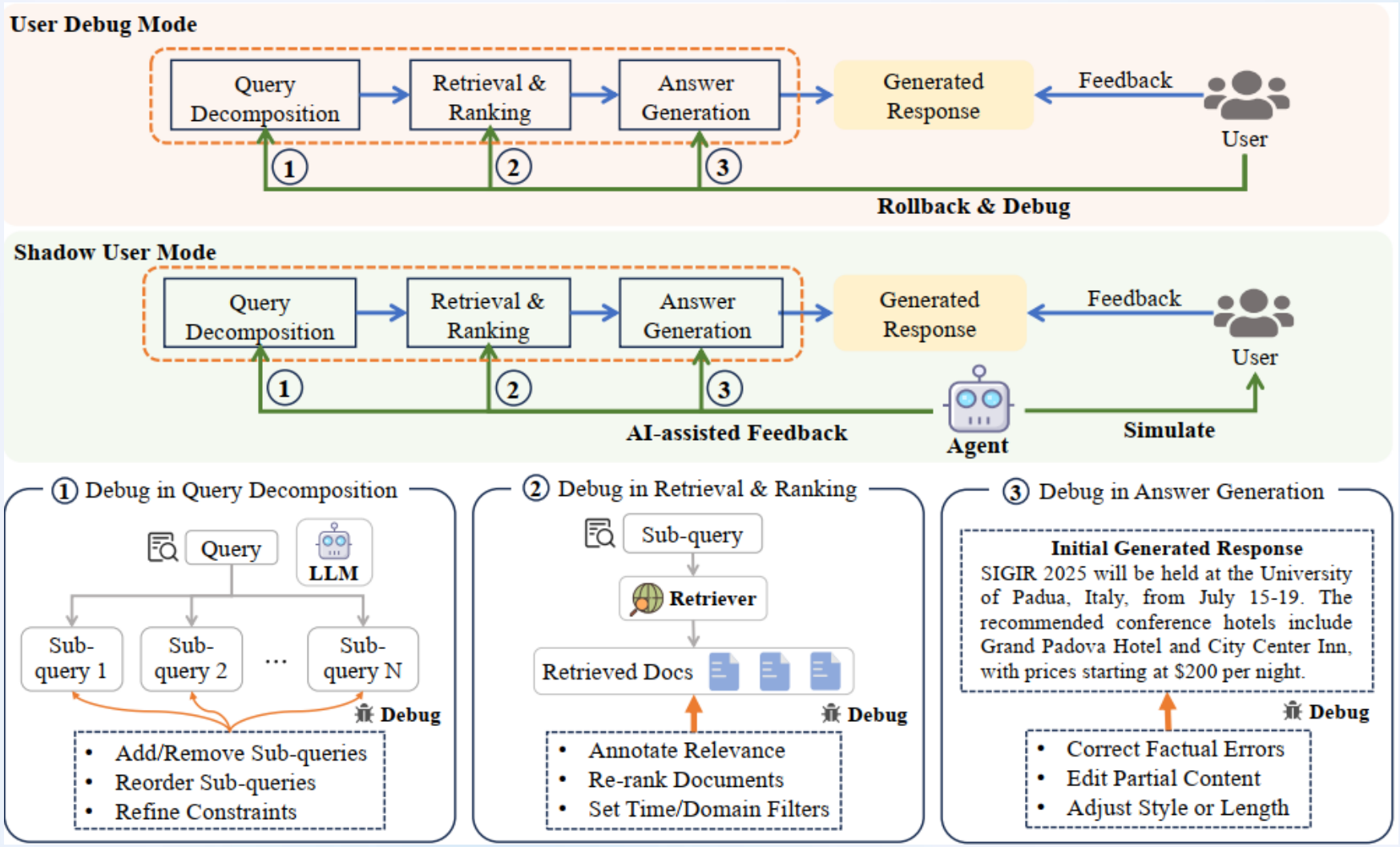
NExT-Search：方法

主动式：用户调试模式 (User Debug Mode)

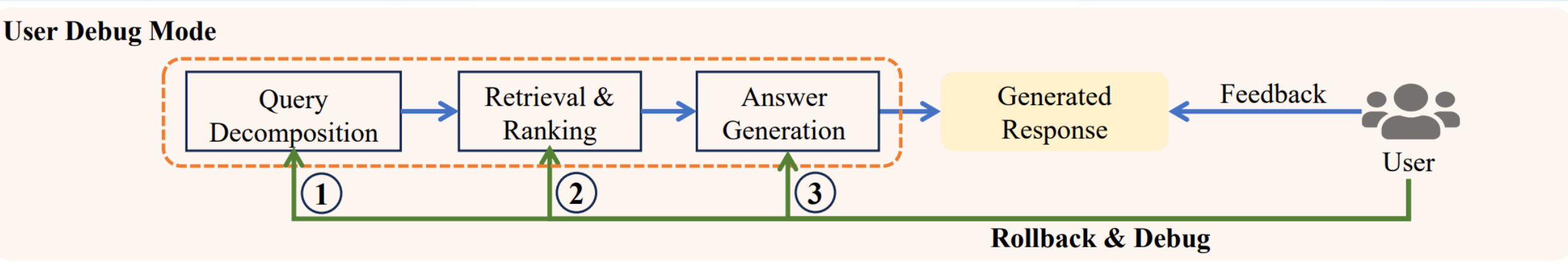
- 积极引导用户参与多阶段反馈与干预。

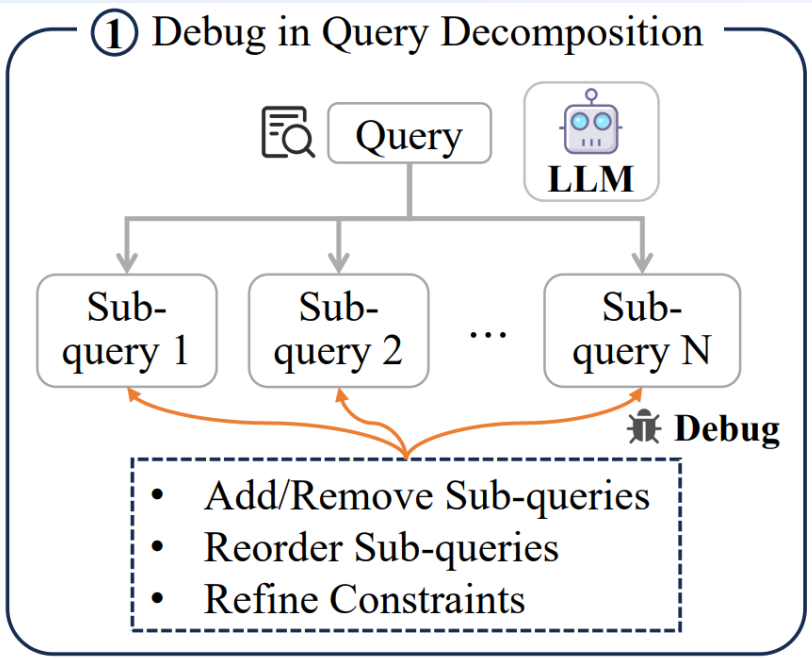
被动式：影子用户模式 (Shadow User Mode)

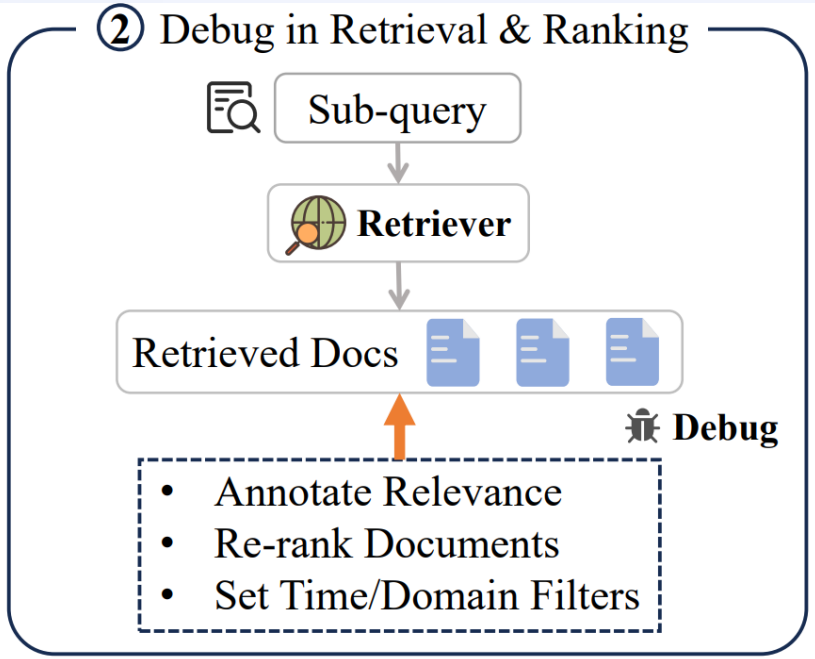
- 利用个性化智能体模拟用户偏好并提供反馈。

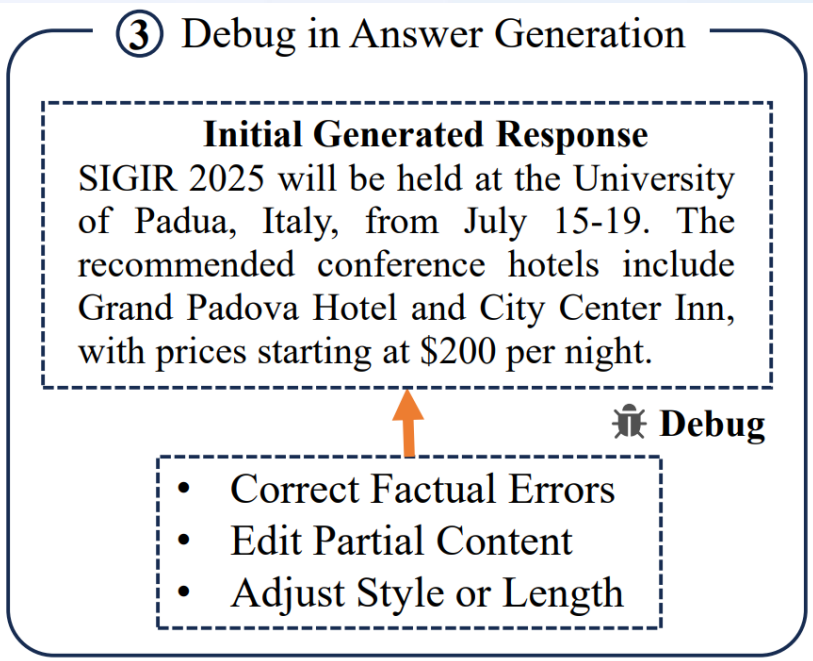


NExT-Search：用户调试模式（User Debug Mode）

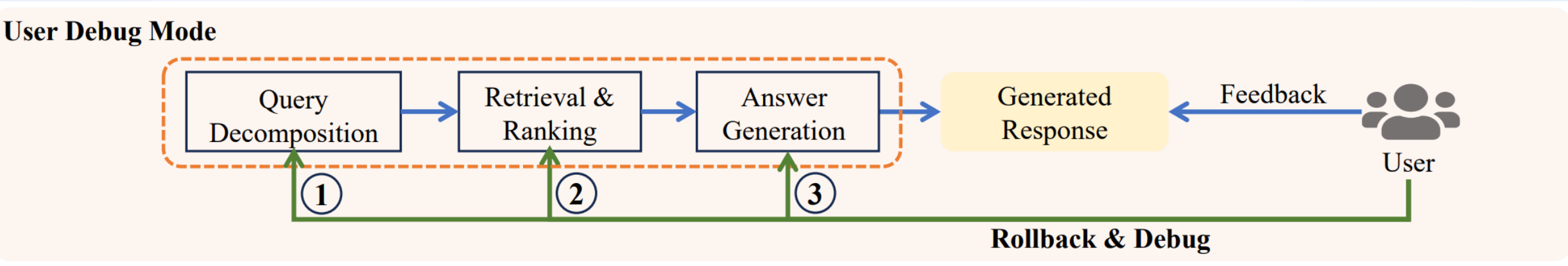




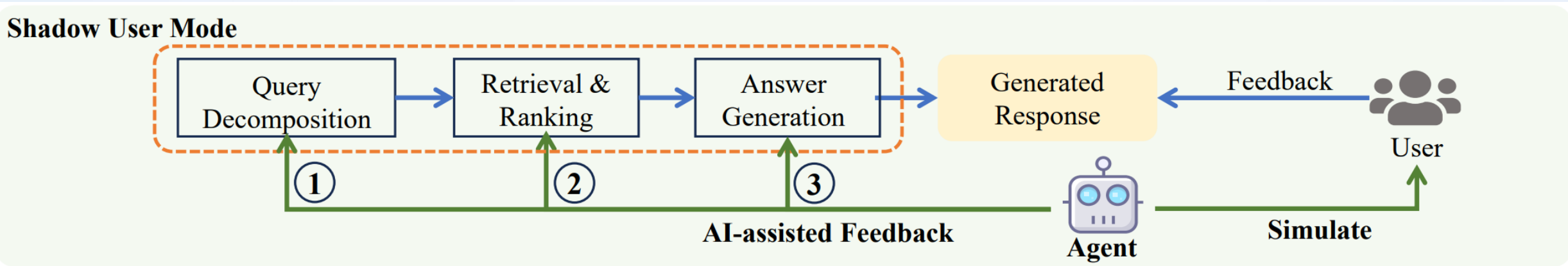




NExT-Search：影子用户模式（Shadow User Mode）



若用户偏好最少化交互，该如何应对？

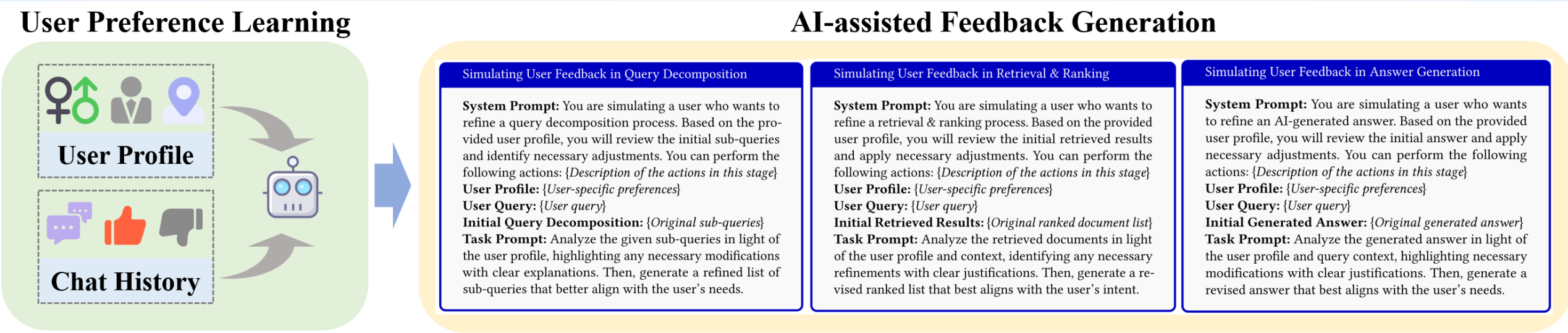


采用个性化用户智能体模拟用户行为，并生成 AI 辅助式反馈

NExT-Search： 个性化用户智能体

智能体的作用是什么？

- 从过往交互中学习用户偏好
- 在用户未干预时生成 AI 辅助式反馈



每位用户对应一个智能体，其会随用户偏好持续进化

NExT-Search：双反馈模式的协同作用

双模式，互补角色

👨‍🔧 用户调试模式（User Debug Mode）

- 精细化、高质量的人工反馈
- 用户在子查询、检索及回答阶段介入
- 支持完全掌控，但需付出更高的操作成本

🕵️ 影子用户模式（Shadow User Mode）

- 通过用户模拟实现的 AI 辅助式反馈
- 从过往行为中学习，降低交互负担
- 借助可信赖的智能体逐步替代人工介入

协同作用与影响

- 无论是人工交互还是模拟交互，每个会话都会产生反馈
- 调试模式 = 高价值但稀疏的信号，影子模式 = 可规模化的信号

NExT-Search： 激励用户参与

挑战

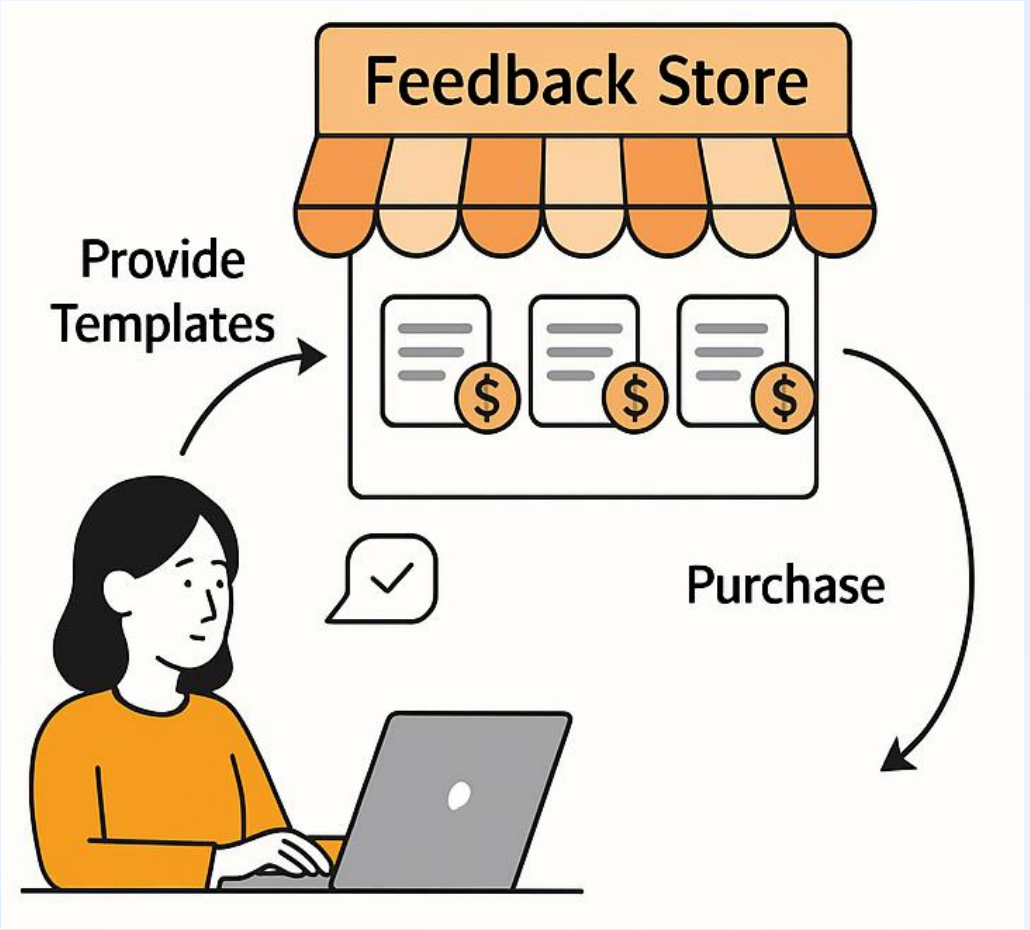
- 用户深度参与调试耗时费力，激励机制至关重要。

反馈商店（Feedback Store）

一个供用户进行以下操作的平台：

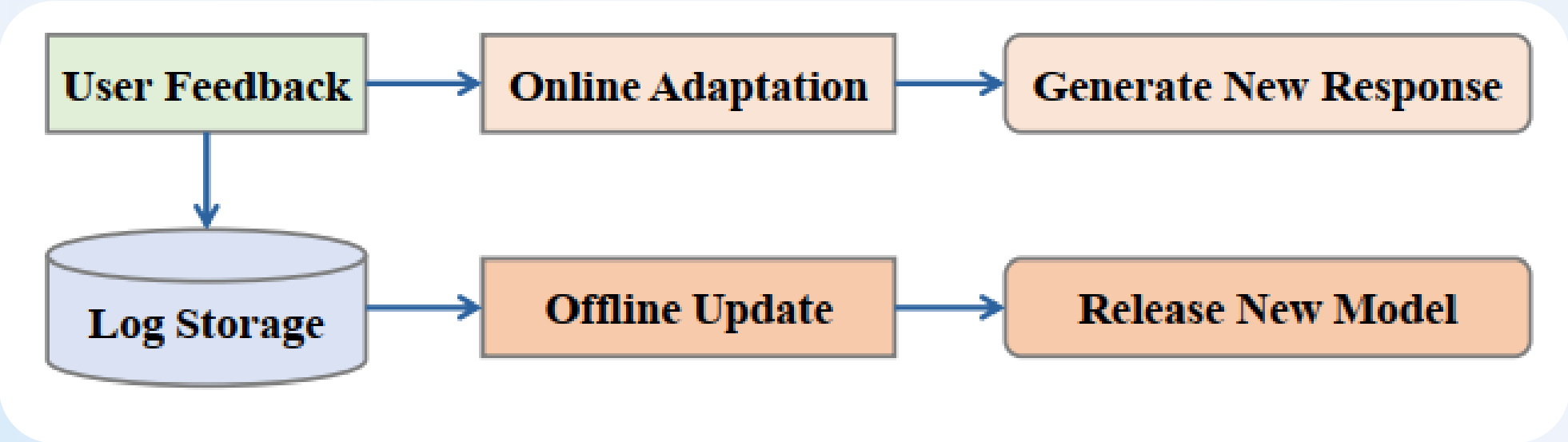
- **提供可复用的调试模板**
- **获取奖励（按使用次数、曝光量或基于表现发放）**
- **购买他人模板以节省时间**

激励贡献行为，实现复用价值，重建以反馈为驱动的搜索生态系统



NExT-Search： 利用用户反馈

- 在线适配（Online Adaptation）：在当前会话内，基于用户反馈实时优化响应内容
- 离线更新（Offline Update）：通过累积的用户日志驱动模型重新训练，强化自改进式飞轮



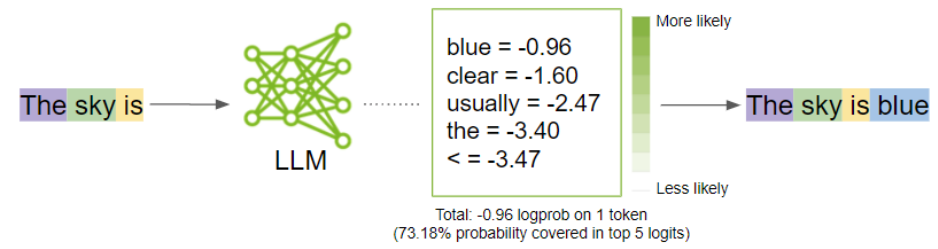
03

在大模型视角下
高效利用外部知识的 RAG 实例

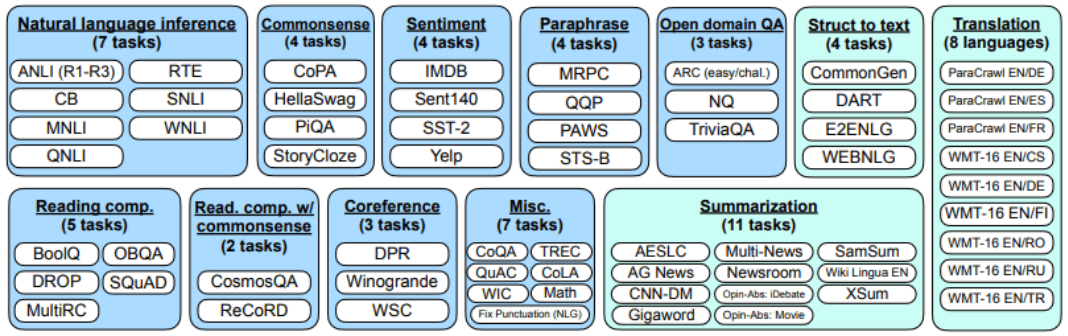
How 2：大模型视角下的检索增强大模型

大模型如何鲁棒地对抗输入的噪音知识，并在参数内外知识之间做出选择

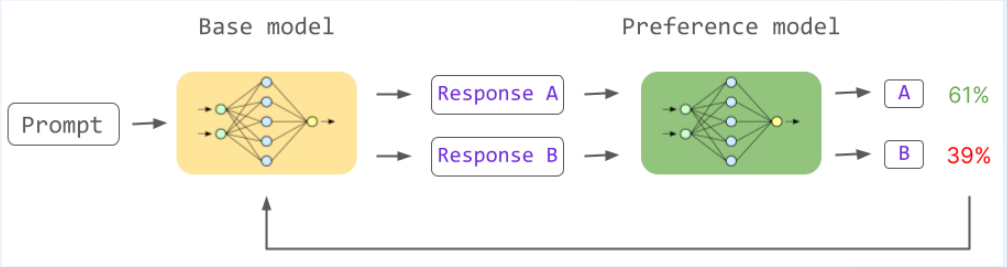
① Pretraining阶段 – Next Token Prediction



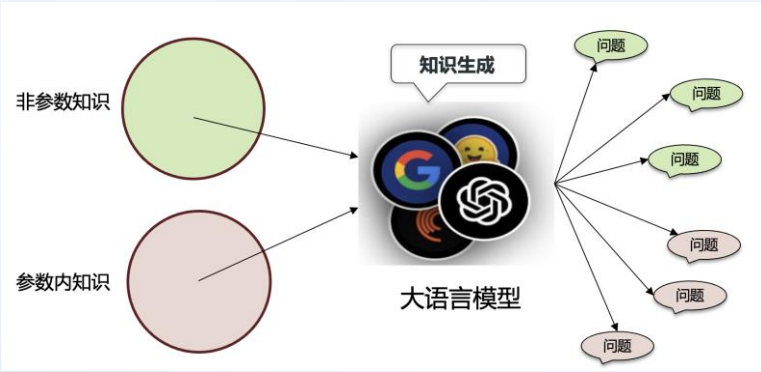
② Instruction Tunning阶段 – Multi-task Learning



③ RLHF阶段 – Alignment



利用检索到的信息（没有特殊优化）



大模型视角下的研究现状

通过微调来对齐大语言模型在检索增强生成中的能力

- **有监督指令微调：**在领域特定数据集上构造检索-问题-答案三元组，利用构造的有监督三元组进行指令微调，教会大模型如何使用检索到的文档。如FID，RetRobust。

- **动态检索增强生成微调：**微调大语言模型使其主动进行动态决策，判断是否进行检索增强生成，如Active-RAG，Self-RAG，Rowen。

均需要有监督的数据

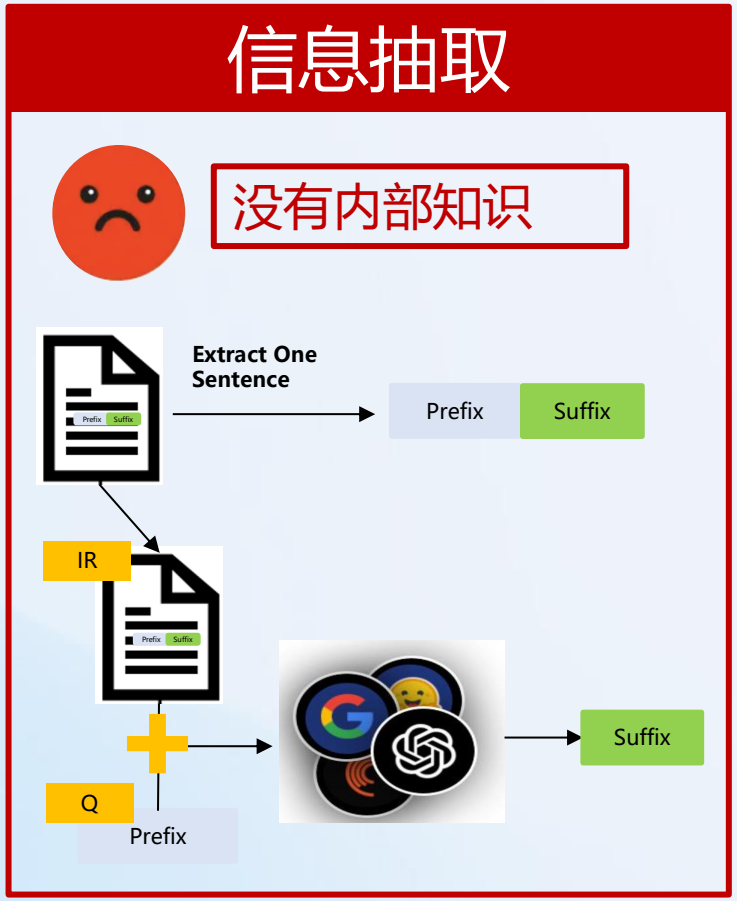
有监督的数据是否是必须的？

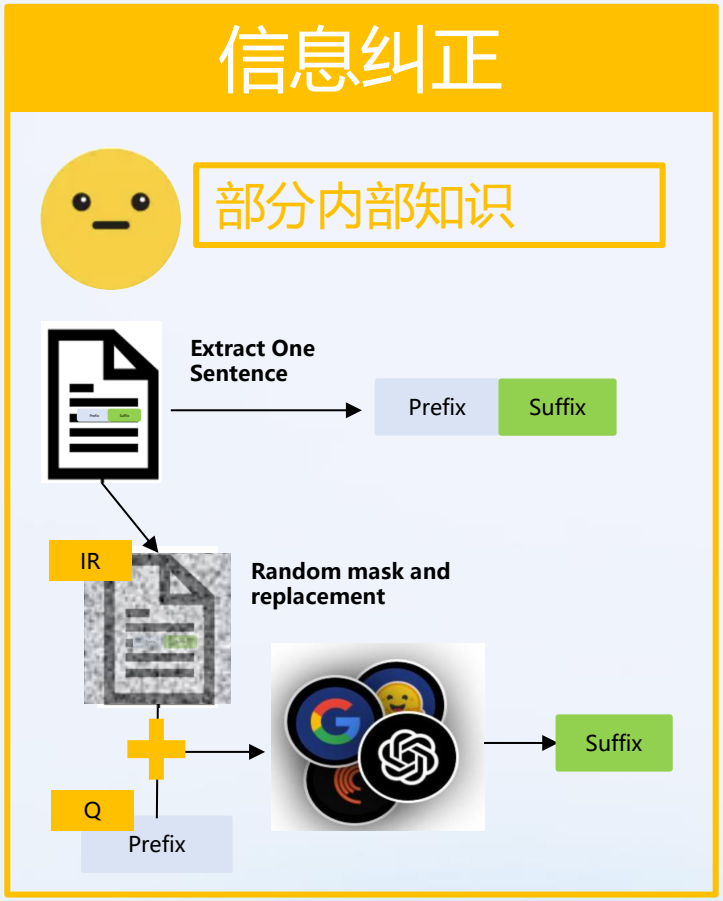
视角一：构建无监督数据和任务

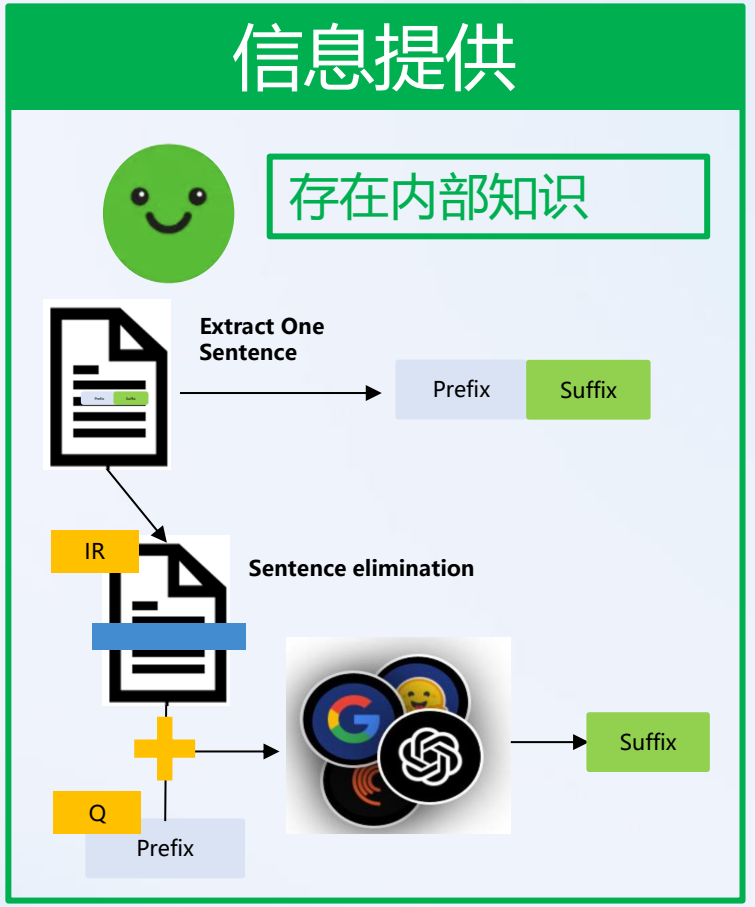
INFO-RAG：无监督检索增强训练

ACL 2024

构建检索增强生成（RAG）的自监督任务：信息抽取、信息纠正和信息提供



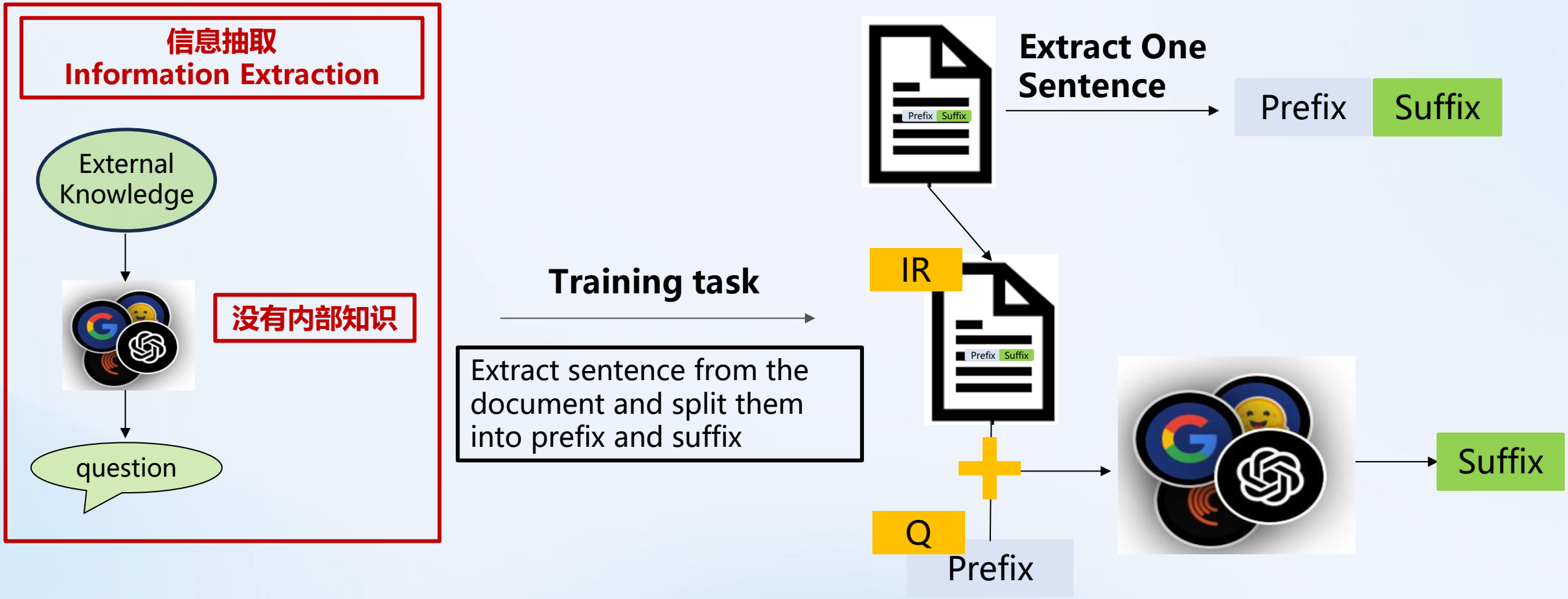




Unsupervised Information Refinement Training of Large Language Models for Retrieval-Augmented Generation. *The 62nd Annual Meeting of the Association for Computational Linguistics (ACL'24)*

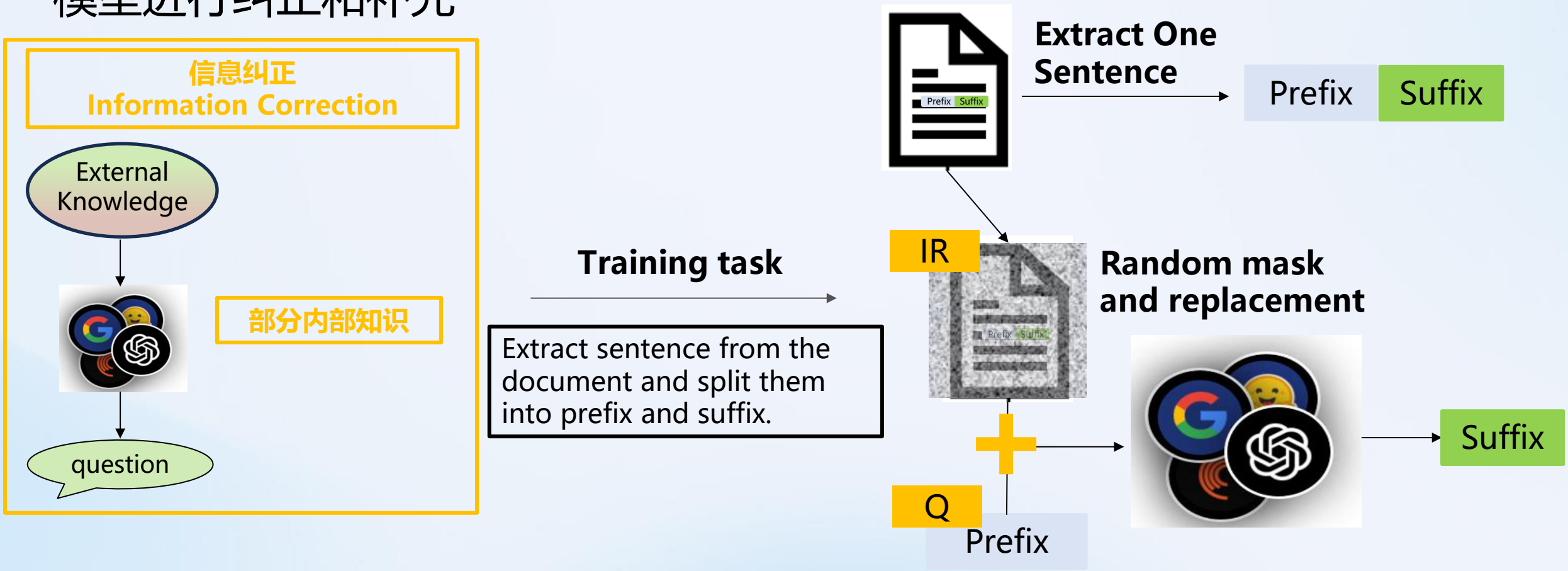
INFO-RAG: 方法

所有正确答案都包含在检索到的文本中，大模型只需从中提取即可



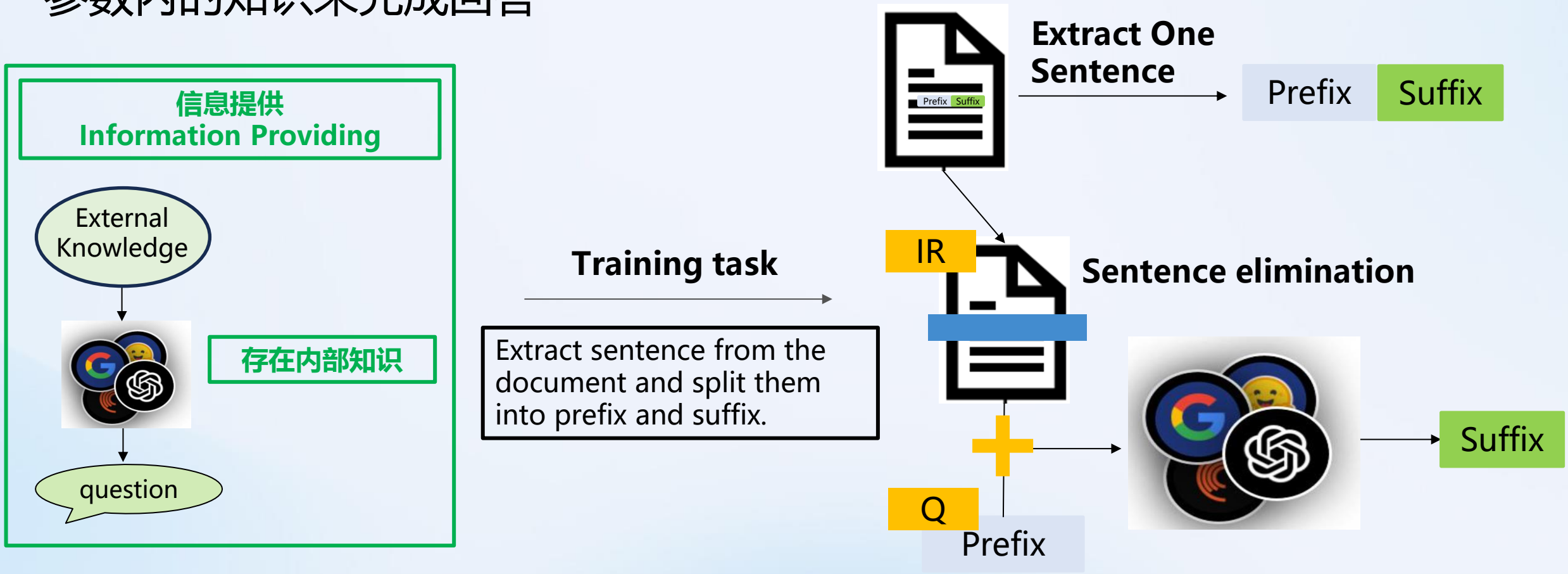
INFO-RAG: 方法

检索到的文本只包含部分答案，甚至可能包含一些错误的答案，需要由大模型进行纠正和补充

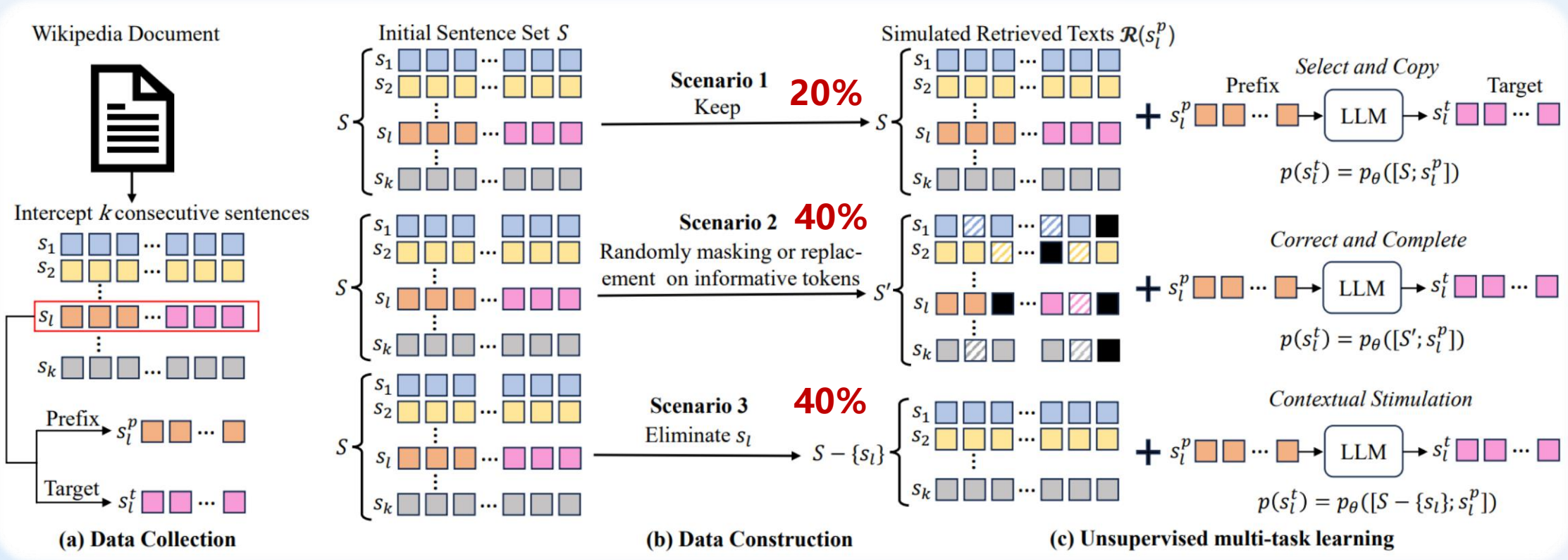


INFO-RAG: 方法

这些检索文本只是与问题在语义上相关，大模型需要借助这些信息激发其参数内的知识来完成回答



INFO-RAG：方法



(a) 从维基百科文档中将文档片段构造成给定prefix和prefix相关的上下文来预测target的无监督样本对

(b) 将这些无监督的样本构造出对应三种场景的任务形式

(c) 混合三种场景的任务进行无监督多任务训练

INFO-RAG：实验

	Soft-Filling		ODQA		Multi-Hop QA		LFQA	Dialog	LM	Code Gen		Overall
	Accuracy		Accuracy		Accuracy		ROUGE	F1	ROUGE	CodeBLEU		
	T-REx	ZS	NQ	WebQ	Hotpot	Musique	ElI5	Wow	WikiText	Python	Java	
LLaMA-2-7B	55.60	54.08	**46.82**	43.52	39.40	25.95	15.18	7.85	60.77	21.44	22.99	35.78
+ INFO-RAG	**65.91**	**57.01**	45.74	**44.68**	**46.56**	**30.19**	**17.18**	**9.09**	**62.91**	**26.75**	**32.06**	**39.83**
LLaMA-2-7B-chat	60.63	55.03	49.42	46.72	50.03	42.69	27.81	10.21	60.26	22.46	23.90	40.83
+ INFO-RAG	**65.77**	**58.32**	**53.93**	**49.13**	**52.01**	**44.45**	**28.15**	**10.49**	**63.24**	**27.25**	**28.79**	**43.78**
LLaMA-2-13B	60.08	50.77	47.40	44.62	42.12	25.78	14.80	7.04	62.20	21.52	29.16	36.86
+ INFO-RAG	**62.80**	**55.63**	**47.82**	**45.42**	**51.48**	**35.02**	**17.48**	**7.20**	**64.14**	**29.00**	**35.50**	**41.04**
LLaMA-2-13B-chat	62.53	56.81	50.36	45.47	61.23	47.06	27.07	11.19	60.52	22.34	30.96	43.23
+ INFO-RAG	**65.39**	**59.05**	**54.04**	**51.07**	**61.91**	**47.93**	**27.24**	**11.38**	**63.92**	**31.98**	**38.12**	**46.55**

INFO-RAG做为一种无监督训练方式
可以应用到现存的大模型中并进一步提升其在各个任务上检索增强的能力

INFO-RAG：实验

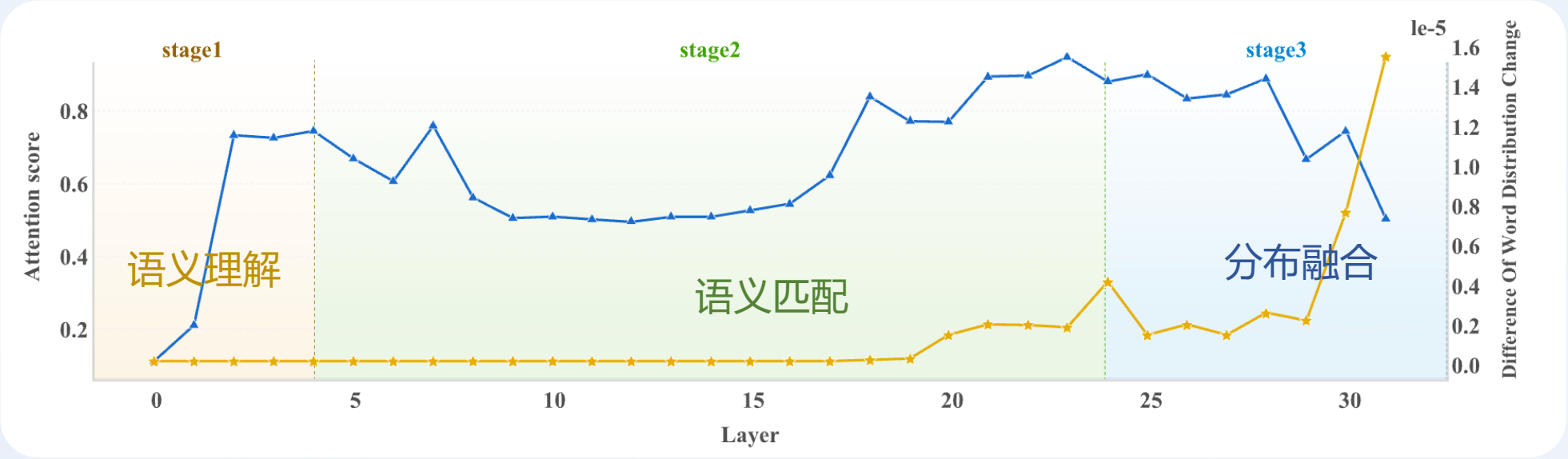
	T-REx			ZS			NQ			WebQ		
	has-ans.	replace	no-ans.	has-ans.	replace	no-ans.	has-ans.	replace	no-ans.	has-ans.	replace	no-ans.
LLaMA-2-7B	67.19	38.37	6.49	64.41	12.78	2.44	**65.54**	16.91	3.41	60.64	25.68	7.90
+ INFO-RAG	**79.80**	**41.79**	**7.04**	**68.10**	**13.55**	**3.26**	64.43	**22.68**	**4.70**	**62.70**	**26.48**	**8.96**
LLaMA-2-7B-chat	73.79	40.56	4.87	66.71	14.19	1.63	68.72	20.81	4.50	66.86	28.63	5.62
+ INFO-RAG	**80.01**	**42.92**	**5.42**	**69.64**	**15.02**	**2.65**	**70.99**	**23.14**	**5.62**	**68.73**	**29.74**	**9.12**
LLaMA-2-13B	72.26	39.47	7.76	60.14	19.71	4.69	**65.94**	18.45	4.42	62.09	26.63	9.27
+ INFO-RAG	**75.80**	**44.08**	**8.48**	**65.94**	**23.21**	**4.90**	64.98	**27.60**	**8.02**	**63.51**	**28.24**	**9.88**
LLaMA-2-13B-chat	75.96	43.79	5.59	67.03	16.58	1.42	69.37	30.72	6.16	65.07	31.88	5.47
+ INFO-RAG	**79.25**	**48.59**	**6.67**	**70.26**	**25.02**	**3.87**	**73.73**	**33.85**	**8.39**	**70.59**	**37.48**	**11.25**

INFO-RAG对于三种检索增强的场景都有明显的提升
对检索文档中的噪音具有较强的鲁棒性

视角二：探索模型内部知识选择机制

Tok-RAG：可解释检索增强生成

探索模型内部参数知识与外部检索知识的融合过程：语义理解、语义匹配和分布融合



$p(x_i|x_{1:i-1})$

从大语言模型预训练的知识分布中采样隐变量

知识融合

$p_R(x_i|x_{1:i-1})$

从检索到的文本概率中采样

A Theory for Token-Level Harmonization in Retrieval-Augmented Generation, ICLR 2025

Tok-RAG：基本理论框架

理论基础：大语言模型的文本生成过程是隐式的隐变量推断

$$p(x_i|x_{1:i-1}) = \int_{\mathcal{Z}} p(x_i|x_{1:i-1}, z)p(z|x_{1:i-1})\, dz,$$

给定前缀，预测当前token的分布

基于采样隐变量的分布预估

隐变量采样

引入外部检索隐变量 z^*

分析框架：利用隐变量推断来分析大语言模型RAG的知识融合过程

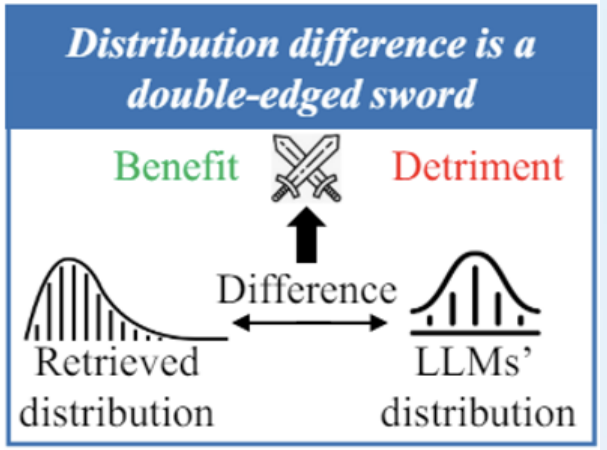
$$\begin{aligned} p(x_i|R, x_{1:i-1}) &= \int_{\mathcal{Z}} p(x_i|R, x_{1:i-1}, z)p(z|R, x_{1:i-1})\, dz \qquad (2) \\ &= \int_{\mathcal{Z}-\{z^*\}} p(x_i|R, x_{1:i-1}, z)p(z|R, x_{1:i-1})\, dz + p(x_i|R, x_{1:i-1}, z^*)p(z^*|R, x_{1:i-1}). \end{aligned}$$

Tok-RAG：理论

参数知识与检索知识分布差异带来了RAG中的收益与损失

$$\underbrace{\mathrm{KL}(p_R(r)\|p(r|z))}_{\textbf{benefit}} - \underbrace{\mathrm{KL}(p_R(r)\|p(r|z^*))}_{\textbf{detriment}} \propto \frac{1}{\mathcal{D}}$$

收益 **损失**



$p_R(r)$

外部检索知识的分布

差异大 收益大

差异大 损失大

$p(r|z)$ 内部参数知识的实际分布

表示内外部知识是否互补

$p(r|z^*)$ 内部参数知识的在给定外部主题之后的分布

表示内外部知识是否冲突

$$p(x_i|x_{1:i-1})$$

$$p(x_i|x_{1:i-1},R)$$

$$p_R(x_i|x_{1:i-1})$$

Tok-RAG：理论

$p(x_i|x_{1:i-1})$

内部参数分布**不加检索**
下一个词的概率

$p(x_i|x_{1:i-1}, R)$

内部参数分布**加检索**
下一个词的概率

$p_R(x_i|x_{1:i-1})$

外部检索分布
下一个词的概率

$S_1 = \mathrm{Sim}(p(x_i|x_{1:i-1}), p(x_i|x_{1:i-1}, R))$

检索带来的分布变化的小

$>$

$S_2 = \mathrm{Sim}(p_R(x_i|x_{1:i-1}), p(x_i|x_{1:i-1}, R))$

变化来自检索分布的比例小

不使用检索

Tok-RAG：理论

$p(x_i|x_{1:i-1})$

内部参数分布**不加检索**
下一个词的概率

$p(x_i|x_{1:i-1}, R)$

内部参数分布**加检索**
下一个词的概率

$p_R(x_i|x_{1:i-1})$

外部检索分布
下一个词的概率

$S_1 = \mathrm{Sim}(p(x_i|x_{1:i-1}), p(x_i|x_{1:i-1}, R))$

检索带来的分布变化的大

$<$

$S_2 = \mathrm{Sim}(p_R(x_i|x_{1:i-1}), p(x_i|x_{1:i-1}, R))$

变化来自检索分布的比例大

使用检索

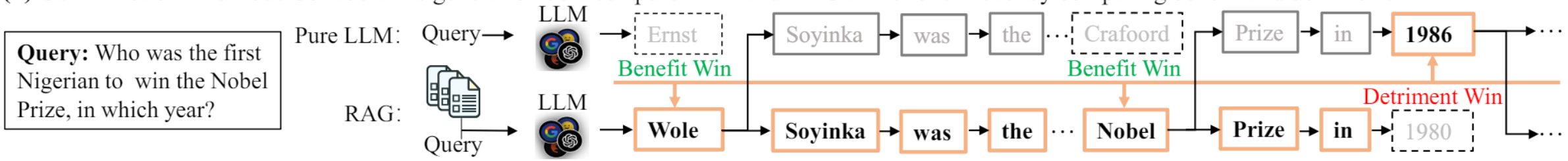
Tok-RAG：方法

$$w_{LLM}: p(x_i|x_{1:i-1}) \quad w_{RAG}: p(x_i|x_{1:i-1}, R) \quad w_{IR}: p_R(x_i|x_{1:i-1})$$

$$S_1 = \cos(w_{RAG}, w_{LLM}) \qquad S_2 = \cos(w_{RAG}, w_{IR})$$

$$s = \begin{cases} \text{benefit win} & \text{if } \cos(\mathbf{w}_{RAG}, \mathbf{w}_{IR}) \geq \cos(\mathbf{w}_{RAG}, \mathbf{w}_{LLM}), \\ \text{detriment win} & \text{if } \cos(\mathbf{w}_{RAG}, \mathbf{w}_{IR}) < \cos(\mathbf{w}_{RAG}, \mathbf{w}_{LLM}), \end{cases}$$

(b) Our Practical Method: Collaborative generation between pure LLM and RAG at the token-level by comparing benefit and detriment.



Tok-RAG：实验

Methods	Train LLM	Add Module	TriviaQA						WebQ						Squad					
			Ratio of Hard Negative Passages						Ratio of Hard Negative Passages						Ratio of Hard Negative Passages					
			100%	80%	60%	40%	20%	0%	100%	80%	60%	40%	20%	0%	100%	80%	60%	40%	20%	0%
Standard RAG	no ✔	no ✔	43.8	67.0	71.3	76.2	78.2	81.9	23.9	35.8	40.6	43.4	48.4	53.1	8.6	31.0	43.2	53.0	58.8	67.2
NLI+RAG	no ✔	need ✗	50.8	61.2	68.2	73.0	76.4	79.1	30.7	40.3	44.5	47.5	50.9	52.8	9.9	21.1	33.7	43.4	51.7	60.5
CRAG	no ✔	need ✗	48.2	68.3	72.5	76.7	81.5	82.2	25.6	37.4	41.9	46.2	51.5	54.9	7.4	28.7	39.6	50.7	53.2	61.1
RetRobust	need ✗	no ✔	49.2	67.3	72.9	77.5	79.4	82.3	30.0	38.9	42.5	48.2	49.8	54.3	10.5	30.8	43.3	52.5	58.4	66.0
Self-RAG	need ✗	no ✔	43.0	68.7	73.5	76.4	80.8	82.2	18.3	34.8	42.2	47.2	51.3	57.0	5.5	27.8	38.9	46.4	52.5	58.3
INFO-RAG	need ✗	no ✔	49.7	68.4	73.2	77.9	80.0	82.5	29.7	38.0	43.9	48.1	49.4	54.8	10.7	30.1	43.5	53.7	59.2	67.5
X-RAG (Ours)	no ✔	no ✔	**53.5**	**72.9**	**77.6**	**81.3**	**83.4**	**85.7**	**32.9**	**43.8**	**47.3**	**50.0**	**52.9**	**57.3**	**12.8**	**31.3**	**44.5**	**54.1**	**60.8**	**68.1**

在实际的开放域问答任务的检索增强生成中，X-RAG不需要借助额外的模块或训练LLM，
便能够超越主流的鲁棒RAG框架及训练方法，如RetRobust，INFO-RAG，Self-RAG等

Tok-RAG：实验

LLMs	Methods	# Generation Times	Wikitext AUC	Wikitext F1	ASQA AUC	ASQA F1	Bio AUC	Bio F1	NQ AUC	NQ F1
OPT-6.7B	Logprobs	2	65.25	64.33	68.96	67.55	65.24	64.59	55.31	51.41
	Uncertainty	2	64.12	63.50	66.14	63.96	65.78	64.60	56.03	52.15
	Consistency-Lexical	10	64.01	62.17	69.42	67.04	65.41	65.28	55.06	51.13
	Consistency-Semantic	10	65.93	64.22	70.11	69.50	65.76	64.37	56.24	52.88
	X-RAG (Ours)	2	**68.64**$^{+}$	**66.88**$^{+}$	**72.28**$^{+}$	**72.05**$^{+}$	**66.27**$^{+}$	**66.04**$^{+}$	**57.92**$^{+}$	**52.90**$^{+}$
Mistral-7B	Logprobs	2	73.52	72.90	68.05	66.86	65.22	64.39	57.04	57.23
	Uncertainty	2	73.72	72.71	67.47	65.63	65.59	65.83	57.19	57.10
	Consistency-Lexical	10	72.15	70.44	69.16	67.33	64.79	64.33	56.95	54.37
	Consistency-Semantic	10	73.98	72.26	70.05	69.54	65.68	65.12	57.43	56.12
	X-RAG (Ours)	2	**75.85**$^{+}$	**74.11**$^{+}$	**71.51**$^{+}$	**71.47**$^{+}$	**66.37**$^{+}$	**66.04**$^{+}$	**58.52**$^{+}$	**57.56**$^{+}$
LLaMA-2-7B	Logprobs	2	73.47	72.95	68.50	68.04	62.11	60.94	67.40	69.24
	Uncertainty	2	73.98	73.01	68.72	67.63	63.67	63.50	68.03	69.15
	Consistency-Lexical	10	73.51	71.62	70.09	68.45	62.49	61.98	68.17	70.09
	Consistency-Semantic	10	74.96	74.23	71.23	69.38	63.77	62.10	69.72	71.14
	X-RAG (Ours)	2	**81.89**$^{+}$	**80.42**$^{+}$	**76.96**$^{+}$	**76.80**$^{+}$	**64.08**$^{+}$	**64.19**$^{+}$	**70.50**$^{+}$	**72.45**$^{+}$

在判断token级别的正确性方面，X-RAG超越了目前主流的基于对数概率，不确定度以及自一致性的幻觉检测方法。

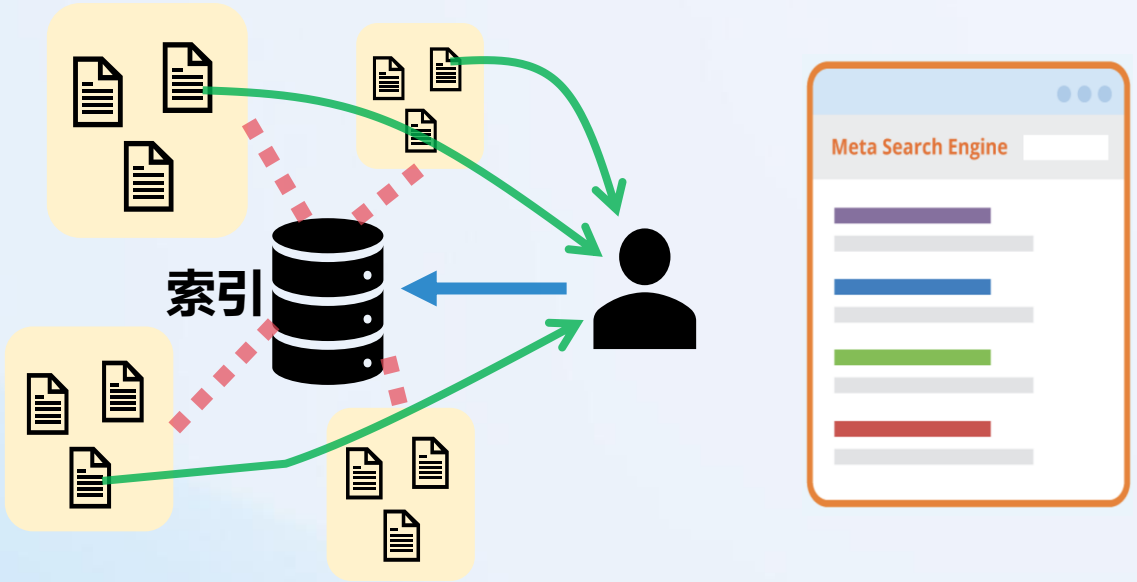
04

未来方向与行业启示

信息流动的角度

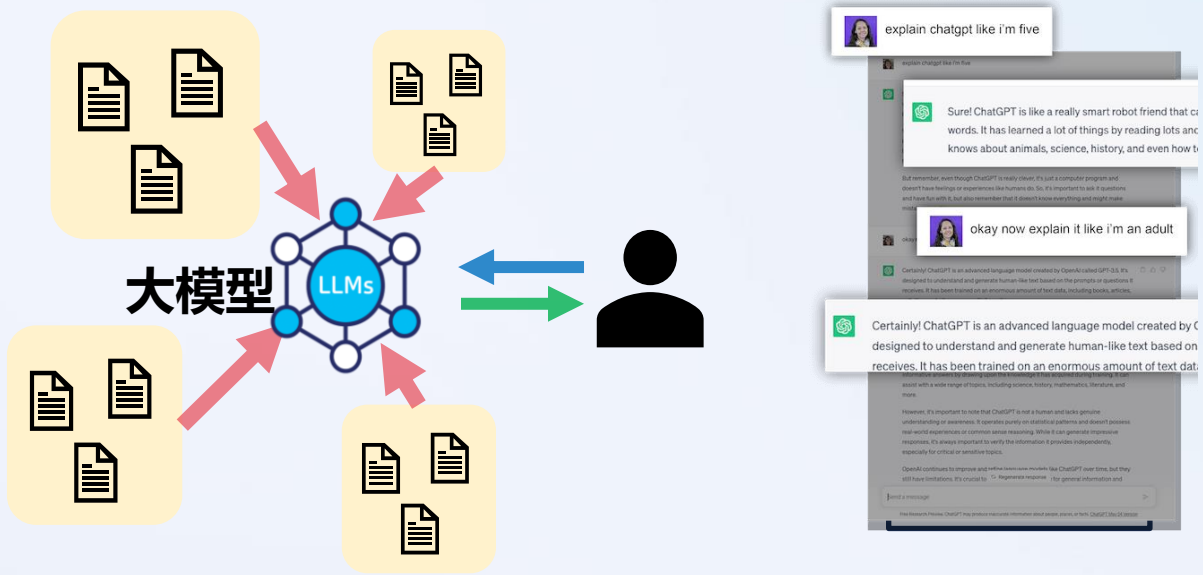
搜索引擎 + 大语言模型 → 搜索引擎 X 大语言模型

搜索引擎



① 分布式信息　② 原生态信息

大语言模型

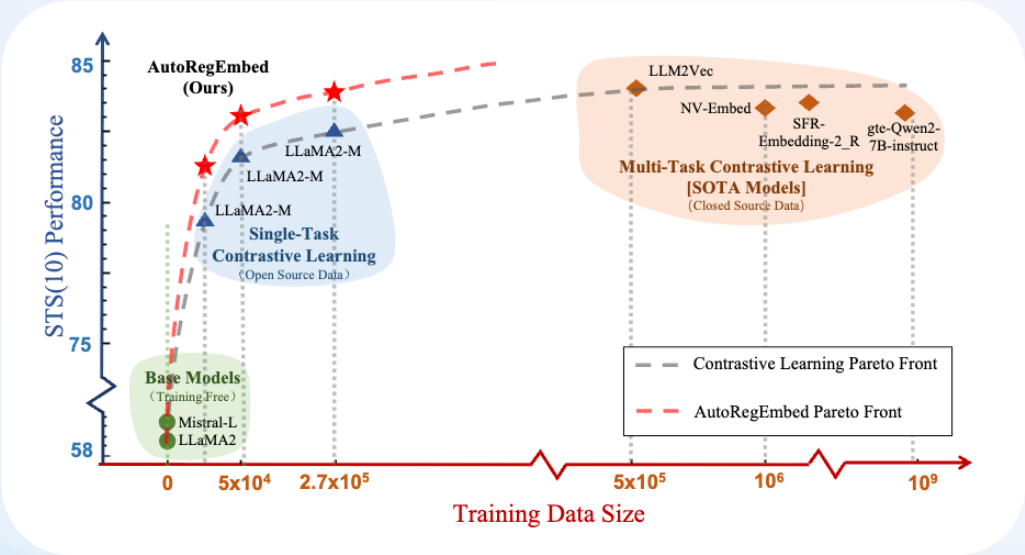


① 集中式信息　② 加工态信息

Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks, WWW 2024

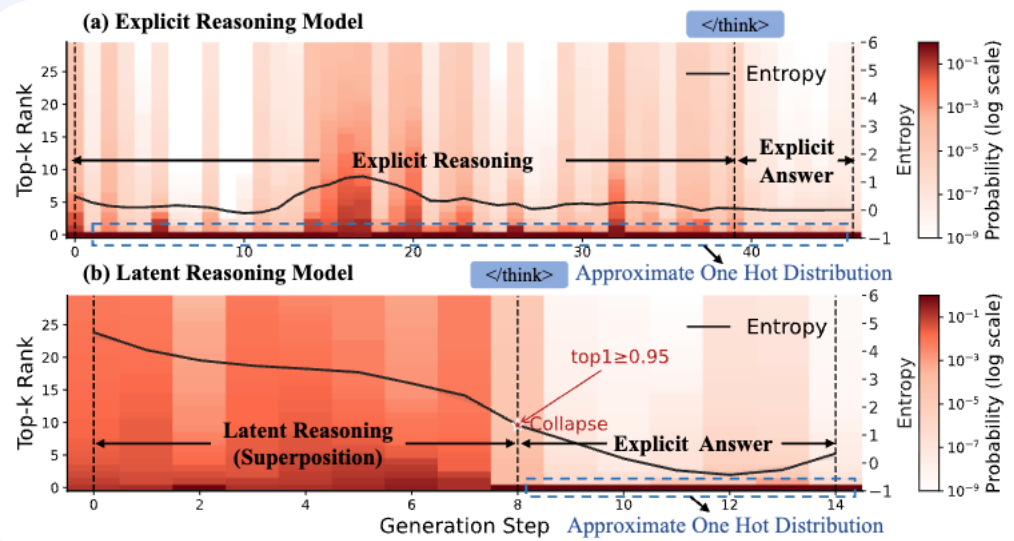
信息压缩的角度

更好的利用大语言模型内部压缩的大量信息



从大语言模型中抽取检索稠密向量效果更好

Following the autoregressive nature of llm embeddings via compression and alignment, EMNLP 2025



隐式推理的效果逼近显式推理

Latent Reasoning in LLMs as a Vocabulary-Space Superposition, Arxiv 2025

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

突破参数知识边界：自增强优化的检索增强大模型技术