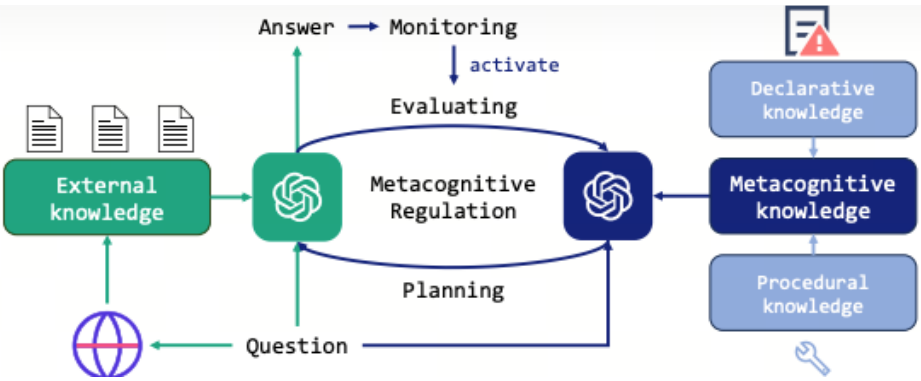
知乎



智源研究院2024：运用高级思考策略（元认知检索）提升的大语言模型（LLM）性能

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

1 人赞同了该文章

Introduction

大型语言模型（LLMs）因其能理解并生成人类语言的能力，在多种NLP任务中表现出色。然而，它们生成虚构内容的问题引起了关注。为保证内容真实，研究者引入检索系统，让LLMs从外部知识库获取信息。传统的一次性检索方法对明确任务有效，但在多步推理复杂任务中不足。因此，研究转向了多时序检索框架，它不再仅限于一次检索，而是通过循环过程迭代获取和使用知识，如分解问题、利用生成内容做动态查询，以提高生成的准确性和可靠性。尽管先前的RAG技术提高了答案质量，但它们受限于固定推理步骤，无法自我诊断错误并优化。

我们认识到，这是由于缺乏元认知能力，即自我反思和自我调整。人类在解决复杂问题时，会运用元认知。因此，我们提出MetaRAG（Metacognitive Retrieval-Augmented Generation），它结合了RAG和元认知。MetaRAG通过监测、评估和规划，让模型具备内省式推理，能动态检测和纠正错误，从而实现对生成内容的精准优化。实验证明，这种改进显著提升了模型性能。

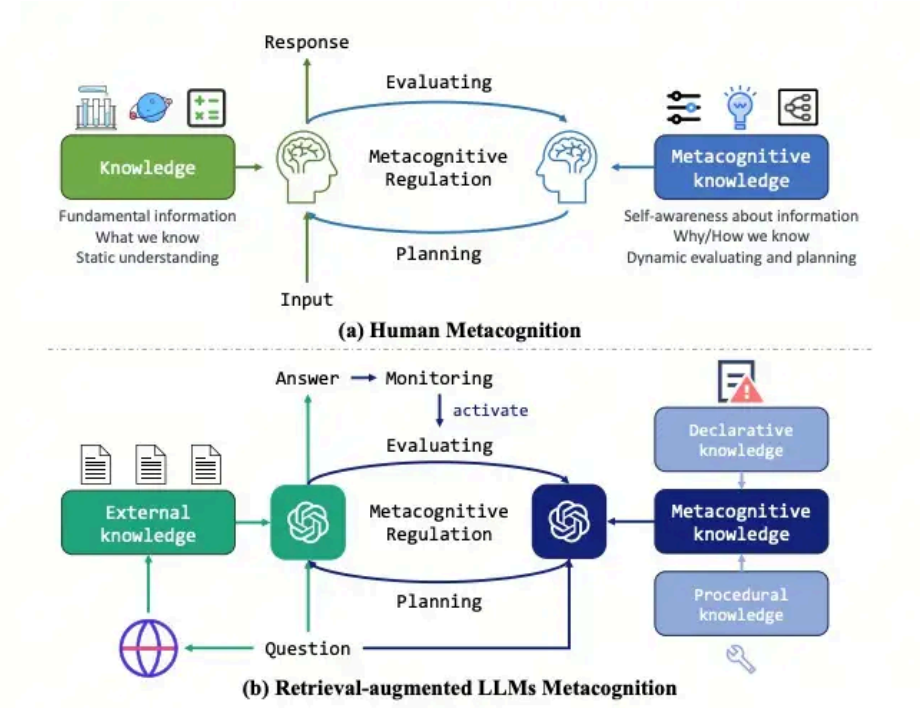


Figure 1: The correspondence between the metacognitive processes in humans and retrieval-augmented LLMs

元认知，源自认知心理学，关注个体对自己认知活动的理解和评价。图(a)展示了元认知的两个组成部分：元认知知识（理解自身认知能力及局限）和元认知调节（主动管理认知过程）。元认知

知乎

借鉴元认知原理，我们构建了MetaRAG（Meta-Cognitive Retrieval-Augmented Generation）框架，如图(b)所示，它包含一个认知-元认知互动结构。认知层⁺处理问题和参考以生成答案，而元认知层则监督和质疑这个过程，寻找潜在的错误。通过分析不同知识情境下的模型表现，我们发现模型出错主要有三原因：知识不充分、知识冲突和错误推理。

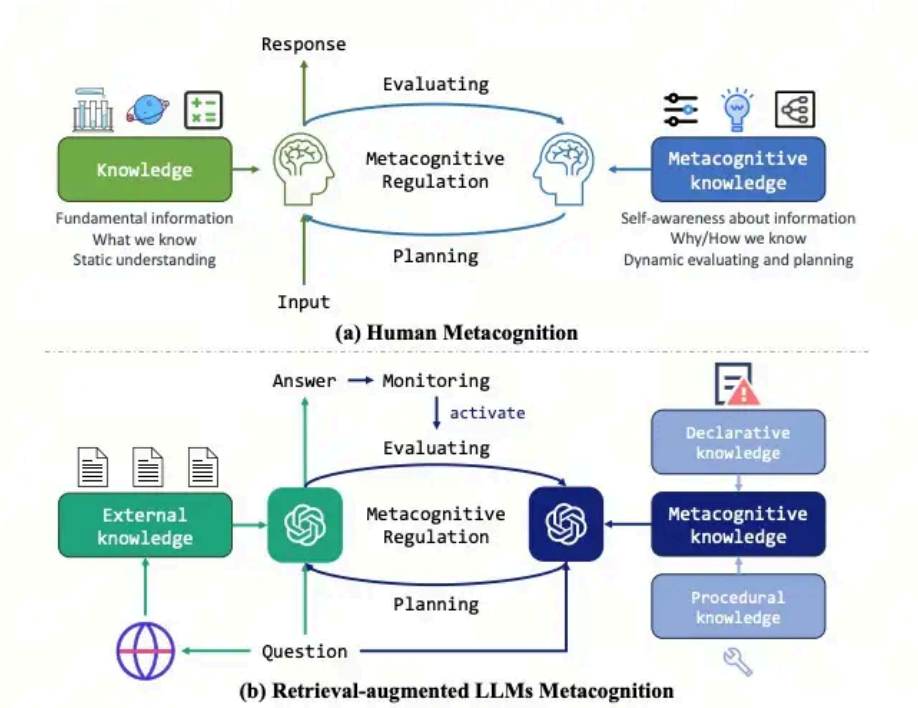


Figure 1: The correspondence between the metacognitive processes in humans and retrieval-augmented LLMs

MetaRAG通过元认知，旨在两个关键领域改进RAG任务：一是确保从外部检索和LLM内部知识的匹配与连贯性；二是保证多步推理的逻辑性和有效性。这样，MetaRAG能自我检测并针对性地调整，从而提高生成答案的精确度。首先的**监测**环节检查响应质量，决定是否启动元认知评估；其次的**评估**阶段，利用元认知知识⁺分析响应的错误，分类为四种情况，涉及陈述性知识和程序性知识；最后的**规划**阶段，根据评估结果给出个性化建议，以修正认知问题。

本研究的主要创新在于融合了LLMs和人类的内省思维，设计了MetaRAG框架，专为多跳问答任务服务。我们通过实证研究揭示了多跳问答中的三大问题：知识不充分、知识冲突和错误推理。通过这三步元认知调控，MetaRAG不仅能识别问题，还能有效改进，特别是在推理能力和表现上明显优于现有基准。

Preliminary

我们首先定义了检索增强生成的任务，这是为了克服单一检索在复杂问题处理中的局限。然后，我们关注了这一任务在多跳问答中的具体问题，揭示了模型在理解和生成过程中可能存在的错误来源。通过对这些局限性的深入理解，我们提出了融入元认知的方法，以帮助模型自我诊断和优化其推理过程，从而提高答案的准确性和质量。

Task Definition

1. 信息匹配度低：尽管能检索大量信息，但可能无法满足问题的深度需求，导致答案不完整。
2. 推理复杂性：多步推理需要模型理解并整合复杂关系，一次检索往往无法胜任。
3. 知识融合问题：整合不同来源且可能矛盾的信息是个挑战，影响答案质量和一致性。
4. 动态理解适应性：面对复杂问题，RAML需提升对新查询语义的理解能力。
5. 评价标准缺失：现有指标不足以全面评估模型⁺在多步问答中的表现，需要开发专门的评估方法。 1 因此，未来的研究焦点将致力于改进RALM，以克服这些问题，提升在多跳问答任务中的效能。

$$y = \mathrm{LLM}_{\mathrm{QA}}([D_q, q], \mathrm{Prompt}_{\mathrm{QA}})$$

知乎

信息，由LLM进行精确或近似答案的生成：

$$\mathbf{RALM}(q, D_q; [q, D_q]) = \mathbf{LLM}_{\mathbf{QA}}([q, D_q])$$

未来工作将集中于优化RAML，解决信息不匹配、推理复杂性、知识融合、动态理解适应性及评价指标问题，以提升其在多跳问答任务的表现。

Task Exploration

我们通过实验研究考察了在不同知识背景下的检索增强语言模型性能。目标是区分模型依赖LLM内在知识还是外部检索。我们采用人工评估，分别评估两类情况： (a) 评价LLM直接提供的课本答案准确性。 (b) 评估检索文档的完整性，看它们能否提供充足信息。 将问题分为四类：纯内源（仅依赖LLM）、单次检索、混合（LLM与检索结合）和无外在知识。这样做的目的是量化RAML在各种条件下的表现，并理解其行为，以便优化模型构造和训练策略。

在这个实验设置中，我们关注的是模型在没有外部知识支持的情况下表现。这意味着无论依赖LLM还是检索，都无法得出有效的解答，归类为"无知识"条件。这样的分析旨在量化RAML在缺乏直接信息来源时的能力，为模型优化提供实证数据和深入理解。

在某些情况下，答案可能存在于检索文档中，但不是由LLM（如$\mathbf{LLM}_{\mathbf{QA}}$）直接生成。这种条件下，我们考察的是模型仅依赖检索信息的能力。研究将揭示RAML在面对纯粹依赖外部信息时的性能，以此评估其对知识获取和整合的依赖程度。

在这个类别下，答案完全源于LLM的内在知识，不依赖于任何外部检索。我们关注的是LLM本身能否独立提供准确的答案，以评估其内部知识结构和理解能力。

在内外兼备条件下，RAML既利用LLM的内在知识，也参考检索到的文档信息来生成答案。这种设置旨在考察模型在既有内部资源又接入外部资源时的表现，以全面评估其知识整合和推理能力。

(1) 在零知识情境中，模型生成准确答案的难度增加。

(2) 独立于LLM或检索的单一知识来源提升准确性，但存在知识不一致可能导致错误。

(3) 当内外部知识都能解决问题时，答案质量提升，但仍非绝对，因为模型可能基于错误的逻辑推断。这揭示了多跳问答中的三大困境：缺乏、冲突及错误推理。 为了应对这些问题，我们关注元认知策略[+]的作用，以期通过这些策略优化模型在处理这些问题时的性能。

Metacognitive RAG

1. 监测：实时检查模型在处理问题时的行为和决策，确保其遵循正确的路径。
2. 评估：对生成的答案进行深度评估，识别潜在的错误或不一致性。
3. 规划：如果发现问题，元认知空间会指导模型调整策略，优化其推理过程。

Monitoring: Assessing Answer Satisfaction

监测主要功能是追踪和记录认知过程，但并非所有认知活动都伴随元认知评估。在复杂问题面前，只有当存在不确定性时，元认知才会介入。在多跳问答任务中，LLM可能因知识不足而产生错误。因此，监测的目标是判断答案质量，必要时启动元认知评估。我们通过专家模型作为标准，对LLM和检索答案进行比较，如若差异大或不符合预期，将启动元认知评估。具体来说，给定问题和相关文档集，我们让专家模型M生成答案，然后通过分析专家答案与LLM答案的相似性来评估其满意度，决定是否需要进一步审查。$y' = M_\phi([D_q, q])$在这个过程中，我们利用模型M（参数为ϕ）产生的专家答案y'来评估。通过比较LLM的输出y与y'，我们使用一个度量$similarity(y, y')$来衡量两者相似性。以此决定是否需要进行元认知评估。如果相似度高，说明答案可能来自正确的知识源，不需要额外干预；若相似度低，可能暗示存在矛盾或信息不匹配，这时引入元认知以检查和纠正可能存在的问题。干预策略将依据设定的阈值和具体问题来制定。

在这个系统中，参数k是个关键阈值，它决定了何时启动元认知评估。若LLM的输出$\mathbf{y}$与专家模型的输出y'通过余弦相似性[+]计算的$\mathbf{similarity}(y, y')$小于$k$，专家模型就会启动元认知过程。这包括元认知评估来验证答案的正确性以及规划可能的调整，以保证答案的精确性和一致性。

知乎

为解决这些问题，我们利用元认知的两种类型：程序性知识和陈述性知识。程序性知识关注的是执行任务所需的具体策略，而陈述性知识涉及问题的实质性内容。在RAML背景下，我们将LLM转变为评估者-批评者，使其不仅生成答案，还对其推理过程进行批判性评估，以便更公正地审视其在解决复杂问题时的局限性。

内部知识评估：通过LLM本身，我们检查它是否能有效地利用内置知识来解答问题。我们让这个版本的LLM作为评估者，以二元形式评估问题，即：

通过：如果LLM能够正确且适用地利用其内部知识给出答案。

未通过：如果答案不合理或与预期不符，显示其内部知识应用存在问题。

$$\mathbf{LLM}_{\text{Eval-Critic}}(q, \mathbf{Prompt}_{\text{Eval}})$$

外部知识评估：通过运用先进的NLU模型TRUE ，我们检验检索到的文档集合D_q是否提供了充足的知识来解答问题。评估方法如下：

满足：如果TRUE 确认文档集包含了足够的信息来支持LLM生成准确答案。

不满足：若TRUE 判定文档不足以解释问题，说明外部知识源可能不够充分。

$$f\left(\left[\{d_i\}_{i=1}^{|D|}\right], q\right)$$

(b) 问题集中在MetaRAG中的陈述性知识，目标是检测推理中的常见误区。我们将这些错误分为3类： 1. 逻辑谬误⁺：源于违反逻辑原则，如无效演绎或归纳推理。 2. 信息缺失：源于系统缺乏必需信息，可能导致决策失误，如关键事实理解错误或遗漏。 3. 知识不精确：由于知识库可能存在不准确或过时信息，由此引发的推理错误。 为应对这些问题，我们的策略将着重提升模型的逻辑推理技巧，确保完整信息获取，同时定期更新和验证知识库，以降低这类错误发生概率。

- 不完整推理：这是多跳问答中常见的问题，表现为模型未能充分利用所有相关信息，或无法形成逻辑连贯的思维链⁺以得出准确答案。为克服此问题，我们计划优化模型的推理能力，确保全面考虑所有线索，同时加强对知识的整合和更新，以减少此类不完整推理的产生。
- 答案冗余：当模型过度依赖单一信息，未能精炼或整合多个相关数据点，导致生成的答案内容过多且重复，这就是答案冗余。为避免这个问题，我们需要改进模型的处理策略，使其能更精准地整合信息，确保答案简洁且准确。
- 歧义理解：当模型未能准确解析查询中的细微含义，错误地参考了相关但不准确的信息，导致生成的答案偏离了真实意图，这是歧义理解的表现。为克服这一问题，我们需要强化模型的语义理解和解析能力，确保对查询有深入且精确的理解。

Planning: Strategizing Answer Refinement

在元认知规划中，针对知识不足问题的解决策略是生成新的查询。当评估者-批评者LLM识别到这个问题时，它的任务是创造有针对性的新问题，以从语料库⁺中获取缺失的特定信息。新查询的设计应满足两个条件：(1) 跨越原始问题，聚焦于额外需求的信息；(2) 分解为明确的子问题，以便更精细地定位所需知识。利用LLM的内省能力，我们推断并构造这些查询，以驱动获取更完整和准确的知识。

$$q' = \mathbf{LLM}_{\text{Eval-Critic}}([q, D_q, y], \mathbf{Prompt}_{\text{QG}})$$

在处理冲突知识时，元认知规划包括两方面的应对措施。首先，面对知识不一致的情况，我们需要教LLM如何在多种来源的矛盾知识间做出平衡，选择最相关且全面的信息。例如，当历史与现代资料并存，模型需学会判断哪个版本更可信。其次，对于矛盾信息澄清，LLM需要学会解析和解决这些冲突，可能需要对信息源进行深入分析，或者通过逻辑推理来消除误解。通过这些策略，我们在评估阶段发现问题后，旨在指导模型在面对这类问题时能有效地调整和优化，提升答案的准确性和一致性。

当只能利用LLM内部知识时，我们面临的问题是模型可能受限于其训练数据，无法获取或理解外部信息。在这种情况下，评估方法如下： 1. 知识验证：检查LLM的答案是否基于其训练数据内的逻辑和信息，排除依赖外部资源的直接证据。

知乎

在外部知识唯一可用的情况下，避免模型依赖误导，我们的策略是： 1. 严格依赖验证：确保LLM完全基于提供的参考资料来生成答案，排除任何潜在的内在推断或自我解释。

1. 排除误导证据：通过仔细分析模型的推理过程，消除可能源自误导信息的迹象。
2. 反馈调整：如果发现答案与参考资料不符，及时纠正模型，引导它重新审视问题，必要时修正其对信息的理解。

通过这样的方法，我们确保LLM在处理外部信息时保持清醒，避免产生基于幻觉的错误答案。

修正错误推理： 尽管模型能够正确运用内外部知识，但仍可能出现推理错误。为此，我们采取两步策略来改进： 1. 详尽推理校验：在生成答案后，对推理过程进行深度审核，确保每个步骤的逻辑正确，且答案源于正确的内外部信息。如有错误，立即纠正并引导模型学习避免类似错误的策略。 2. 灵活推理调整：在复杂问题中，允许模型根据需要动态调整推理路径。如果发现内部证据不再适用，模型能自动识别并修正，转向利用更相关或更新的外部知识。

这样，即使在全面利用知识的基础上，我们也力求避免因推理错误导致的答案失真，提升模型的稳定性和准确性。

Experimental Setup

Datasets and Evaluation Metrics

为了验证我们的多跳推理模型，我们在两个相关数据集，即HotpotQA和2WikiMultiHopQA上进行了实验，它们均基于维基百科✦构建，提供了丰富的外部文档资源。鉴于实验资源的约束，我们从每个数据集的验证集选取了大约500个样本进行测试。评估上，我们在答案层面上使用精确匹配（EM）来确认预测与标准答案的一致性，而在令牌层面，我们遵循 提出的指导，通过计算F1分数、精度（Prec.）和召回（Rec.）来全面评估模型的性能。

Baselines

在比较中，我们选择了两种典型的封闭型模型以及六种基于检索增强的模型作为基准。这些基准包括： 1. 标准提示：直接指示LLM简单响应。 2. 链式思考：展示推理过程，鼓励深度思考。 3. 标准RAG：使用查询检索文档，然后让LLM综合信息。 4. ReAct：结合推理与行动，内建检索功能✦。 5. Flare：在生成答案时主动检索。 6. IR-CoT：交替使用检索和因果推理（CoT）。 7. 自我询问：通过整合中间步骤处理复杂问题。 8. 反思：引入评估者以强化语言代理。

为了保证公正的对比，我们在所有模型上实施统一的配置，包括相同的演示上下文、提示格式、检索器和文档库设置。

Implementation Details

在认知处理阶段，我们选用高性能的'gpt-35-turbo-16k' LLM，通过API反复调用，温度设置为0。鉴于数据集主要基于维基百科，我们利用维基百科的完整内容作为文档库，每篇文章切分为100个令牌的段落。检索文档时，我们采用BM25算法和E5检索器，取前5篇作为外部参考。 元认知层面，我们利用预训练的T5-large模型作为专家模型，利用sentence transformers的工具进行相似度评估。我们设定阈值为0.4作为基础判准，以保证精确度，同时设定最大迭代次数为5次。

Results and Analysis

Main Results

知乎

Method	Retr.	Multi.	Critic	HotpotQA				2WikiMultihopQA			
				EM	F1	Prec.	Rec.	EM	F1	Prec.	Rec.
Without retrieval (Closebook)											
Standard Prompting	-	-	-	20.0	25.8	26.4	28.9	21.6	25.7	24.5	31.8
Chain-of-Thought	-	-	-	22.4	34.2	33.9	46.0	27.6	37.4	35.8	44.3
With retrieval (BM25+E5)											
Standard RAG	✓	-	-	24.6	33.0	34.1	34.5	18.8	25.2	25.6	26.2
ReAct	✓	✓	-	24.8	41.7	42.6	44.7	21.0	28.0	27.6	30.0
Flare	✓	✓	-	29.2	42.4	42.8	43.0	28.2	39.8	40.0	40.8
IR-CoT	✓	✓	-	31.4	40.3	41.6	41.2	30.8	42.6	42.3	40.9
Self-Ask	✓	✓	-	28.2	43.1	43.4	44.8	28.6	37.5	36.5	42.8
Reflexion	✓	✓	✓	30.0	43.4	43.2	44.3	27.8	41.7	41.2	44.2
MetaRAG (ours)	✓	✓	✓	**37.8**†	**49.9**†	**52.1**†	**50.9**†	**42.8**†	**50.8**†	**50.7**†	**52.2**†

(1) MetaRAG在HotpotQA和2WikiMultiHopQA两个数据集上表现出优越性能，超越所有基线方法，证明元认知策略的有效性。相比于仅依赖自我批评的Reflexion，MetaRAG在所有评估指标上均有显著提升，这强调了元认知在提升答案准确性的价值。 (2) 模型配备自我批评机制显著提高了性能，表明赋予LLMs自我评估能力有助于它们更全面地审视自己的回应质量。MetaRAG在此基础上，还考虑了知识条件和多跳推理的精确性，能够定位并改正错误。

(3) 在2WikiMultiHopQA上，MetaRAG的改进尤为显著，相对于Reflexion，提升幅度分别为34.6%和26.0%。数据集特点------存在更多冲突性知识，显示MetaRAG能有效应对不一致情况，通过精细规划策略优化推理，提升精确性。

The Study of Monitoring Phase

为了探究监测对整个框架作用的效应，我们进行了两个实验，对比了不同种类的监测模型，并考察了参数k的不同设置。

Table 2: The comparison of various monitoring expert models with different parameter size (Param.) on 2WikiMultihopQA.

Expert model	Param.	EM	F1	Prec.	Rec.
Large Language Models					
LLaMA2-chat	13B	40.4	47.6	47.6	48.8
ChatGLM2	6B	39.8	48.8	48.5	50.5
Fine-tuned QA Models					
SpanBert-large	0.34B	42.0	50.4	50.3	51.8
T5-large	0.77B	42.8	50.8	50.7	52.2

我们研究了不同监测模型对整体框架的影响，特别比较了两种类型：一种是大型语言模型（LLMs如LLaMA2-chat和ChatGLM2），它们展示了零样本评估的能力；另一种是经过问答领域微调的模型（如SpanBERT-large和T5-large）。结果显示，尽管微调过的问答模型如T5-large在参数量上相对较少，但在监测阶段表现更优，这证明了它们能以高效的方式提供精准反馈。LLMs在元认知中自我监督的有效性也得到了体现，它们提升了模型的性能。至于具体比较，T5-large略胜于SpanBERT-large，这可能归因于生成模型在任务适应性上的优势。

我们在监测阶段研究了不同相似度阈值k对模型性能的影响。该阈值控制了元认知介入的难度，阈值越高，启用元认知的可能性越大。我们试验了0.2到0.8，每0.1步幅，在2WikiMultiHopQA上考察了元认知参与的比例和答案质量。 结果显示，当阈值为0.2，约有15%的问题会启动元认知，这使得性能相对于Reflexion有了约20%的提升。随着阈值的提升，需要元认知处理的问题比例逐渐增加，达到0.8时，这一比例升至84%。值得注意的是，最优性能并不出现在阈值最高的位置，反而在0.4时达到峰值。这提示我们并非所有问题都适合通过元认知来解决，过度思考有时可能反而干扰了直接有效的判断，这与人类通常的经验相符。

原文《Metacognitive Retrieval-Augmented Large Language Models》

发布于 2024-04-26 10:42 · IP 属地北京