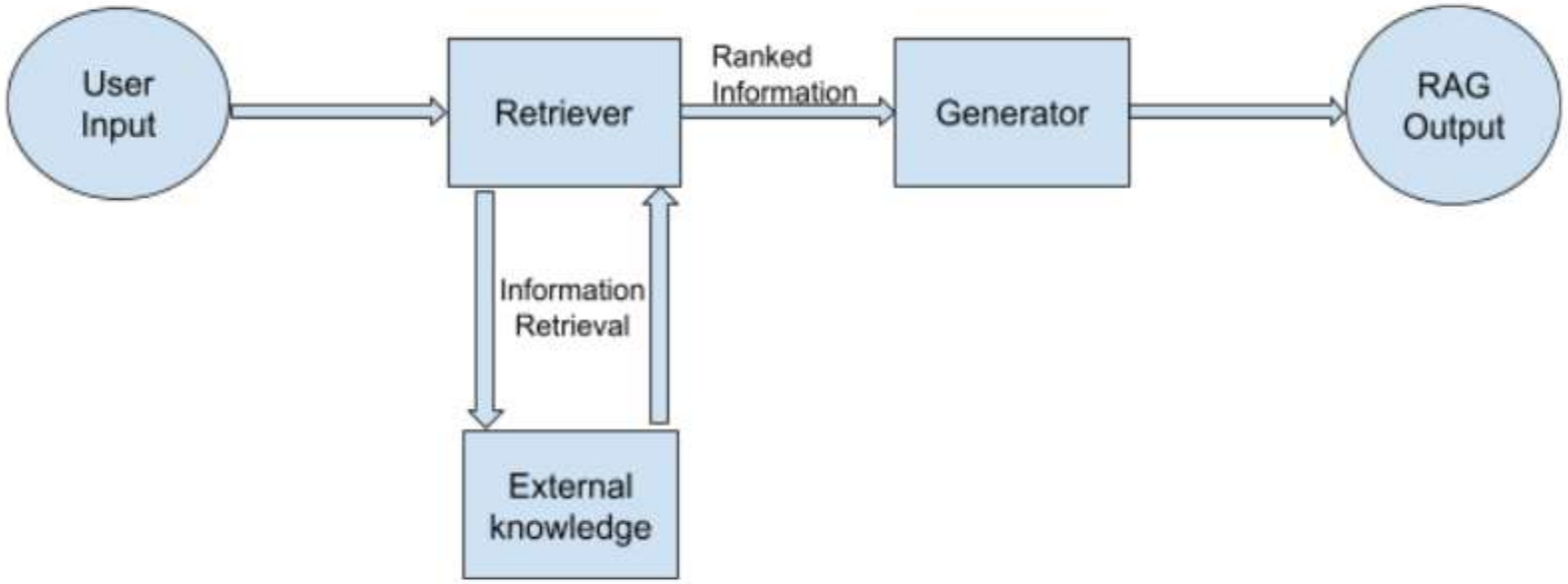
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

ACL 2025

CTRLA: ADAPTIVE RETRIEVAL-AUGMENTED GENERATION VIA INHERENT CONTROL

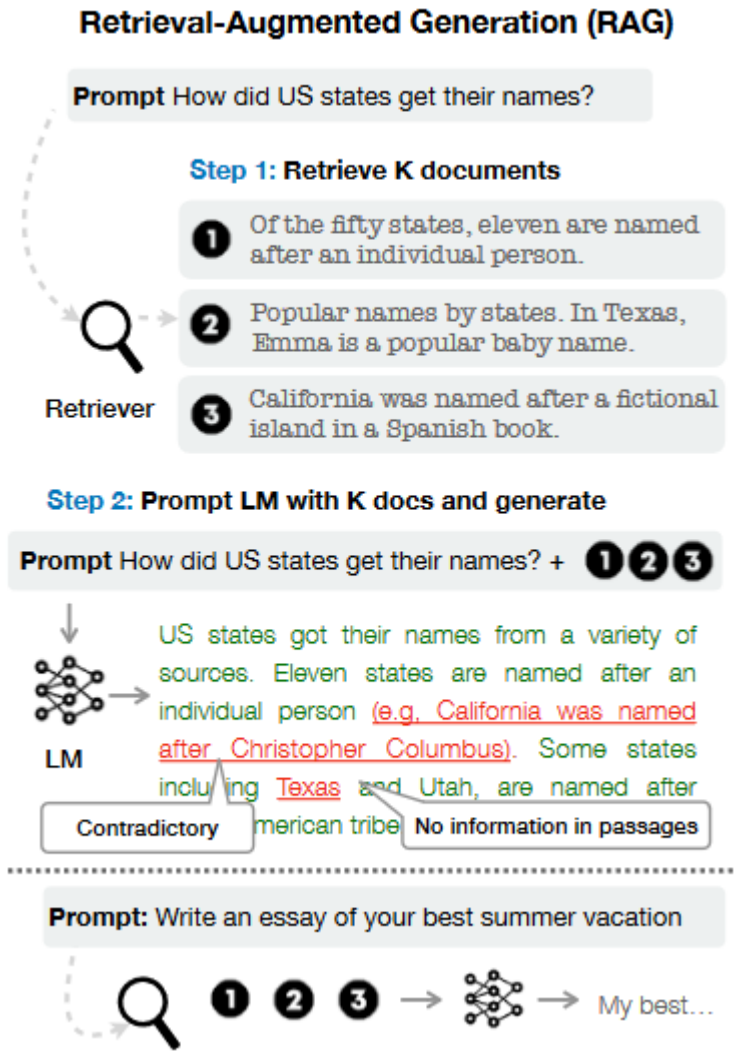
Huanshuo Liu[1,✉], **Hao Zhang**[1,✉,✉], **Zhijiang Guo**[1,✉], **Jing Wang**[2]
Kuicai Dong[1], **Xiangyang Li**[1], **Yi Quan Lee**[1], **Cong Zhang**[1], **Yong Liu**[1]
[1]Noah's Ark Lab, Huawei Technologies Co., Ltd
[2] Individual Researcher

背景



RAG的基本架构和流程

背景

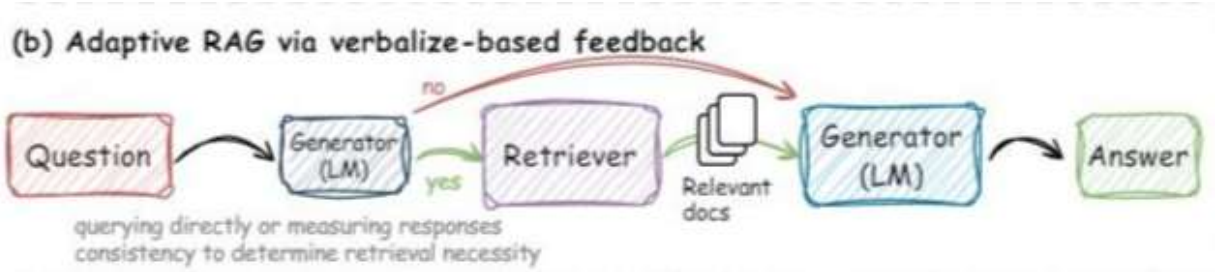


如何平衡模型的内部知识以及外部知识

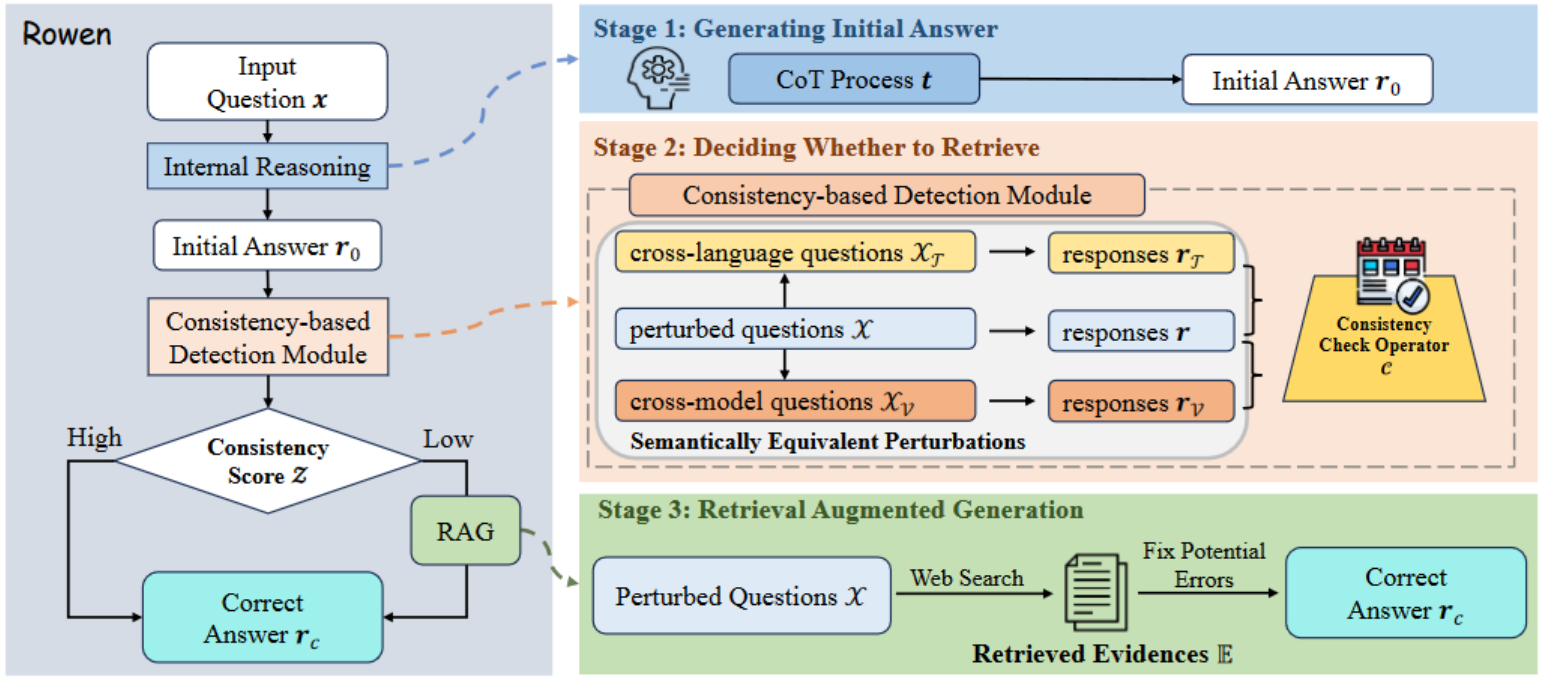
对于需要结合外部知识的问题，检索出相应知识后，可将检索出的知识与问题结合为prompt一同输入到大模型得到答案。

对于无需结合外部知识的问题，例如一些开放性问题和用内部知识可解决的问题，可能会引入无关信息，从而降低最终答案的质量。

背景-过往工作

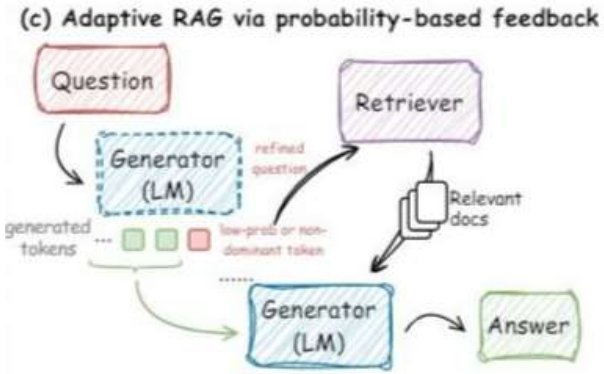


基于显式的模型的response feedback：例如，通过不同的方式扰动模型的一致性就代表模型知道答案，不需要搜prompt和问题，看模型的回答是否具有一致性。有一致性就代表模型知道答案，不需要搜

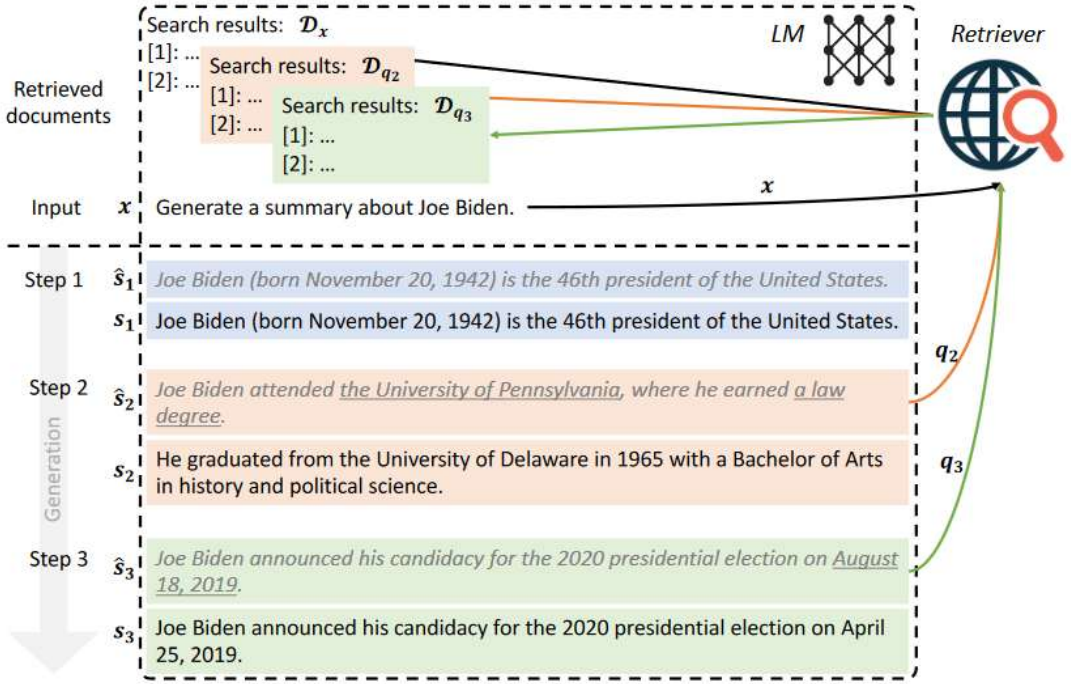


ROWEN

背景-过往工作

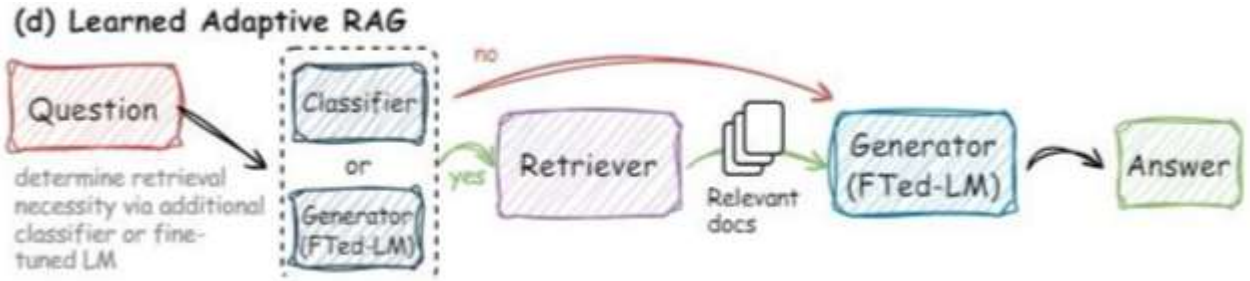


通过probability feed back：假定模型对于输出的token的概率本身就代表了他的信心。如果概率高，就代表信心足，就不需要搜



FLARE

背景-过往工作



通过Finetune或者额外训练一个classifier：这两种方式希望为模型注入自己判断是否要搜的能力。例如，判断问题类型，如果这类问题引入外部知识有益就搜，没什么帮助就不搜

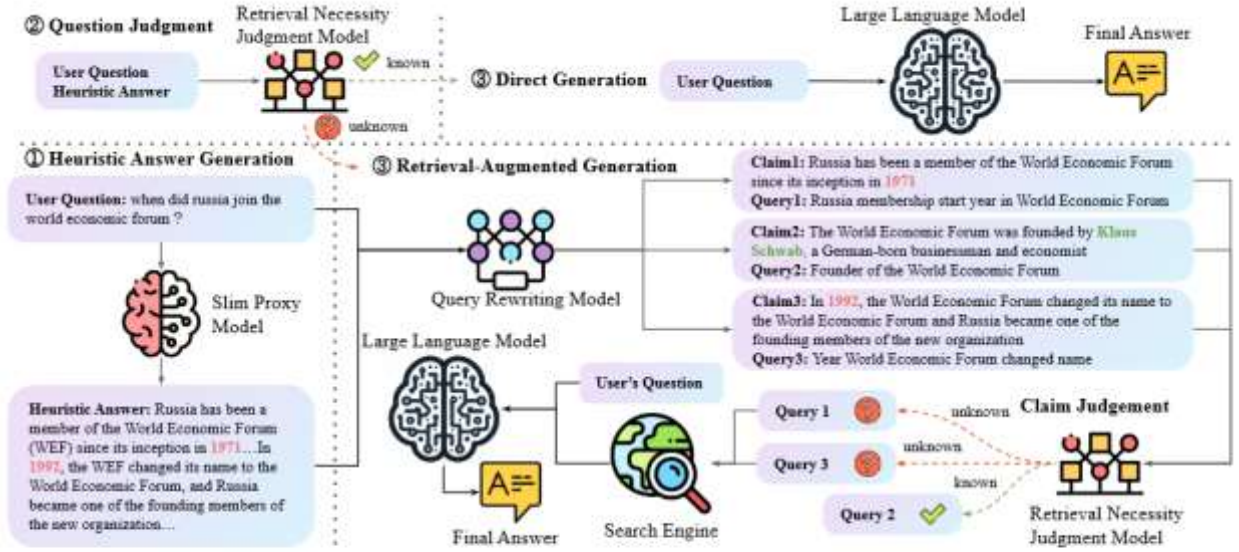
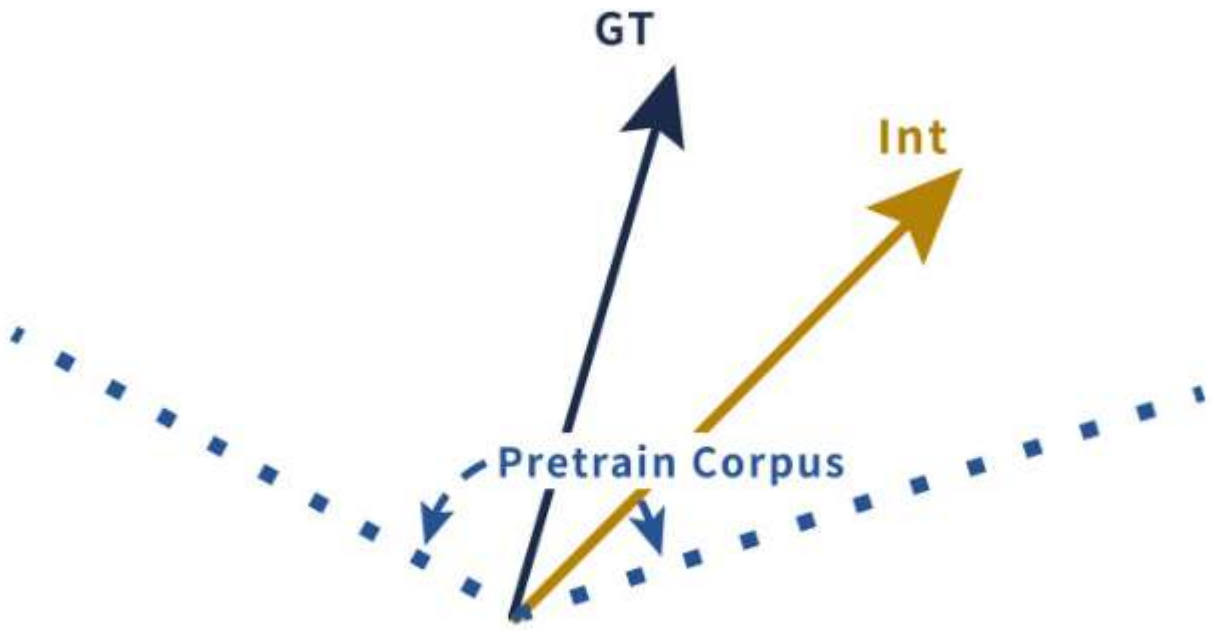


Figure 1: A display of the main process of SlimPLM. Solid lines with arrows represent the flow of data, while dashed lines with arrows signify control signals from the retrieval necessity judgment model. Step 1 and step 2 are mandatory in the pipeline, but step 3 involves choosing between direct generation and RAG.

Slim-PLM

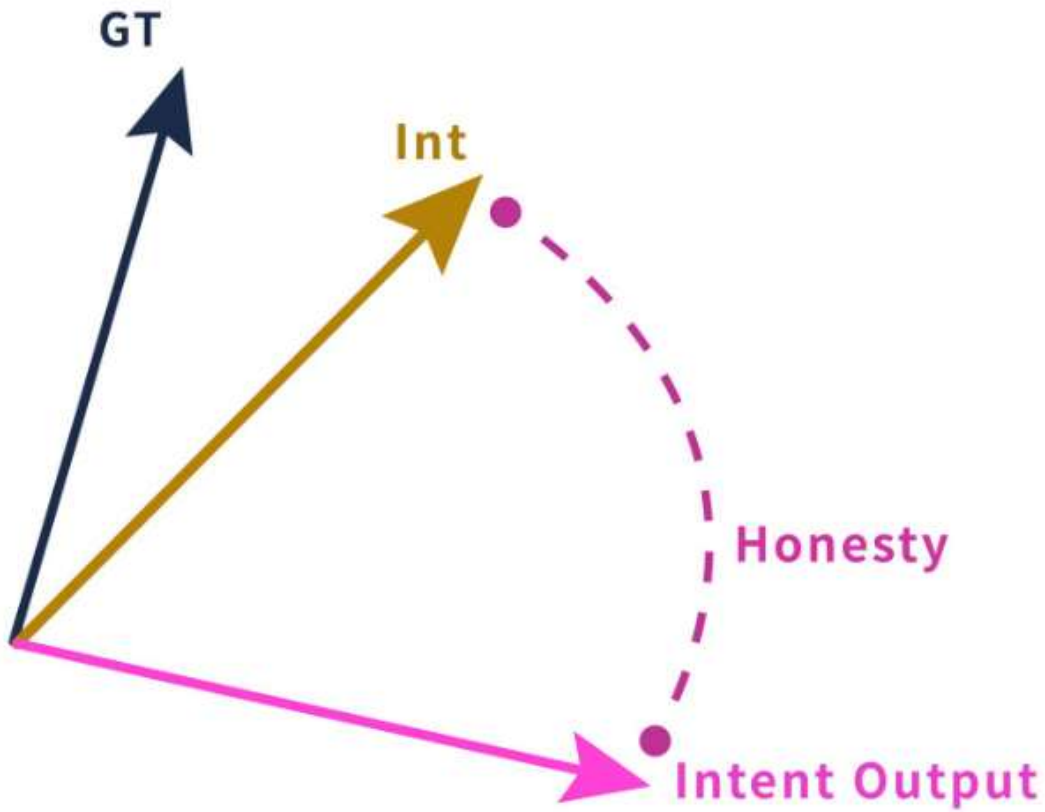
核心问题-输出正确答案的条件

1.GroundTruth=Internal Knowledge



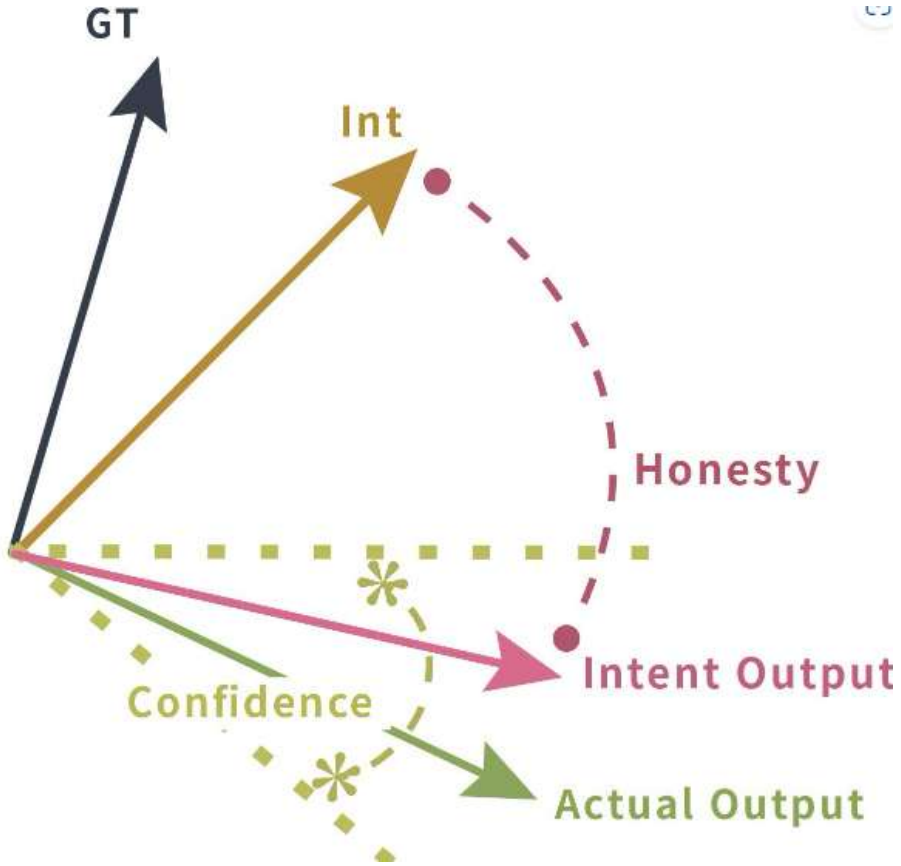
核心问题-输出正确答案的条件

2. Internal Knowledge =Intent Output



核心问题-输出正确答案的条件

3. Intent Output=Actual Output



核心问题-过往方法的问题

现有的方法都没有考虑LLM的输出与内部认知不对齐的情况

1.基于模型输出的一致性：大模型输出的一致性不好评估，且进行一次输出需要模型输出多次结果，响应延迟很高，还未排除一致但仍需要额外知识的情况
2.基于token输出概率：token的输出概率并不代表着模型的输出的confidence。且输出我不知道或缺乏信息等确定的回答，但仍然需要信息。
3.通过Finetune或者额外训练一个classifier：非常依赖数据的质量，而且额外训练的成本高

方法-解决思路

为了引导LLM朝向诚实，并监控其置信度，本文提取了LLM的表示空间中与诚实和置信度方向一致的特征。通过调整诚实方向以输出更多的诚实输出。同时，通过测量当前表示在置信特征上的投影来量化置信度，称这种方法为置信度监测。

方法-提取诚实、置信度方向向量

构建正反prompt，分别代表自信/不自信，诚实/不诚实
例如：你是一个诚实的人，请描述一下si

Prompt 3.1: Instruction for Honesty and Confidence Feature Extraction

[INST] Pretend you're a <honest/dishonest> | <confident/unconfident> person making statements about the world. **[/INST]** <a statement s_i>

$$\mathcal{I}_h^+ \oplus s_i \text{ and } \mathcal{I}_h^- \oplus s_i$$

方法-提取诚实、置信度方向向量

对于k个token,L个layer

$$\{\{\boldsymbol{r}_{i,k}^{l,+}\}_{k=1}^{n}\}_{l=1}^{L} \text{ and } \{\{\boldsymbol{r}_{i,k}^{l,-}\}_{k=1}^{n}\}_{l=1}^{L}$$

将向量相减得到诚实和置信度的方向向量

$$\boldsymbol{v}_{i,k}^{l} = \boldsymbol{r}_{i,k}^{l,+} - \boldsymbol{r}_{i,k}^{l,-}$$

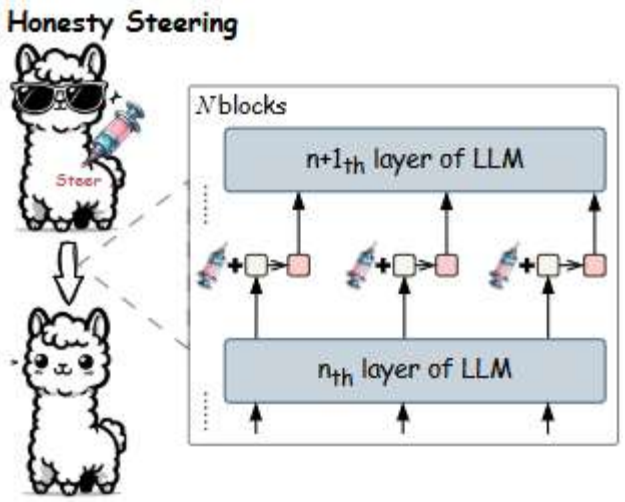
得到共计S组L层*k个方向向量

$$\{\{\{\boldsymbol{v}_{i,k}^{l}\}_{k=1}^{n}\}_{l=1}^{L}\}_{i=1}^{|\mathcal{S}|}$$

最后通过主成分分析提取出每层的诚实和置信度方向向量

$$\boldsymbol{v}_h = \{\boldsymbol{v}_h^l\}_{l=1}^{L}. \qquad \boldsymbol{v}_c = \{\boldsymbol{v}_c^l\}_{l=1}^{L}.$$

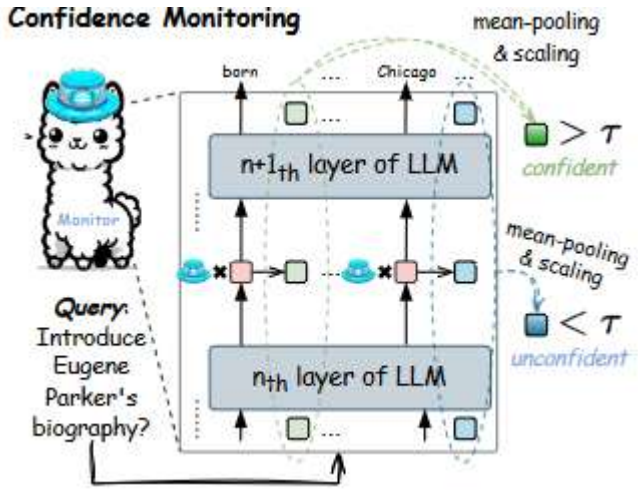
方法-向诚实倾斜



然后对于每一层，将诚实方向向量通过参数加到每一层的表示上来控制诚实倾斜

$$\hat{\boldsymbol{R}}_k = \boldsymbol{R}_k + \lambda \cdot \boldsymbol{v}_h = \{\boldsymbol{r}_k^l + \lambda \cdot \boldsymbol{v}_h^l \mid \forall\, l \in [1, \ldots, L]\},$$

方法-置信度估计



每层的表示通过点积，均值池化和缩放得到置信度分数

$$\tilde{m}_k = \texttt{meanpool}([m_{k,1}, \ldots, m_{k,L}]) = \texttt{meanpool}\left([\boldsymbol{r}_k^{l,\top} \cdot \boldsymbol{v}_c^l]_{l=1}^L\right),$$
$$\bar{m}_k = \texttt{scale}([\tilde{m}_0, \ldots, \tilde{m}_k])[-1] - \tau,$$

每个token得到一个均质池化后的置信度分数，若大于0则表示他的表示方向偏向于有自信的方向，否则反之。

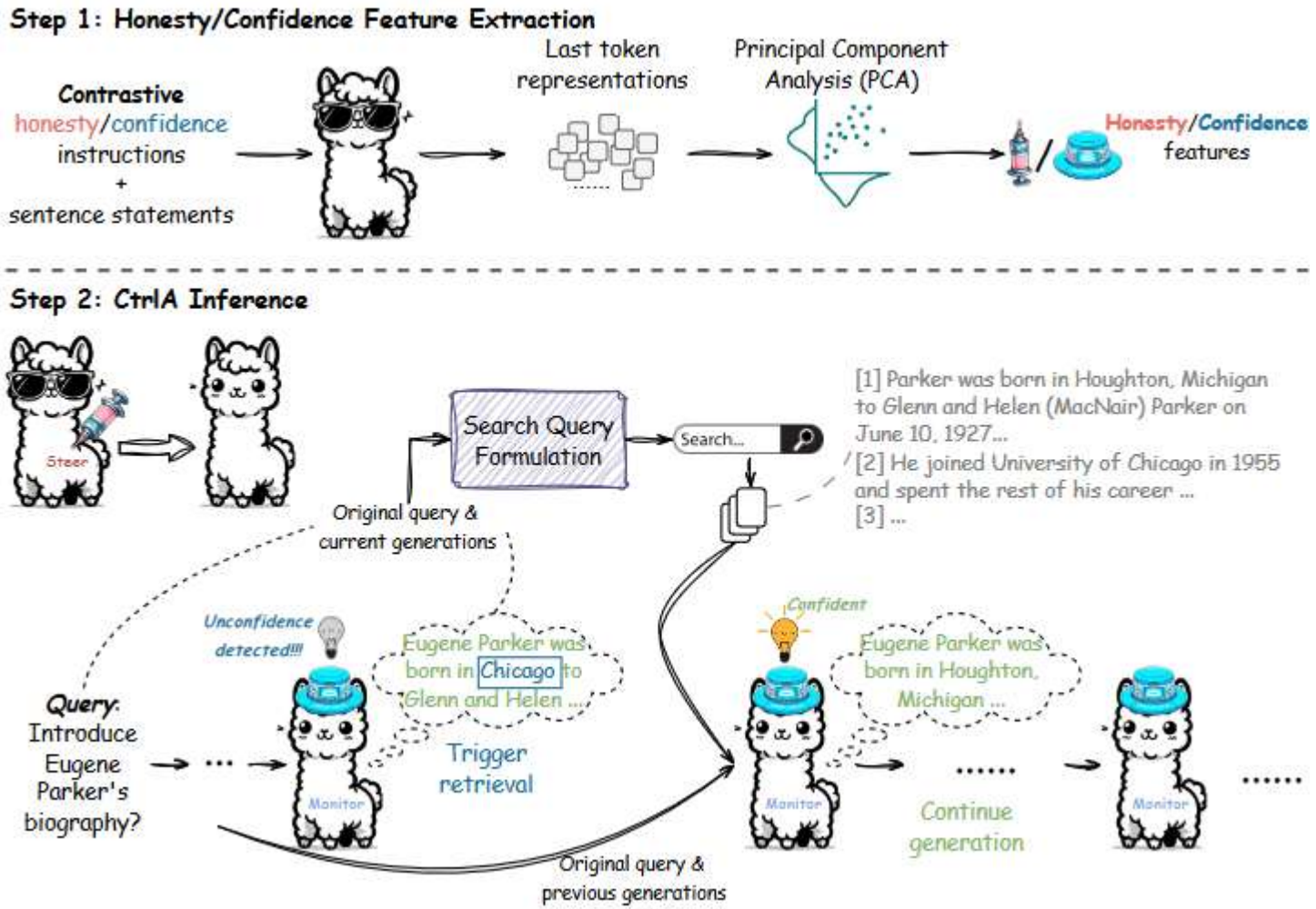
若一个段落有哪个词的阈值小于0，则出发检索机制。

方法-查询构建

上下文增强查询（CAQ）：将原始查询与当前生成片段结合，掩码掉“新且不自信的 token”（避免噪声与意图偏移），生成查询q_t = [原始查询；掩码后片段]。定义这个新且不自信的token是没有出现在q和之前生成的输出片段中且自信度分数小于零

目标验证查询（TVQ）：让 LLM 基于原始查询和当前生成片段，生成格式规范的验证型查询（如“验证‘Eugene Parker 1955 年加入芝加哥大学’是否正确”），适配密集检索器（如 BGE）对查询格式的要求。也就是给定查询和当前生成片段结合进pormot来引导大模型生成查询来验证其正确性

方法


Step 1: Honesty/Confidence Feature Extraction
Contrastive
honesty/confidence
instructions
+
sentence statements
Last token
representations
Principal Component
Analysis (PCA)
Honesty/Confidence
features
Step 2: CtrlA Inference
Steer
Search Query
Formulation
Search...
[1] Parker was born in Houghton, Michigan to Glenn and Helen (MacNair) Parker on June 10, 1927...
[2] He joined University of Chicago in 1955 and spent the rest of his career ...
[3] ...
Original query &
current generations
Unconfidence
detected!!!
Eugene Parker was born in Chicago to Glenn and Helen ...
Trigger
retrieval
Confident
Eugene Parker was born in Houghton, Michigan ...
Continue
generation
Query:
Introduce
Eugene
Parker's
biography?
Monitor
Original query &
previous generations

实验设置

任务类型	数据集	特点	评估指标
短文本 QA	TriviaQA	包含常识、事实类问题	准确率
短文本 QA	PopQA	侧重长尾、冷门事实问题	准确率
长文本 QA	ASQA	多答案、需要详细解释的问题	str-em、Rouge-L
长文本 QA	Bio	传记类问题，需要生成连贯段落	FactScore、Rouge-L
多跳 QA	2WikiMultihopQA	需要多步推理的复杂问题	EM、F1、检索频率
多跳 QA	HotpotQA	包含干扰信息的多跳推理问题	EM、F1
时效性 QA	FreshQA	包含时间敏感问题（分快慢变化两类）	宽松准确率、严格准确率

实现与检索设置

1. **核心模型**：主要使用 Mistral-7B-Instruct-v0.1 作为基础 LLM，同时在鲁棒性实验中验证了 LLaMA2-7B/13B、Vicuna-13B 等模型
2. **检索系统**：
   - 稀疏检索：采用 BM25 算法
   - 密集检索：使用 BGE-large-en-v1.5 作为检索器
   - 文档库：2018 版维基百科完整语料（约 600 万文档）+ 针对长尾问题补充的网页检索结果

baseline

1. **无检索（wo-RAG）**：仅使用 LLM 自身知识生成回答
2. 固定句子数检索
3. 固定长度检索
4. 查询分解RAG

实验一

Table 1: Overall results of short-form QA. $^{\diamond}$ is our reproduced results. ‡ denotes results in the corresponding work.

Method	TriviaQA	PopQA
wo-RAG$^{\diamond}_{7B}$	53.8	25.7
SR-RAG$^{\diamond}_{7B}$	62.7	51.9
FL-RAG$^{\diamond}_{7B}$	60.8	28.1
FS-RAG$^{\diamond}_{7B}$	54.3	26.9
QD-RAG$^{\diamond}_{7B}$	52.3	29.4
FLARE$^{\diamond}_{7B}$	<u>72.4</u>	48.3
Self-RAG$^{\ddagger}_{7B}$	66.4	54.9
Self-RAG$^{\ddagger}_{13B}$	69.3	55.8
RQ-RAG$^{\ddagger}_{7B}$	-	<u>57.1</u>
QC-RAG$^{\ddagger}_{11B}$	58.2	-
CTRLA$_{7B}$	**76.4**	**61.8**

实验一

Table 2: Overall results of long-form QA. $^{\diamond}$ is our reproduced results. ‡ denotes results in the corresponding work.

Method	ASQA					Bio
	str-em	R-L	EM	F1	mau	FS
wo-RAG$^{\diamond}_{7B}$	18.8	33.7	8.7	13.7	23.8	41.9
SR-RAG$^{\diamond}_{7B}$	32.4	34.9	18.7	25.1	54.7	78.6
FL-RAG$^{\diamond}_{7B}$	24.4	34.4	11.2	16.7	26.5	56.9
FS-RAG$^{\diamond}_{7B}$	25.9	32.9	11.3	16.9	44.8	57.5
QD-RAG$^{\diamond}_{7B}$	18.1	18.6	8.4	12.3	-	22.4
FLARE$^{\diamond}_{7B}$	29.9	35.2	16.2	22.2	50.4	74.8
Self-RAG$^{\ddagger}_{7B}$	30.0	35.7	-	-	74.3	81.2
Self-RAG$^{\ddagger}_{13B}$	31.7	37.0	-	-	71.6	80.2
CTRLA$_{7B}$	**37.0**	**38.5**	**20.4**	**27.3**	**79.2**	**83.4**

Table 3: Overall results of multi-hop QA. † means results reported by DRAGIN/SeaKR. ‡ denotes results in the corresponding work.

Method	2WMQA			HQA		
	EM	F1	Freq	EM	F1	Freq
wo-RAG$^{\dagger}_{7B}$	14.6	22.3	0.00	18.4	27.5	0.00
SR-RAG$^{\dagger}_{7B}$	16.9	25.5	1.00	16.4	25.0	1.00
FL-RAG$^{\dagger}_{7B}$	11.2	19.2	3.34	14.6	21.1	3.81
FS-RAG$^{\dagger}_{7B}$	18.9	26.5	3.83	21.4	30.4	4.15
FLARE$^{\dagger}_{7B}$	14.3	21.3	0.94	14.9	22.1	1.07
Self-RAG$^{\ddagger}_{7B}$	4.6	19.6	-	6.8	17.5	-
DRAGIN$^{\ddagger}_{7B}$	22.4	39.0	2.84	23.7	34.2	3.02
SeaKR$^{\ddagger}_{7B}$	30.2	36.0	-	27.9	39.7	-
CTRLA$_{7B}$	**36.9**	**43.7**	2.01	**34.7**	**44.9**	3.28

实验二

实现了无检索情况下观察诚实探针的效果（TruthfulQA），还对比了在prompt中引导诚实输出的情况，有效，但没那么好：

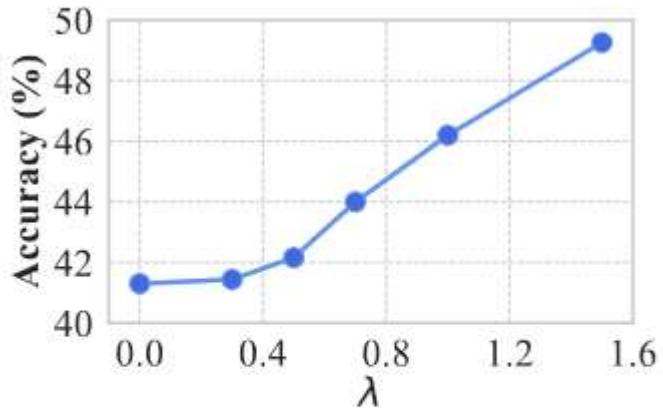


Figure 2: Effects of honesty steering on TruthfulQA.

λ	PopQA	ASQA				2Wiki	
	Acc (%)	str-em	R-L	F1	mau	EM	F1
$\lambda = 0.0$	58.5	36.8	38.1	27.0	76.5	34.9	41.5
$\lambda = 0.3$	**61.8**	**37.0**	**38.5**	**27.3**	**79.2**	**36.9**	**43.7**
HonP	60.2	36.8	38.3	27.0	71.5	34.3	41.0

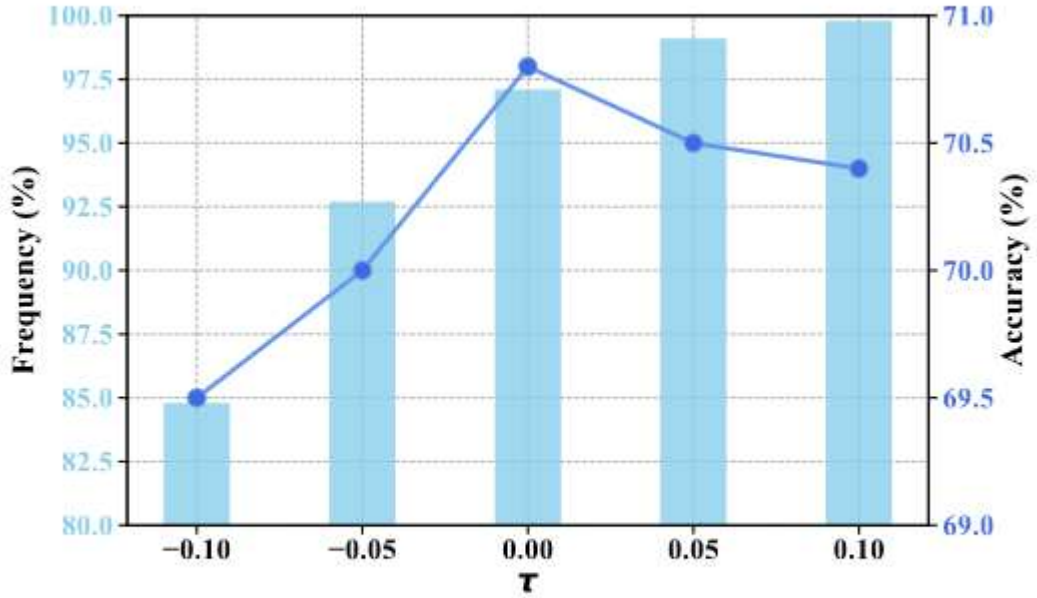


Figure 4: Effects of different choices of τ on TriviaQA.

实验四

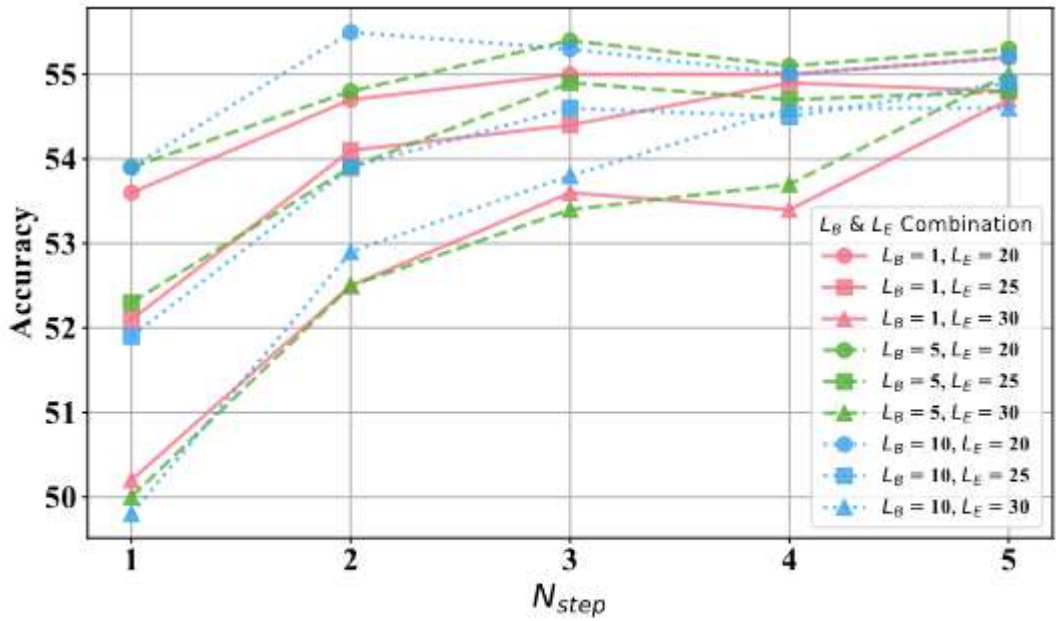


Figure 5: Impacts of honesty steering with respect to the layers and steps on TriviaQA.

Thanks