中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

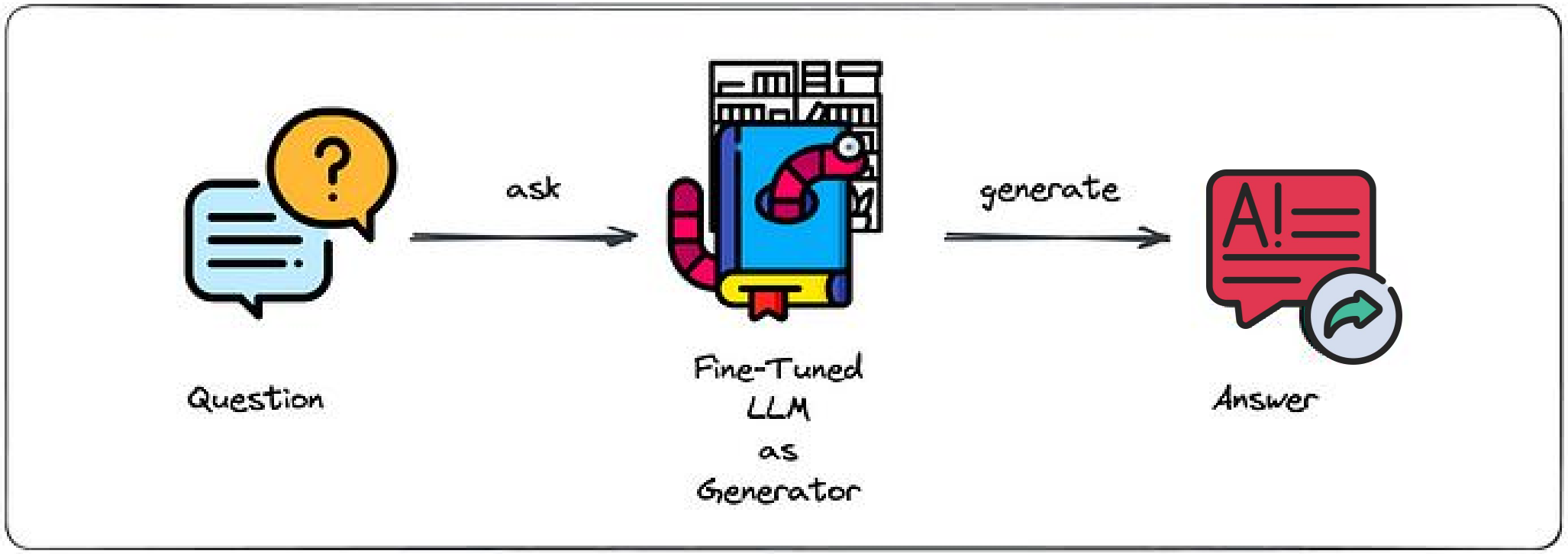
智能算法安全重点实验室
Key Laboratory of AI Safety, CAS

ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability

ICLR 2025, Spotlight
ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, Han Li
2025/06/10

What is Hallucination

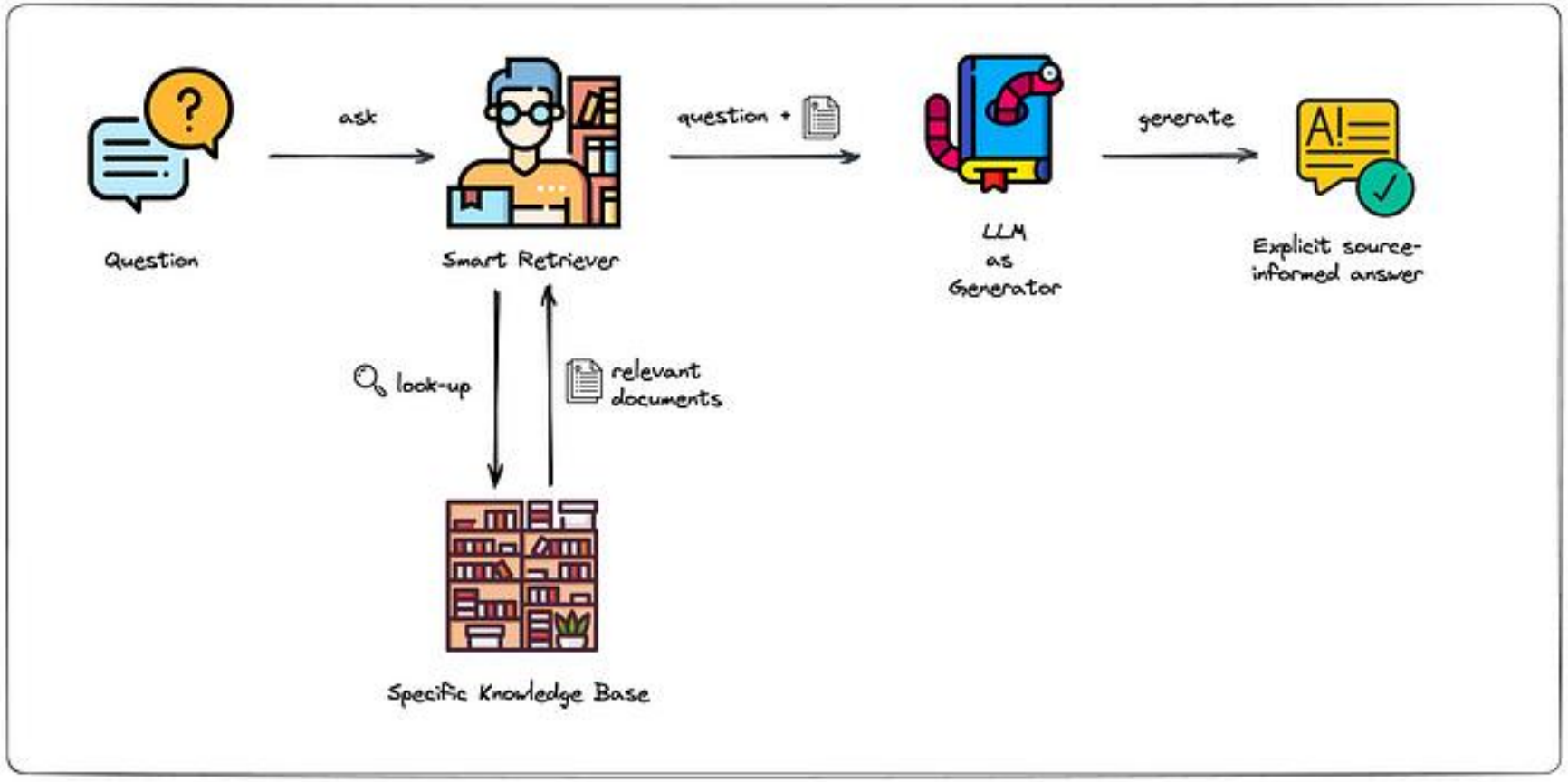
- **Hallucination 示例**



问：马达加斯加的首都是什么？答：安巴拉沃 ✗

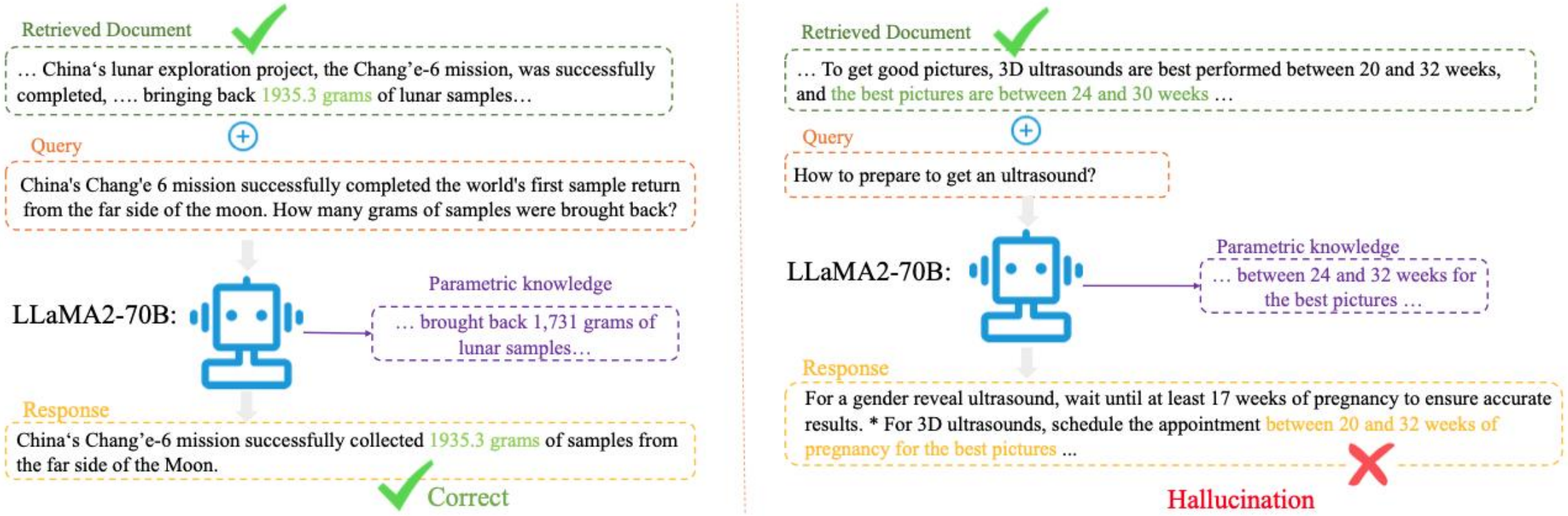
RAG for Hallucination Mitigation

- **RAG 示例**



问：马达加斯加的首都是什么？答：塔那那利佛 ✓

Analyzing RAG Limitations



尽管直接提供给了 LLM 正确答案，LLM 甚至都不会抄

Transformer 结构示意图

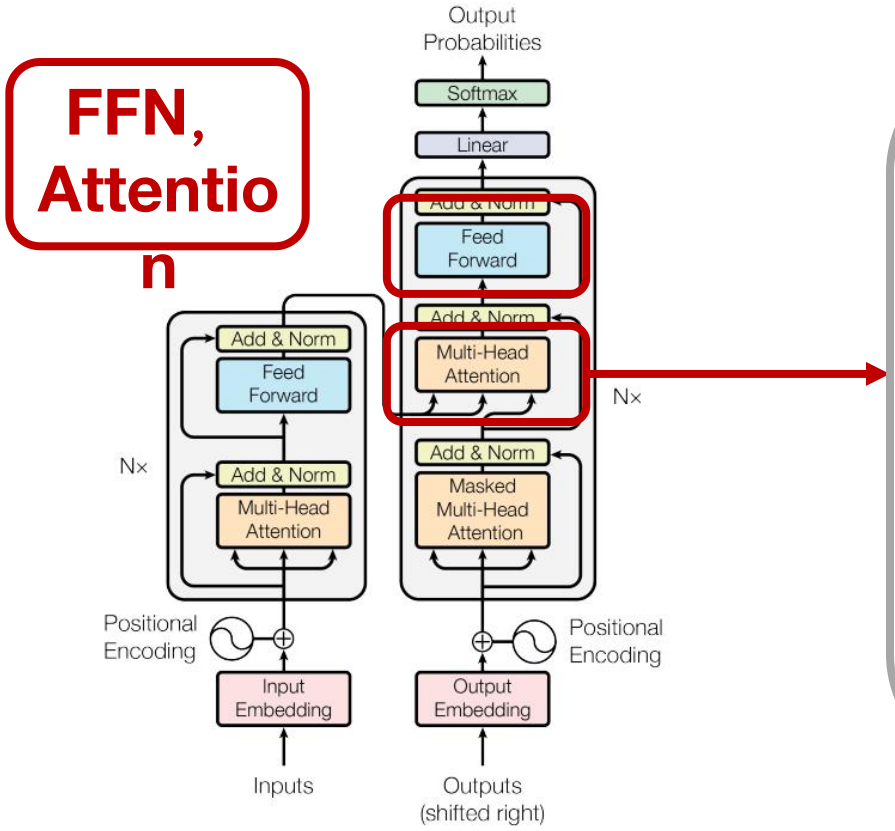
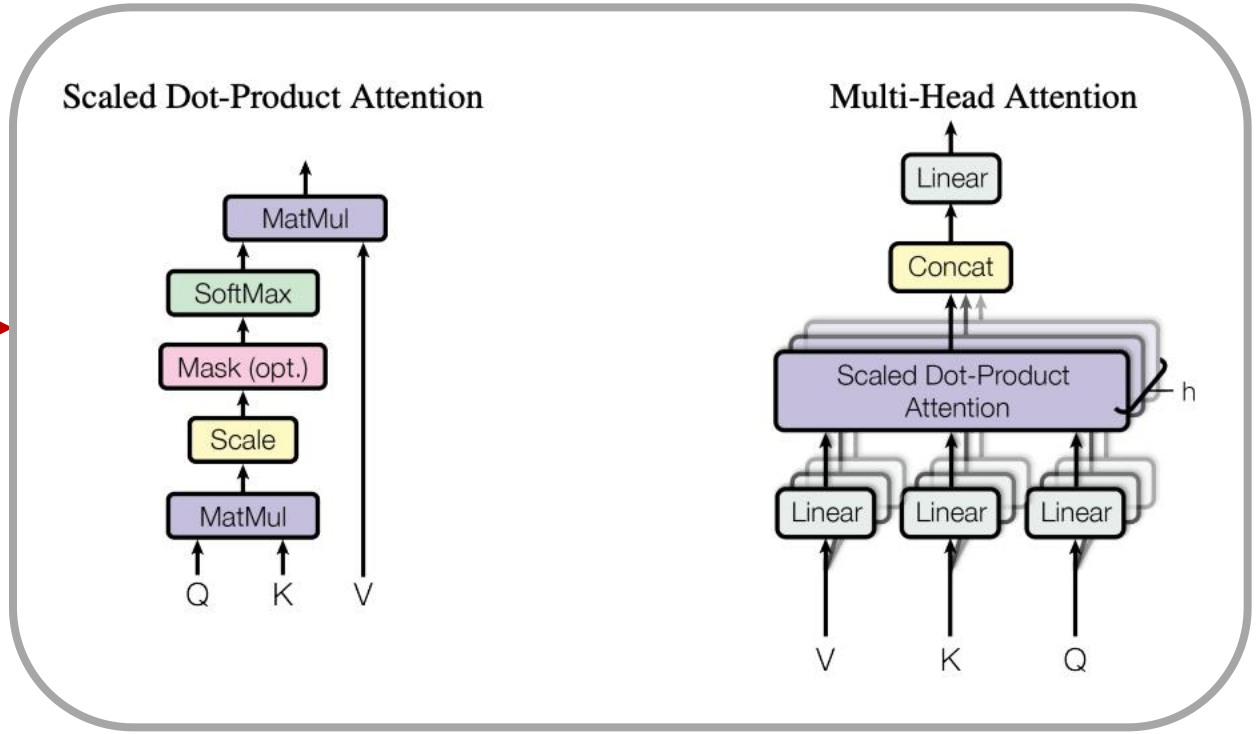


Figure 1: The Transformer - model architecture.

Transformer结构图



Attention计算模块图

Attention 计算公式

- **Attention 计算**

$$\mathrm{Attn}^{l,h}\left(\boldsymbol{X}_{\leq i}^{l-1}\right)=\sum_{j\leq i}a_{i,j}^{l,h}\boldsymbol{x}_j^{l-1}\boldsymbol{W}_V^{l,h}\boldsymbol{W}_O^{l,h}=\sum_{j\leq i}a_{i,j}^{l,h}\boldsymbol{x}_j^{l-1}\boldsymbol{W}_{OV}^{l,h}$$

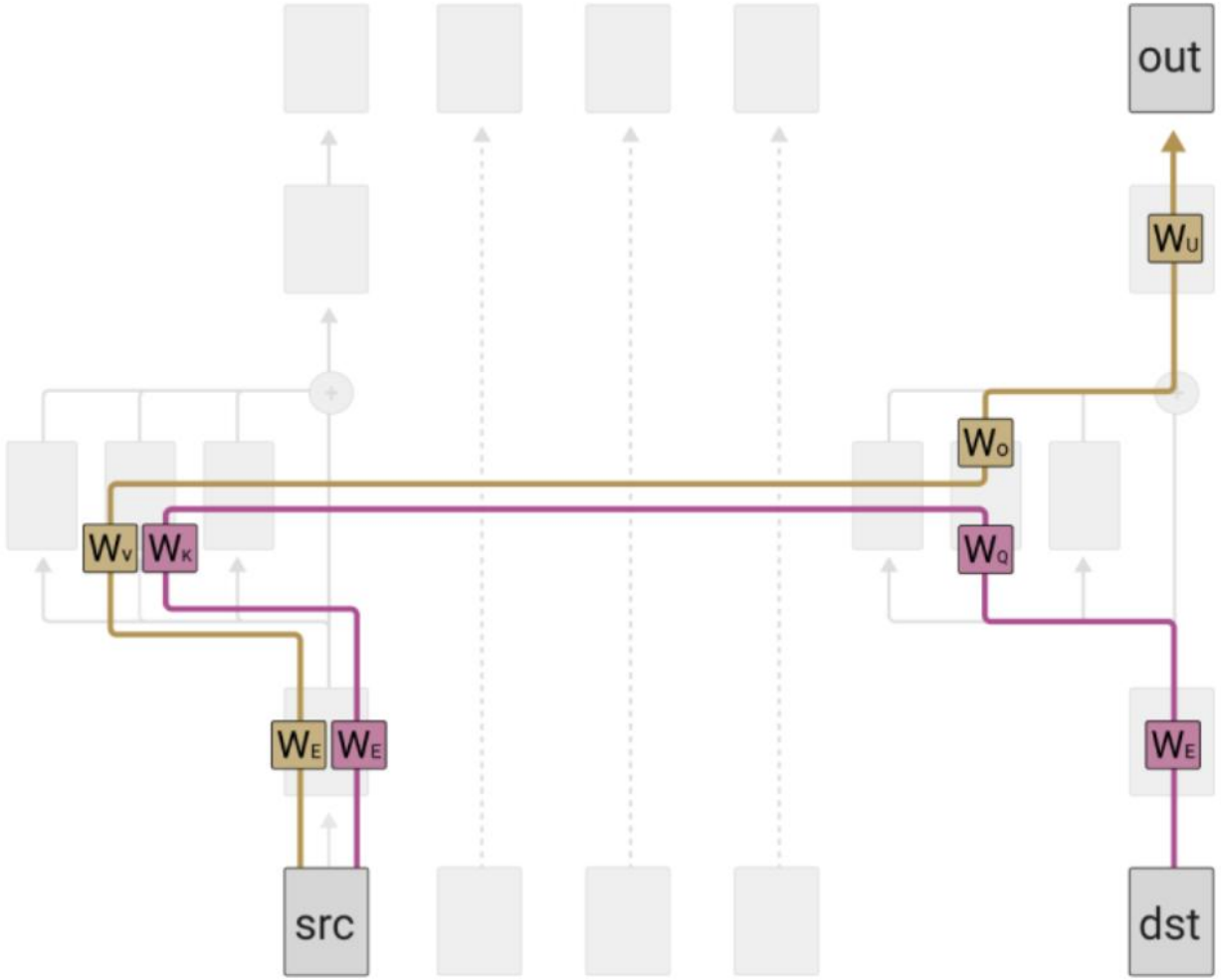
$$\boldsymbol{a}_i^{l,h}=\mathrm{softmax}\left(\frac{\boldsymbol{x}_i^{l-1}\boldsymbol{W}_Q^{l,h}\left(\boldsymbol{X}_{\leq i}^{l-1}\boldsymbol{W}_K^{l,h}\right)^\top}{\sqrt{d_k}}\right)=\mathrm{softmax}\left(\frac{\boldsymbol{x}_i^{l-1}\boldsymbol{W}_{QK}^{h}\boldsymbol{X}_{\leq i}^{l-1\top}}{\sqrt{d_k}}\right)$$

QK、OV Circuits

- **QK、OV Circuits**
  - The QK circuit is formed by tracing the computation of a query and key vector up to their attention head, where they dot product to create a bilinear form.
  - The OV circuit is created by tracing the path computing a value vector and continuing it through up to the logits.

A Mathematical Framework for Transformer Circuits

QK、OV Circuits



The **OV ("output-value") circuit** determines how attending to a given token affects the logits.

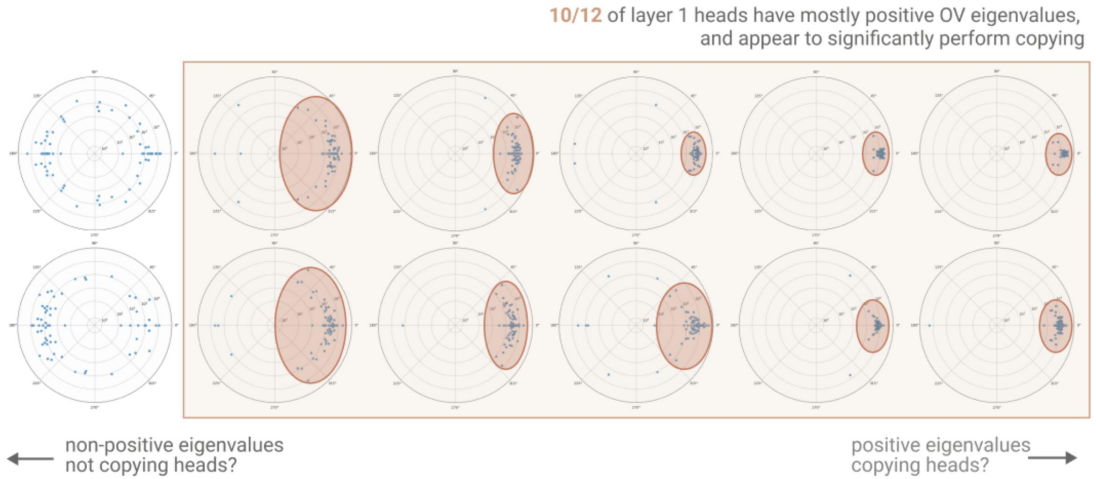
$$W_U W_O W_V W_E$$

The **QK ("query-key") circuit** controls which tokens the head prefers to attend to.

$$W_E^T W_Q^T W_K W_E$$

QK、OV Circuits

- **OV Circuits 存在 Copy 行为 （Copying Head)**
  - ”Copying behavior is widespread in OV matrices and arguably one of the most interesting behaviors.”
  - 启发式研究：
    , most of eigenvalues



A Mathematical Framework for Transformer Circuits

核心预备知识及核心问题

- **核心预备知识**
  1. Attention Head 存在 Copying Head
  2. 深层的 FFN 中存储了 LLM 的内部知识，即 Parametric Knowledge

- **在输入文档包含正确答案的情况下，Hallucination是否来源于：**
  1. 问题1：LLM 忽视了外部知识。即 Copying Head 使用不足。
  2. 问题2：LLM 过分使用了内部知识。即 FFN 使用过度。

数据集介绍

Experiment Setting: We conduct experiments on the Llama2-7B-chat model (Touvron et al., 2023) using the training set of RAGTruth dataset (Niu et al., 2024), a high-quality, manually annotated dataset for RAG hallucinations (Details in Section 5.1). Each data point in RAGTruth consists of a query $\mathbf{q}$, retrieved context $\mathbf{c}$, response $\mathbf{r}$, and the hallucination label h (where 0 is truth and 1 is hallucination). During generation, the input to the LLM f is a sequence of tokens $\mathbf{t} = \langle t_1, t_2, \ldots, t_n \rangle$, including the query $\mathbf{q} = \langle t_1, \ldots, t_q \rangle$, retrieved context $\mathbf{c} = \langle t_{q+1}, \ldots, t_c \rangle$, and a partial generated response $\hat{\mathbf{r}} = \langle t_{c+1}, \ldots, t_n \rangle$.


```
{
  "id": "1472",
  "source_id": "11316",
  "model": "mistral-7B-instruct",
  "temperature": 0.925,
  "labels": [
    {
      "start": 219,
      "end": 229,
      "text": "Gaza Strip",
      "meta": "HIGH INTRO OF NEW INFO\nIt is not mentioned in the original source that Gaza St
      "label_type": "Evident Baseless Info"
    }
  ],
  "split": "train",
  "quality": "good",
  "response": "The Palestinian Authority has officially become the 123rd member of the Interna
}
```

External Context Score (ECS)

- 问题1的检测方法：
  - 计算 response r 和 context c 中 attention 最高的文本的语义相似度
- 计算指标：ECS

For the last token t_n, the attention weights on the context are $\boldsymbol{a}_{n,q:c}^{l,h}$, where $\boldsymbol{a}_n^{l,h}$ is obtained from Equation 8. We select the top $k\%$ tokens with the highest attention scores as attended tokens:

$$\mathcal{I}_n^{l,h} = \arg\operatorname{top}_{k\%}(\boldsymbol{a}_{n,q:c}^{l,h}). \tag{2}$$

Inspired by (Luo et al., 2024; Chen et al., 2024a), which validated that the hidden states of LLM can serve as token semantic representations, we compute the token-level **E**xternal **C**ontext **S**core (ECS) based on the cosine-similarity between the mean-pooling of the last layer hidden states of attended tokens and the hidden state of token t_n:

$$\mathcal{E}_n^{l,h} = \frac{\boldsymbol{e} \cdot \boldsymbol{x}_n^L}{\|\boldsymbol{e}\|\|\boldsymbol{x}_n^L\|}, \quad \boldsymbol{e} = \frac{1}{|\mathcal{I}_n^{l,h}|} \sum_{j \in \mathcal{I}_n^{l,h}} \boldsymbol{x}_j^L. \tag{3}$$

The response-level ECS is the average of token-level scores:

$$\mathcal{E}_{\mathbf{r}}^{l,h} = \frac{1}{|\mathbf{r}|} \sum_{t \in \mathbf{r}} \mathcal{E}_t^{l,h}. \tag{4}$$

Parametric Knowledge Score (PKS)

- 问题2的检测方法：
  - 计算 输入内容 与经过FFN层后的 输出内容的差异。
- 计算指标：PKS

Parametric Knowledge: Considering FFNs store parametric knowledge, to assess how LLM use Parametric Knowledge (as discussed in Section 2.1), we use the LogitLens to map residual stream states before (i.e., $\boldsymbol{x}_n^{\text{mid},l}$, calculated from Equation 9) and after the FFN layer (i.e., $\boldsymbol{x}_n^l$, calculated from Equation 10) to vocabulary distributions. The difference in vocabulary distributions represents the parametric knowledge added by the FFN layer to the residual stream, which is measured by Jensen-Shannon divergence (JSD), gives the token-level **P**arametric **K**nowledge **S**core (PKS):

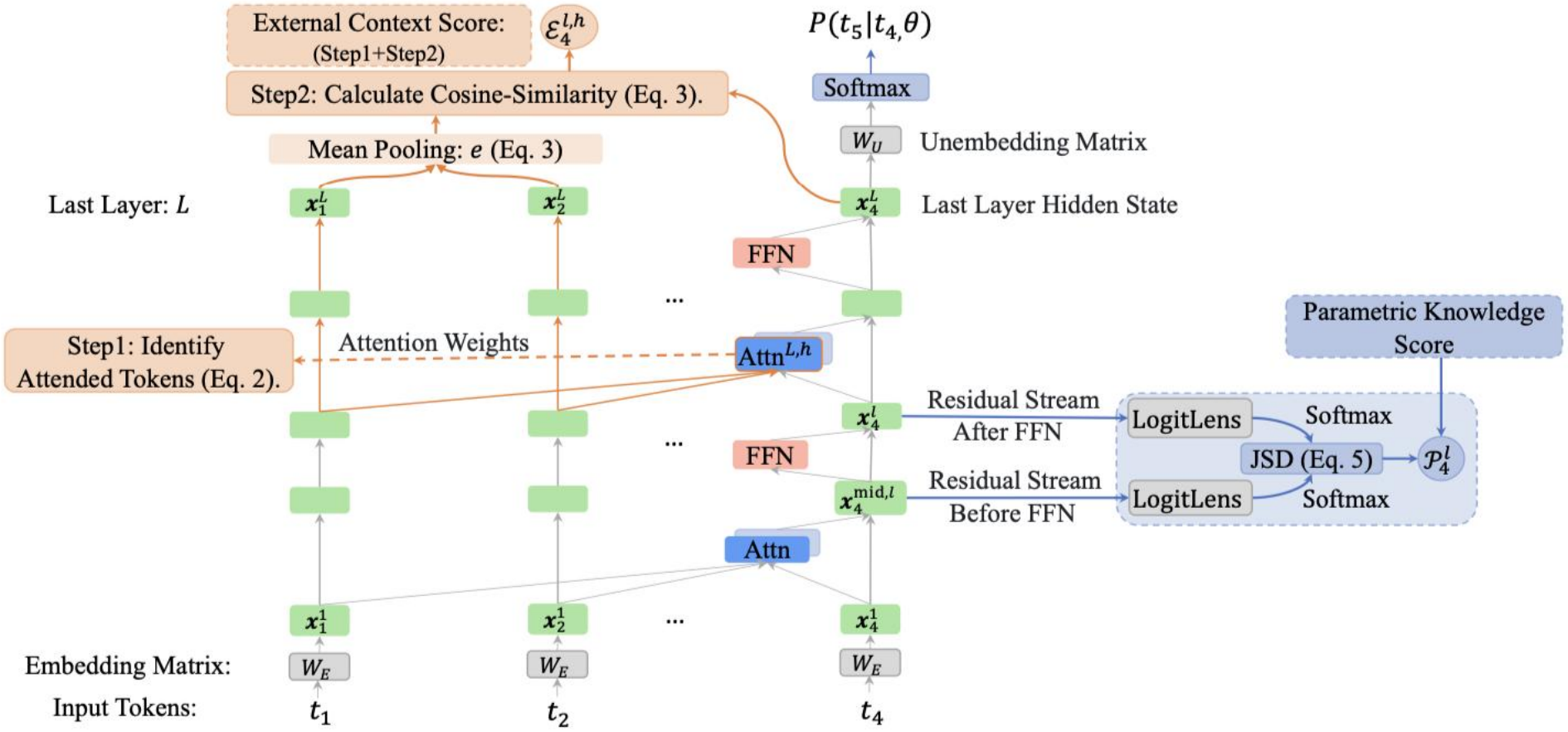
$$\mathcal{P}_n^l = \text{JSD}\left(q(\boldsymbol{x}_n^{\text{mid},l}) \parallel q(\boldsymbol{x}_n^l)\right), \tag{5}$$

where $q(\boldsymbol{x}) = \text{softmax}(\text{LogitLens}(\boldsymbol{x}))$. The response-level PKS is the average of token-level scores:

$$\mathcal{P}_{\mathbf{r}}^l = \frac{1}{|\mathbf{r}|}\sum_{t\in\mathbf{r}} \mathcal{P}_t^l. \tag{6}$$

Pipeline

- **Pipeline**



Research Questions

- **RQ1：** Relationship Between LLM Utilization of External Context, Parametric Knowledge, and Hallucinations
- **RQ2：** Can the relationship identified in RQ1 be validated from a causal perspective?
- **RQ3：** Hallucination Behavior Analysis from the Parametric Knowledge Perspective

RQ1 实验结果

- ECS 实验结果

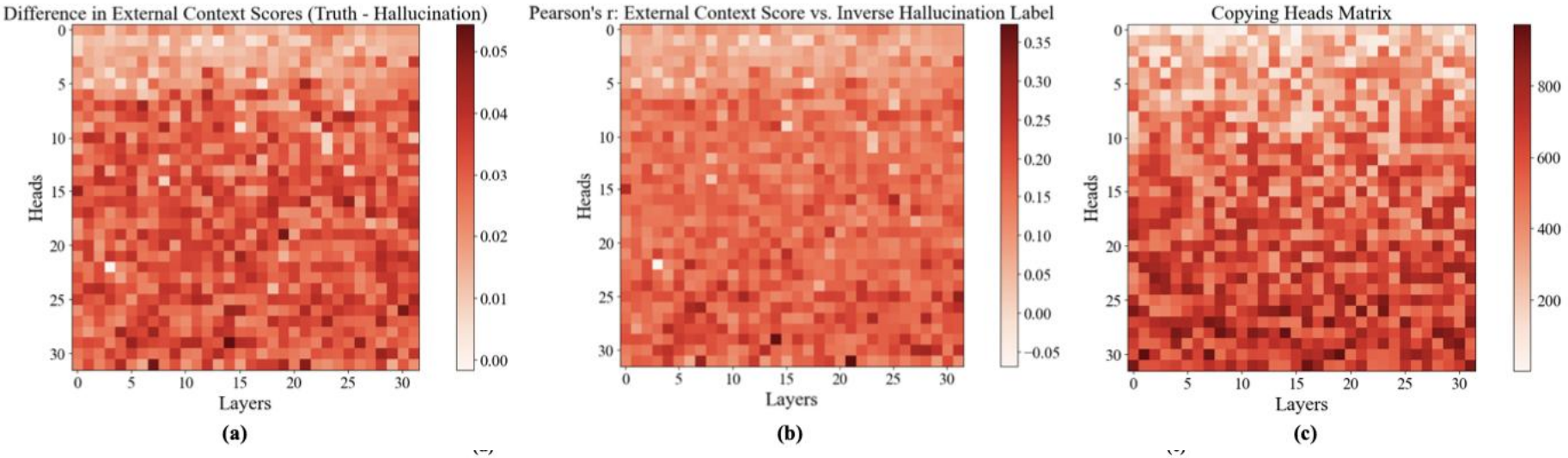


Figure 4: Relationship Between LLM Utilization of External Context, Parametric Knowledge, and Hallucinations. Top shows the internal mechanism of LLM's utilization of external context and the occurrence of hallucinations, where the Pearson correlation coefficient between (c) and (a) is 0.41, and between (c) and (b) is 0.46, indicating correlations among them. Bottom illustrates the

Fig a:

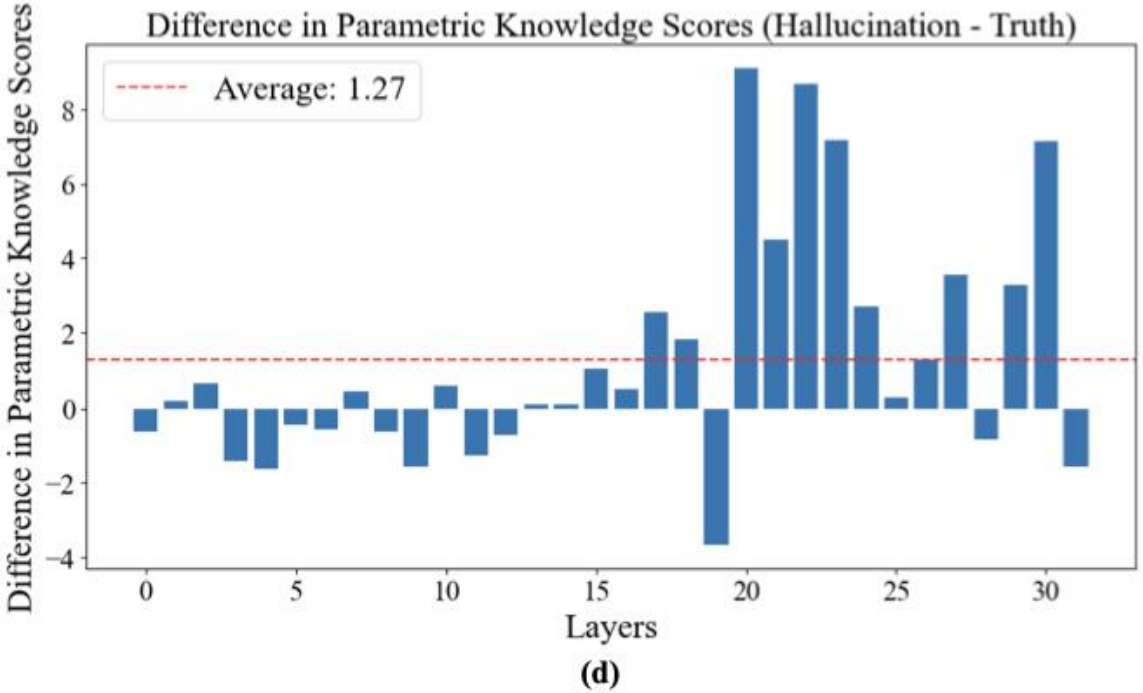
$$\Delta\mathcal{E}^{l,h} = \mathcal{E}_T^{l,h} - \mathcal{E}_H^{l,h} = \frac{1}{|\mathcal{D}^T|}\sum_{\mathbf{r}\in\mathcal{D}^T}\mathcal{E}_{\mathbf{r}}^{l,h} - \frac{1}{|\mathcal{D}^H|}\sum_{\mathbf{r}\in\mathcal{D}^H}\mathcal{E}_{\mathbf{r}}^{l,h}.$$

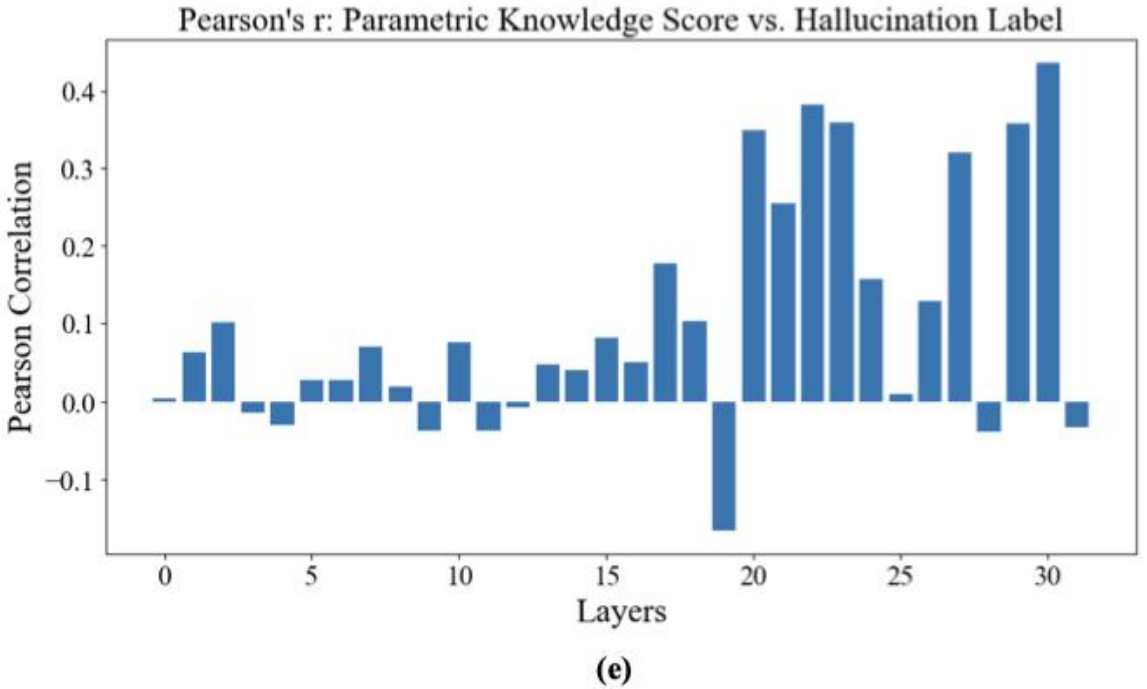
ECS 实验结果

- **Copying Head Scores**
  - **性质：** , most of eigenvalues
  - **寻找 Copying head 的方法：**
    - 核心特征：大 + **特征值方差小** 两个特征
    - 特征值方差求法：略
    - Copying Head Scores：Rank(-) + Rank(variance)

PKS 实验结果

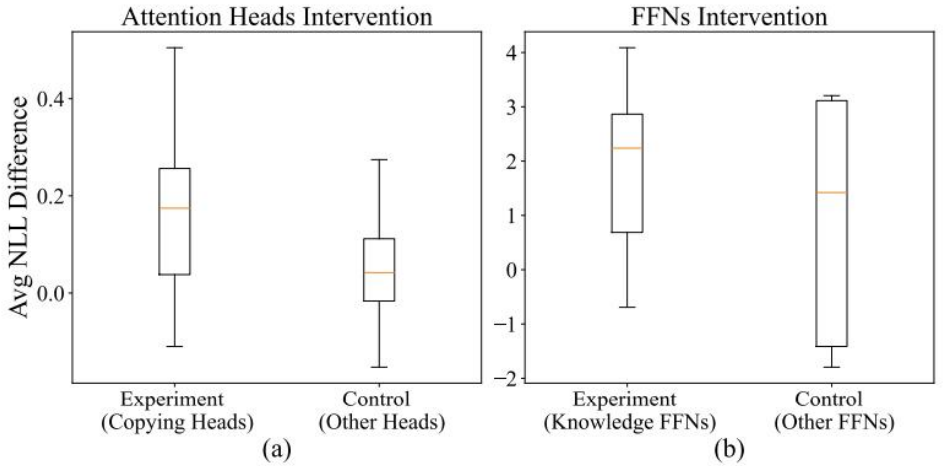
- **实验结果**





RQ2 实验结果

- **RQ2：** Can the relationship identified in RQ1 be validated from a causal perspective?



Attention Heads Intervention: As described in Figure 3 (b), we applied noise to the attention scores $\boldsymbol{a}_i^{l,h}$ of the experimental group to evaluate their impact on hallucinations. Specifically, the attention scores were sampled from a standard normal distribution:

$$a_{i,j}^{l,h} \sim \mathcal{N}(0,1), \quad \tilde{\boldsymbol{a}}_i^{(l,h)} = \text{softmax}\left(\boldsymbol{a}^{(l,h)}\right). \tag{13}$$

This approach simulates the removal of meaningful attention patterns, allowing us to assess how Copying Heads' focus on external context impacts hallucinations.

FFN Modules Intervention: To investigate the role of Knowledge FFNs in hallucinations, we amplified the effect of the FFN modules by increasing their contribution to the residual stream tenfold ($k = 10$). This intervention highlights the influence of parametric knowledge on the generation process.

RQ3 实验结果

- **RQ3：** Hallucination Behavior Analysis from the Parametric Knowledge Perspective

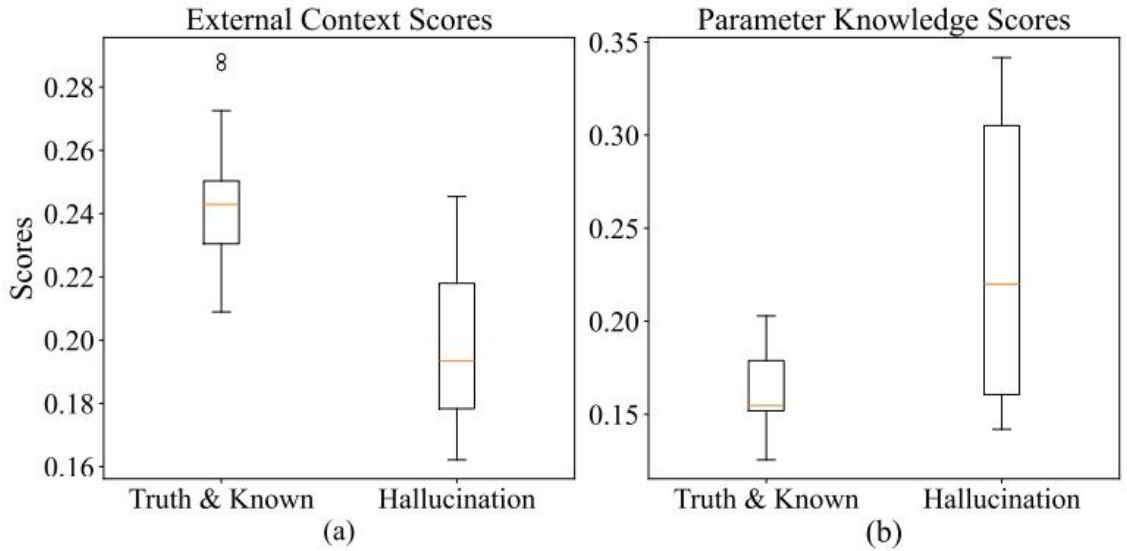


Figure 5: (Left) Intervention Result for Attention Heads and FFNs. (Right) External Context Scores and Parametric Knowledge Scores (scaled by $1e^5$) comparing Truth & Known (where LLM knows the truthful answer) and Hallucination (where LLM is unknown about the answer and hallucinated).

Hallucination 检测方法 ReDeEP (Token)

- **检测方法**

$$\mathcal{H}_t(\mathbf{r}) = \frac{1}{|\mathbf{r}|}\sum_{t\in\mathbf{r}}\mathcal{H}_t(t), \quad \mathcal{H}_t(t) = \sum_{l\in\mathcal{F}}\alpha\cdot\mathcal{P}_t^l - \sum_{l,h\in\mathcal{A}}\beta\cdot\mathcal{E}_t^{l,h},$$

where α and β are regression coefficients for external context and parametric knowledge with $\alpha, \beta > 0$, and this linear regression leverages the high Pearson correlation identified in § 3.

Hallucination 检测方法 ReDeEP (Chunk)

- **检测方法**

As the *Token-level Hallucination Detection* computes scores for each token, it is computationally expensive and lacks full contextual consideration. To improve efficiency and accuracy, we propose *Chunk-level Hallucination Detection* as a more suitable method for RAG hallucination detection. Our approach is inspired by the common chunking operation in RAG Fan et al. (2024); Finardi et al. (2024), where the retrieved context $\mathbf{c}$ and the response $\mathbf{r}$ are divided into manageable segments $\langle \tilde{\boldsymbol{c}}_i \rangle_{i=1}^{N}$ and $\langle \tilde{\boldsymbol{r}}_j \rangle_{j=1}^{M}$. For the chunk-level external context score $\hat{\mathcal{E}}^{l,h}$, we first calculate chunk-level attention weights $W_{i,j}^{l,h} = \text{Mean-Pooling}\left(A_{\tilde{\mathbf{c}}_i,\tilde{\mathbf{r}}_j}^{l,h}\right)$, where A is the original token-level attention weight matrix, then determine the highest attention chunk pairs $(\tilde{\mathbf{c}}, \tilde{\mathbf{r}})$. Using an embedding model (emb), we compute the external context score for each chunk as follows:

$$\tilde{\mathcal{E}}_{\mathbf{r}}^{l,h} = \frac{1}{M} \sum_{\tilde{\mathbf{r}} \in \mathbf{r}} \tilde{\mathcal{E}}_{\tilde{\mathbf{r}}}^{l,h}, \quad \tilde{\mathcal{E}}_{\tilde{\mathbf{r}}}^{l,h} = \frac{\text{emb}(\tilde{\mathbf{r}}) \cdot \text{emb}(\tilde{\mathbf{c}})}{\| \text{emb}(\tilde{\mathbf{r}}) \| \| \text{emb}(\tilde{\mathbf{c}}) \|}.$$

For the chunk-level parametric knowledge score $\tilde{\mathcal{P}}^{l}$, we sum the token-level parametric knowledge scores for each chunk:

$$\tilde{\mathcal{P}}_{\mathbf{r}}^{l} = \frac{1}{M} \sum_{\tilde{\mathbf{r}} \in \mathbf{r}} \tilde{\mathcal{P}}_{\tilde{\mathbf{r}}}^{l}, \quad \tilde{\mathcal{P}}_{\tilde{\mathbf{r}}}^{l} = \frac{1}{|\tilde{\mathbf{r}}|} \sum_{t \in \tilde{\mathbf{r}}} \mathcal{P}_{t}^{l}.$$

Finally, the Chunk-level Hallucination Detection score $\mathcal{H}_c(\mathbf{r})$ is defined as:

$$\mathcal{H}_c(\mathbf{r}) = \sum_{l \in \mathcal{F}} \alpha \cdot \tilde{\mathcal{P}}_{\mathbf{r}}^{l} - \sum_{l,h \in \mathcal{A}} \beta \cdot \tilde{\mathcal{E}}_{\mathbf{r}}^{l,h}.$$

消除方法：Add Attention ReduceFFN (AARF)

- **消除方法**

During the generation of token t_n, we compute the hallucination score $\mathcal{H}_t(t_n)$. If $\mathcal{H}_t(t_n) \leq \tau$, we proceed with the normal output computation $f(\mathbf{x})$ (see Equation 12). If $\mathcal{H}_t(t_n) > \tau$, we adjust the weights of Copying Heads $\mathcal{A}$ and Knowledge FFN modules $\mathcal{F}$, shifting focus toward external context and reducing reliance on parametric knowledge:

$$f(\mathbf{x}) = \sum_{l=1}^{L} \sum_{h=1}^{H} \widehat{\text{Attn}}^{l,h}\left(\boldsymbol{X}_{\leq n}^{l-1}\right) \boldsymbol{W}_U + \sum_{l=1}^{L} \widehat{\text{FFN}}^{l}\left(\boldsymbol{x}_n^{\text{mid},l}\right) \boldsymbol{W}_U + \boldsymbol{x}_n \boldsymbol{W}_U,$$

$$\widehat{\text{Attn}}^{l,h}(\cdot) = \begin{cases} \alpha_2 \cdot \text{Attn}^{l,h}\left(\boldsymbol{X}_{\leq n}^{l-1}\right), & \text{if } (l,h) \in \mathcal{A}, \\ \text{Attn}^{l,h}\left(\boldsymbol{X}_{\leq n}^{l-1}\right), & \text{otherwise} \end{cases}, \quad \widehat{\text{FFN}}^{l}(\cdot) = \begin{cases} \beta_2 \cdot \text{FFN}^{l}\left(\boldsymbol{x}_n^{\text{mid},l}\right), & \text{if } l \in \mathcal{F}, \\ \text{FFN}^{l}\left(\boldsymbol{x}_n^{\text{mid},l}\right), & \text{otherwise.} \end{cases}$$

Here, α_2 is a constant greater than 1 for amplifying attention head contributions, and β_2 is a constant between (0, 1) for reducing FFN contributions.

实验结果

- LLaMA2-7B

LLMs	Categories	Models	RAGTruth				Dolly (AC)			
			AUC	PCC	Rec.	F_1	AUC	PCC	Rec.	F_1
LLaMA2-7B	MPE	SelfCheckGPT	–	–	0.3584	0.4642	–	–	0.1897	0.3188
		Perplexity	0.5091	-0.0027	0.5190	0.6749	0.6825	0.2728	0.7719	0.7097
		LN-Entropy	0.5912	0.1262	0.5383	0.6655	0.7001	0.2904	0.7368	0.6772
		Energy	0.5619	0.1119	0.5057	0.6657	0.6074	0.2179	0.6316	0.6261
		Focus	0.6233	0.2100	0.5309	0.6622	0.6783	0.3174	0.5593	0.6534
	ECP	Prompt	–	–	0.7200	0.6720	–	–	0.3965	0.5476
		LMvLM	–	–	0.7389	0.6473	–	–	0.7759	0.7200
		ChainPoll	0.6738	0.3563	0.7832	0.7066	0.6593	0.3502	0.4138	0.5581
		RAGAS	0.7290	0.3865	0.6327	0.6667	0.6648	0.2877	0.5345	0.6392
		Trulens	0.6510	0.1941	0.6814	0.6567	0.7110	0.3198	0.5517	0.6667
		RefCheck	0.6912	0.2098	0.6280	0.6736	0.6494	0.2494	0.3966	0.5412
		P(True)	0.7093	0.2360	0.5194	0.5313	0.6011	0.1987	0.6350	0.6509
	PCE	EigenScore	0.6045	0.1559	0.7469	0.6682	0.6786	0.2428	0.7500	0.7241
		SEP	0.7143	0.3355	0.7477	0.6627	0.6067	0.2605	0.6216	0.7023
		SAPLMA	0.7037	0.3188	0.5091	0.6726	0.5365	0.0179	0.5714	0.7179
		ITI	0.7161	0.3932	0.5416	0.6745	0.5492	0.0442	0.5816	0.6281
	Ours	**ReDeEP(token)**	0.7325	0.3979	0.6770	0.6986	0.6884	0.3266	0.8070	0.7244
		ReDeEP(chunk)	**0.7458**	**0.4203**	**0.8097**	**0.7190**	**0.7949**	**0.5136**	**0.8245**	**0.7833**

实验结果

- LLaMA2-13B

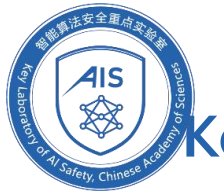
LLaMA2-13B	MPE	SelfCheckGPT	–	–	0.3584	0.4642	–	–	0.1897	0.3188
		Perplexity	0.5091	-0.0027	0.5190	0.6749	0.6825	0.2728	0.7719	0.7097
		LN-Entropy	0.5912	0.1262	0.5383	0.6655	0.7001	0.2904	0.7368	0.6772
		Energy	0.5619	0.1119	0.5057	0.6657	0.6074	0.2179	0.6316	0.6261
		Focus	0.7888	0.4444	0.6173	0.6977	0.7067	0.1643	0.7333	0.6168
	ECP	Prompt	–	–	0.7000	0.6899	–	–	0.4182	0.5823
		LMvLM	–	–	**0.8357**	0.6553	–	–	0.7273	0.6838
		ChainPoll	0.7414	0.4820	0.7874	0.7342	0.7070	0.4758	0.4364	0.6000
		RAGAS	0.7541	0.4249	0.6763	0.6747	0.6412	0.2840	0.4182	0.5476
		Trulens	0.7073	0.2791	0.7729	0.6867	0.6521	0.2565	0.3818	0.4941
		RefCheck	0.7857	0.4104	0.6800	0.7023	0.6626	0.2869	0.2545	0.3944
		P(True)	0.7998	0.3493	0.5980	0.7032	0.6396	0.2009	0.6180	0.5739
	PCE	EigenScore	0.6640	0.2672	0.6715	0.6637	0.7214	0.2948	0.8181	0.7200
		SEP	0.8089	0.5276	0.6580	0.7159	0.7098	0.2823	0.6545	0.6923
		SAPLMA	0.8029	0.3956	0.5053	0.6529	0.6053	0.2006	0.6000	0.6923
		ITI	0.8051	0.4771	0.5519	0.6838	0.5511	0.0646	0.5385	0.6712
	Ours	**ReDeEP(token)**	0.8181	0.5478	0.7440	0.7494	0.7226	0.3776	0.8148	0.7154
		ReDeEP(chunk)	**0.8244**	**0.5566**	0.7198	**0.7587**	**0.8420**	**0.5902**	**0.8518**	**0.7603**

实验结果

- LLaMA3-8B

LLaMA3-8B	**MPE**	SelfCheckGPT	–	–	0.4111	0.5111	–	–	0.2195	0.3600
		Perplexity	0.6235	0.2100	0.6537	0.6778	0.5924	0.1095	0.3902	0.4571
		LN-Entropy	0.7021	0.3451	0.5596	0.6282	0.6011	0.1150	0.5365	0.5301
		Energy	0.5959	0.1393	0.5514	0.6720	0.5014	-0.0678	0.4047	0.5440
		Focus	0.6378	0.3079	0.6688	0.6879	0.6177	0.1266	0.6918	0.6874
	ECP	Prompt	–	–	0.4403	0.5691	–	–	0.3902	0.5000
		LMvLM	–	–	0.5109	0.6986	–	–	0.6341	0.5361
		ChainPoll	0.6687	0.3693	0.4486	0.5813	0.6114	0.2691	0.3415	0.4516
		RAGAS	0.6776	0.2349	0.3909	0.5094	0.6870	0.3628	0.8000	0.5246
		Trulens	0.6464	0.1326	0.3909	0.5053	0.7040	0.3352	0.3659	0.5172
		RefCheck	0.6014	0.0426	0.3580	0.4628	0.5260	-0.0089	0.1951	0.2759
		P(True)	0.6323	0.2189	0.7083	0.6835	0.6871	0.3472	0.5707	0.6573
	PCE	EigenScore	0.6497	0.2120	0.7078	0.6745	0.6612	0.2065	0.7142	0.5952
		SEP	0.7004	0.3713	0.7333	0.6915	0.5159	0.0639	0.6829	0.5385
		SAPLMA	0.7092	0.4054	0.5432	0.6718	0.5019	-0.0327	0.4040	0.5714
		ITI	0.6534	0.3404	0.6850	0.6933	0.5011	0.0024	0.3091	0.4250
	Ours	**ReDeEP(token)**	**0.7522**	**0.4493**	**0.7984**	**0.7132**	0.6701	0.2421	0.8293	0.6901
		ReDeEP(chunk)	0.7285	0.3964	0.7819	0.6947	**0.7354**	**0.3652**	**0.8392**	**0.7100**

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES



智能算法安全重点实验室
Key Laboratory of AI Safety, CAS

Thank you for your attentions!