中国人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

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

ZhongXiang Sun (孙忠祥), Xiaoxue Zang, Kai Zheng,
Jun Xu (徐君), Xiao Zhang, Weijie Yu, Yang Song, Han Li

Gaoling School of Artificial Intelligence
Renmin University of China

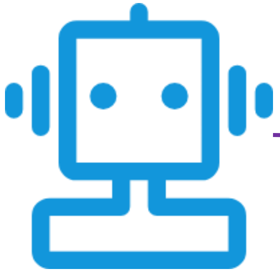
Background: RAG

Retrieved Document

… China's lunar exploration project, the Chang'e-6 mission, was successfully completed, …. bringing back 1935.3 grams of lunar samples…

Query

China's Chang'e 6 mission successfully completed the world's first sample return from the far side of the moon. How many grams of samples were brought back?

LLaMA2-70B: 

Parametric knowledge

… brought back 1,731 grams of lunar samples…

Response

China's Chang'e-6 mission successfully collected 1935.3 grams of samples from the far side of the Moon.

Correct

Can LLMs always answer correctly with relevant documents?

Background: RAG Hallucination

Retrieved Document

… To get good pictures, 3D ultrasounds are best performed between 20 and 32 weeks, and the best pictures are between 24 and 30 weeks …

Query

How to prepare to get an ultrasound?

LLaMA2-70B:

Parametric knowledge

… between 24 and 32 weeks for the best pictures …

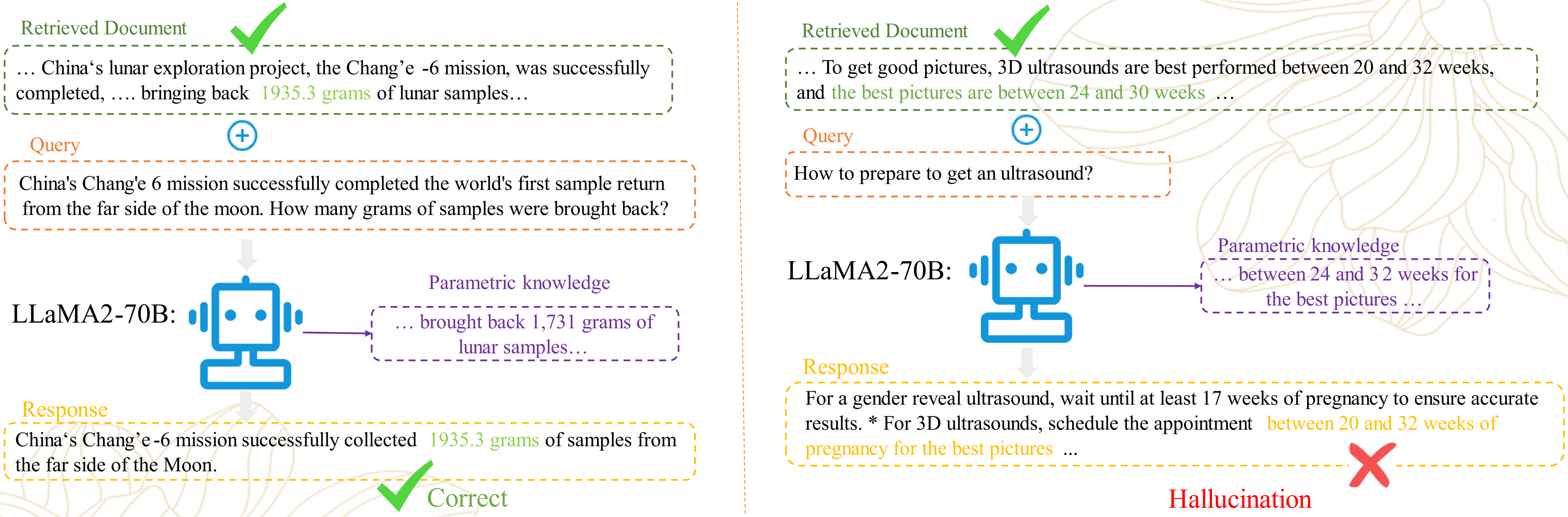
Response

For a gender reveal ultrasound, wait until at least 17 weeks of pregnancy to ensure accurate results. * For 3D ultrasounds, schedule the appointment between 20 and 32 weeks of pregnancy for the best pictures ...

Hallucination

Niu C, Wu Y, Zhu J, et al. RAGTruth. ACL 2024.

Observation



- Recent studies have examined the Conflicts between the external context and the LLM's parametric knowledge of RAG.
- We find these conflicts can lead to hallucinations but do not always cause them.

Research Problem

- Detecting RAG hallucinations
  - Specifically in cases where the retrieved external context is accurate and relevant.

of people born in the 1980s. According to the seventh national census data, the current population of people born in the 1980s is 212 million, with a survival rate of 94.8%, and a death rate of 5.2%.

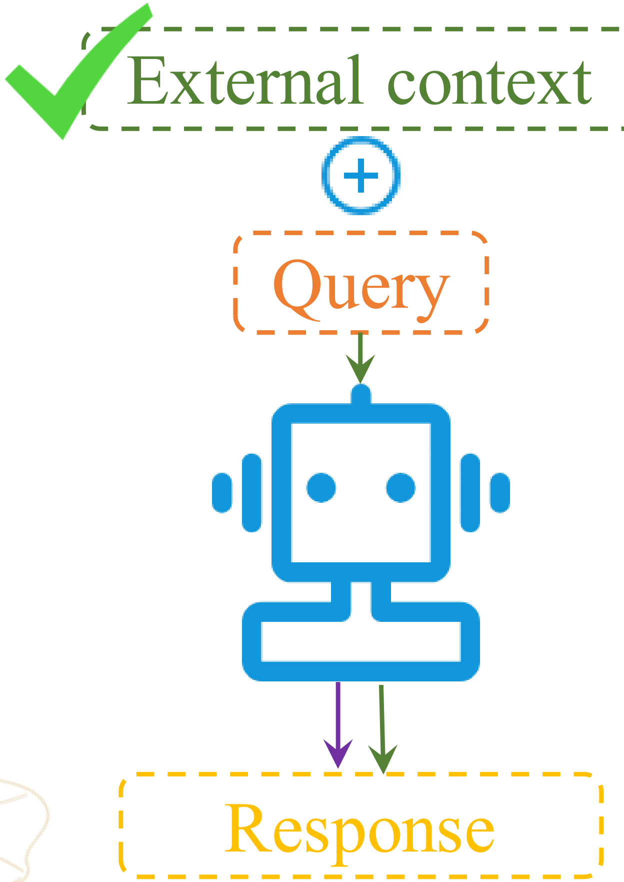
This is relatively straightforward data. In the table on page 10, there are birth and death rates for each decade. For example, the average death rate for those born in the 1970s is 7.11 per thousand, for the 1980s it's 5.99 per thousand, and for the 1990s it's 6.57 per thousand.

The LLM hallucinates the per mille sign (‰) in the retrieved document as a percent sign (%) in its generated response.

RAG Hallucination vs. LLM Hallucination Detection

RAG vs. LLM Hallucination Detection: Causal View

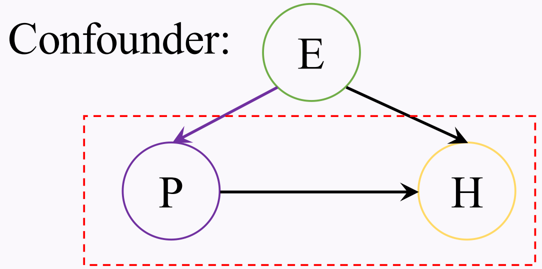
E: External Context P: Parametric Knowledge H: Response Hallucination or not



RAG vs. LLM Hallucination Detection: Causal View

E: External Context P: Parametric Knowledge H: Response Hallucination or not

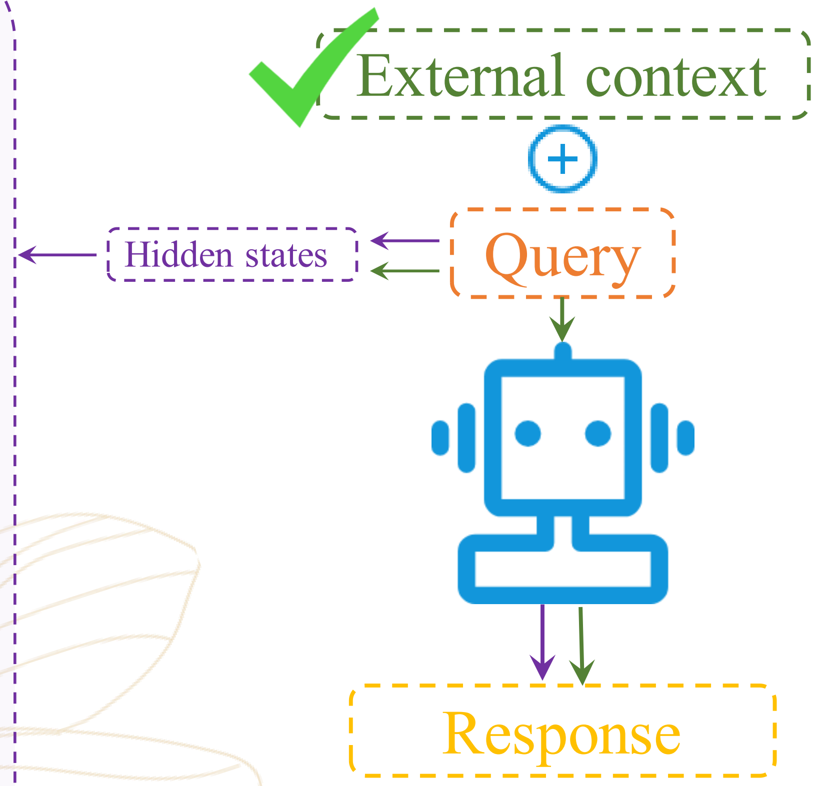
(i) P confounded by E



Causal graph (a)

Methods

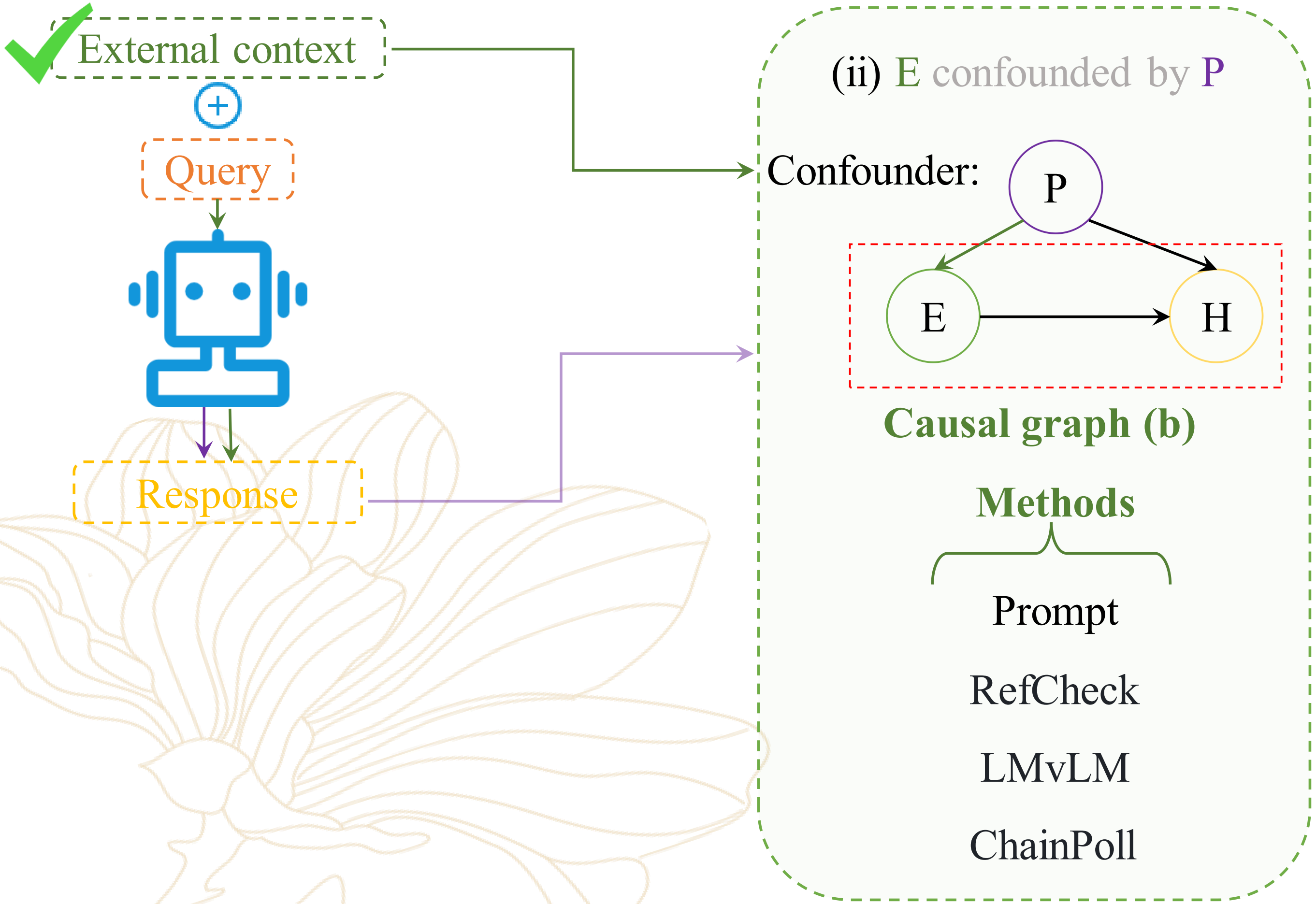
ITI
SEP
SAPLMA
EigenScore



- From a knowledge storage perspective:
  - Hidden states represent the result of querying the parametric knowledge (P) with external context (E), establishing a causal path from E to P
- From a causal perspective:
  - The presence of E as a confounder complicates the accurate prediction of hallucinations based on P alone.

RAG vs. LLM Hallucination Detection: Causal View

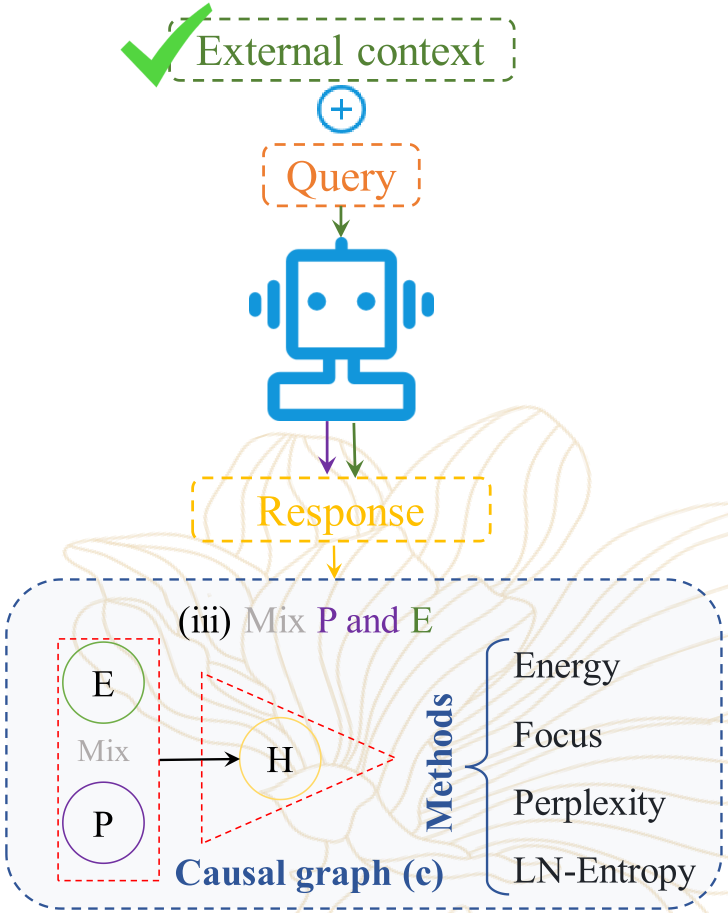
E: External Context　　P: Parametric Knowledge　　H: Response Hallucination or not



- Parametric knowledge (P) is a confounder between the external context (E) and hallucinations (H)
- Due to the unavoidable presence of parametric knowledge in the response

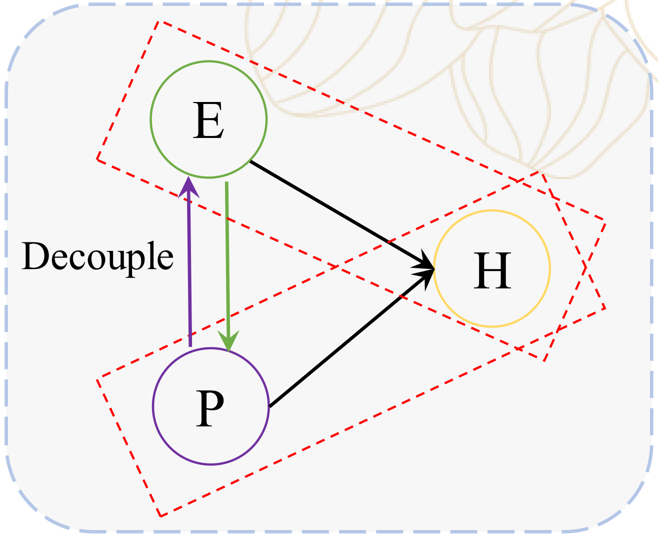
RAG vs. LLM Hallucination Detection: Causal View

E: External Context P: Parametric Knowledge H: Response Hallucination or not



- Mixing of E and P without decoupling their roles obscures their individual contributions.

Decouple & Regression

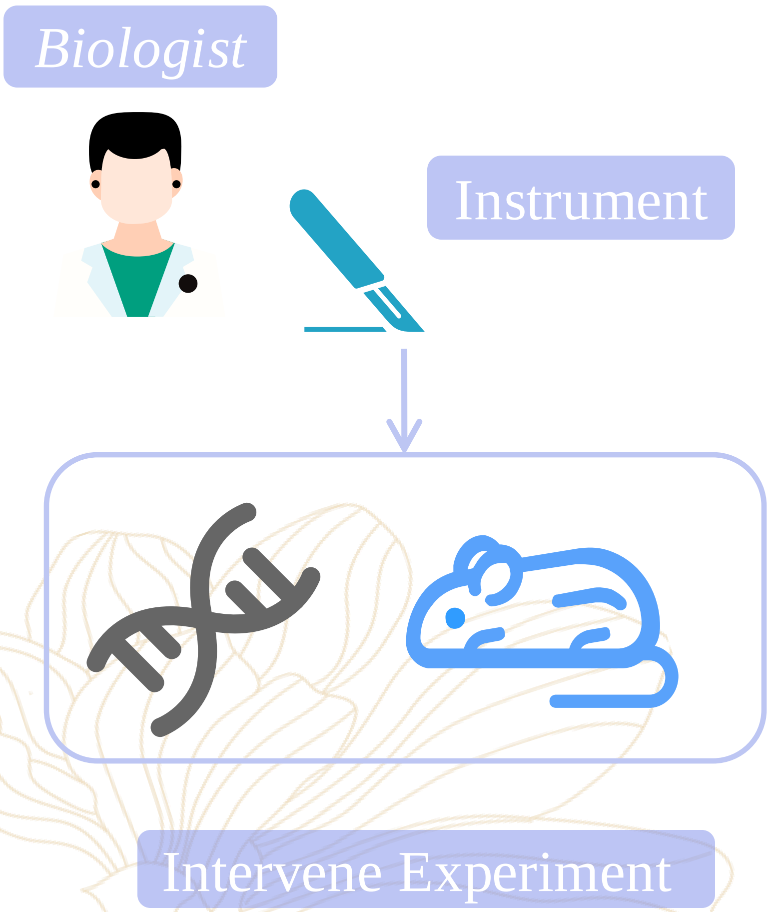


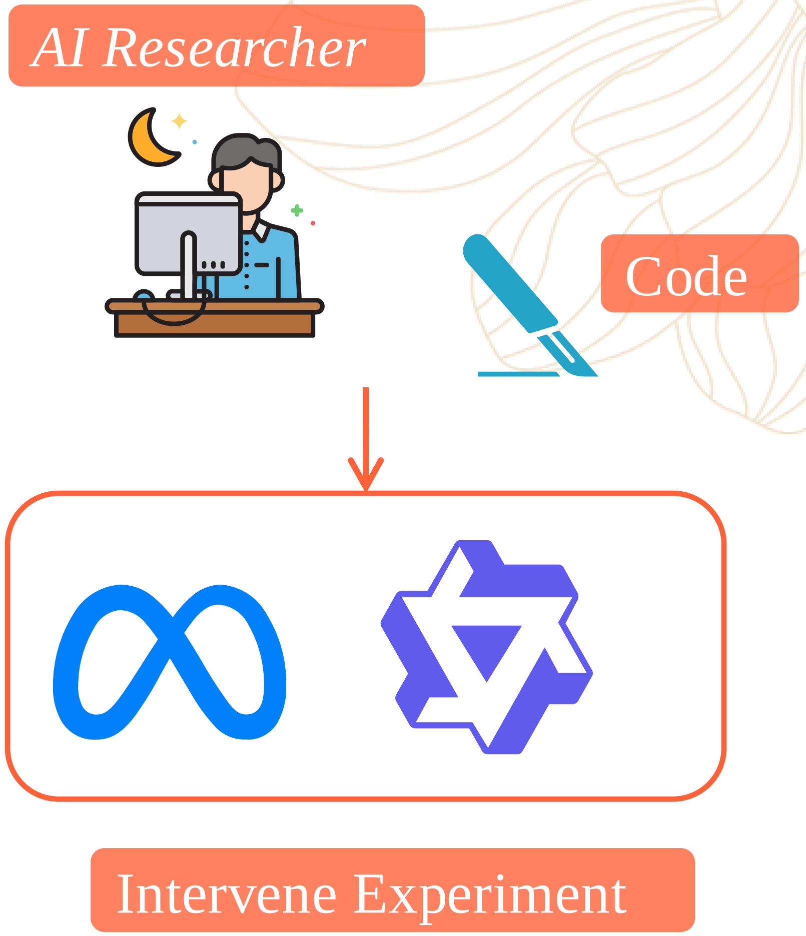
(Ours)

Treat P and E as covariates

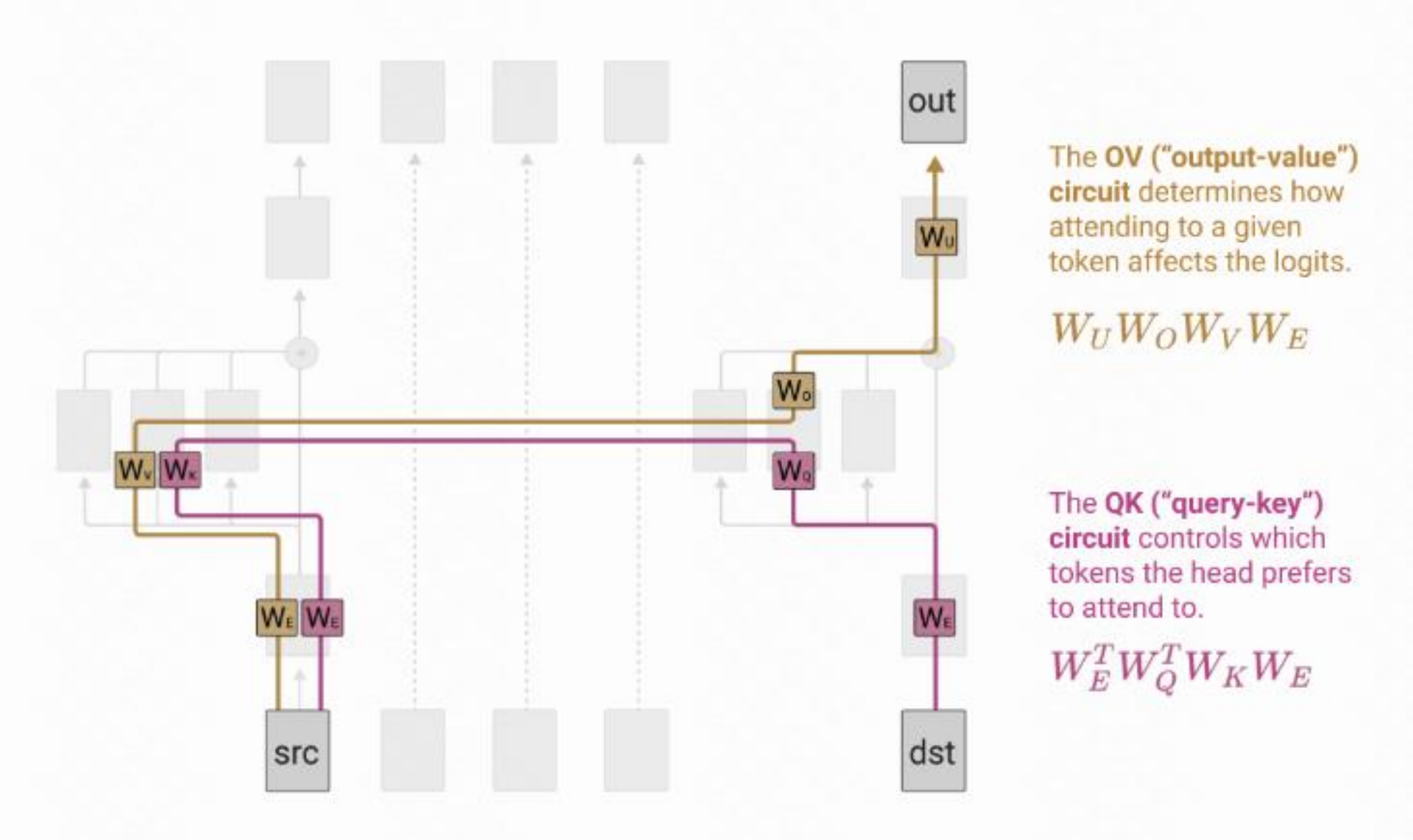
to solve the confounding problem

Mechanistic Interpretability





Mechanistic Interpretability Findings: Copying Heads



Copying Head: Heads with more positive OV matrix is likely for copying information to the logits

[1] https://transformer-circuits.pub/2021/framework/index.html

Mechanistic Interpretability Findings: FFNs

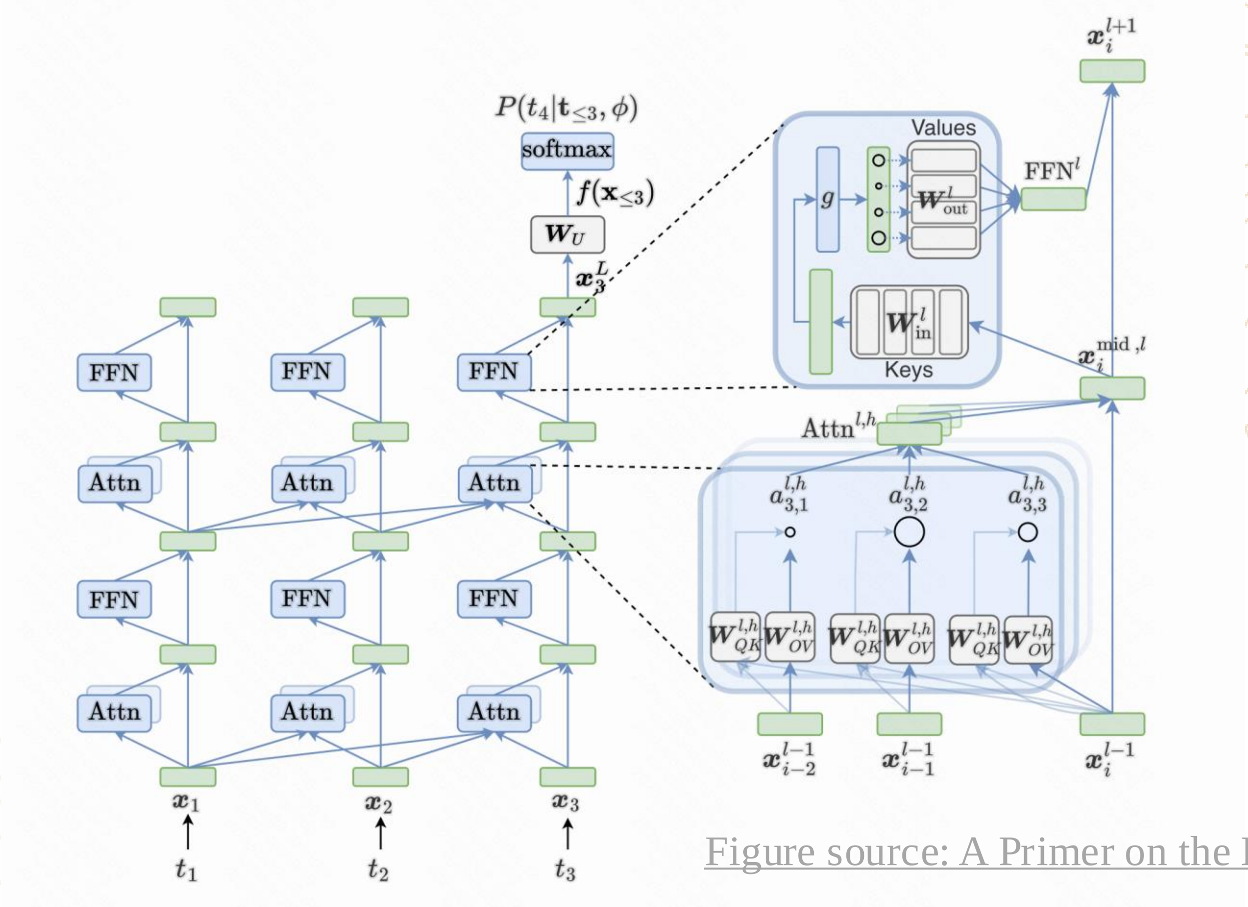


Figure source: A Primer on the Inner Workings of Transformer

- Each FFN layer transforms the hidden state by linearly combining key-value pairs.
- FFNs enabling the model to retrieve and integrate stored information effectively for prediction.

[1] Dai D, Dong L, Hao Y, et al. Knowledge neurons in pretrained transformers[J]

Mechanistic Interpretability Findings: Logit Lens

$$
\begin{aligned}
f(\mathbf{x}) &= \boldsymbol{x}_n^L \boldsymbol{W}_U \\
&= \left(\sum_{l=1}^{L} \sum_{h=1}^{H} \mathrm{Attn}^{l,h}(\boldsymbol{X}_{\leq n}^{l-1}) + \sum_{l=1}^{L} \mathrm{FFN}^{l}(\boldsymbol{x}_n^{\mathrm{mid},l}) + \boldsymbol{x}_n \right) \boldsymbol{W}_U \\
&= \underbrace{\sum_{l=1}^{L} \sum_{h=1}^{H} \mathrm{Attn}^{l,h}(\boldsymbol{X}_{\leq n}^{l-1}) \boldsymbol{W}_U}_{\text{Attention head logits update}} + \underbrace{\sum_{l=1}^{L} \mathrm{FFN}^{l}(\boldsymbol{x}_n^{\mathrm{mid},l}) \boldsymbol{W}_U}_{\text{FFN logits update}} + \boldsymbol{x}_n \boldsymbol{W}_U .
\end{aligned}
$$

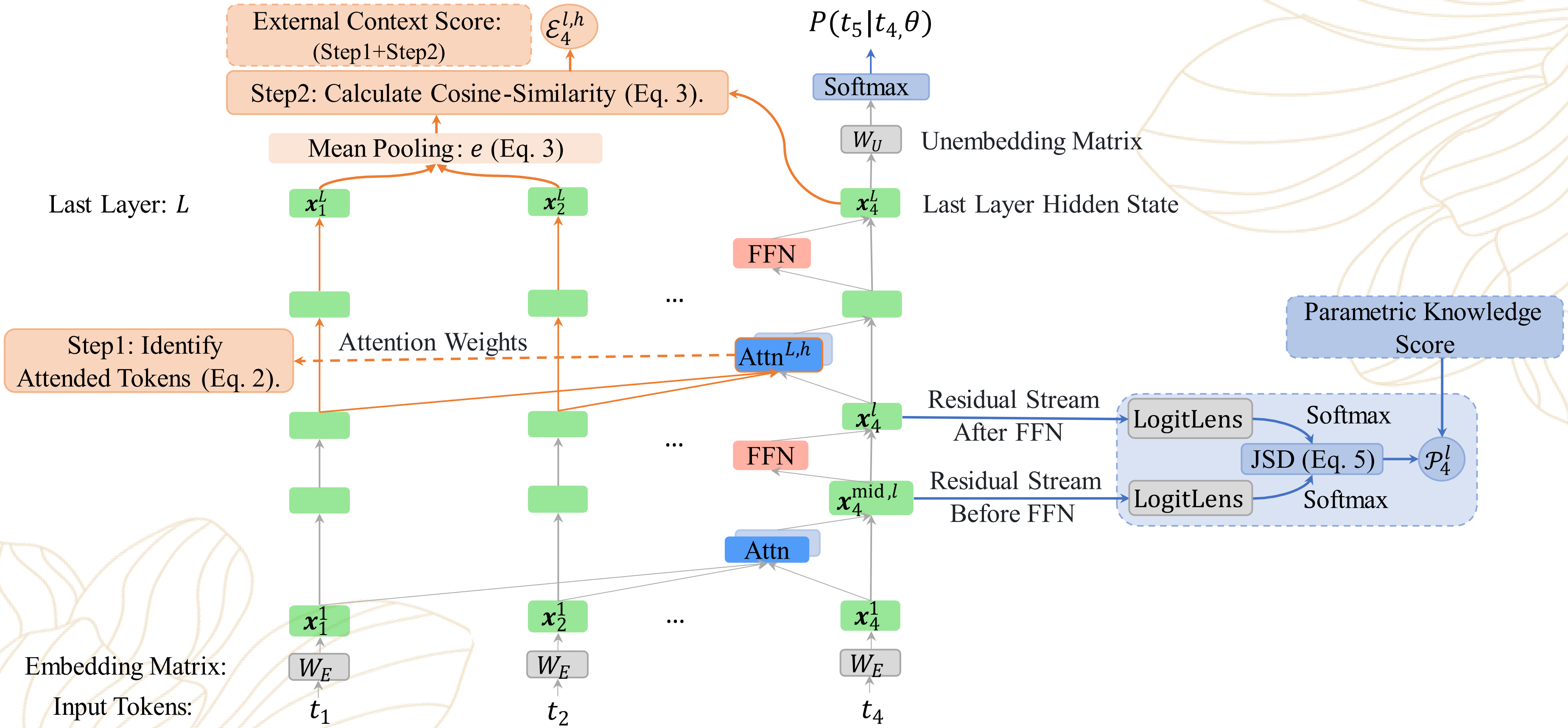
Logit Lens: The LogitLens is a technique that decodes hidden states $\boldsymbol{x}^l$ directly into the vocabulary distribution using the LayerNorm and the unembedding matrix $\boldsymbol{W}_U$ of the LLM for interpretability (nostalgebraist, 2020):

$$\mathrm{LogitLens}\left(\boldsymbol{x}^l\right) = \mathrm{LayerNorm}(\boldsymbol{x}^l)\boldsymbol{W}_U . \quad (1)$$

[1] nostalgebraist. Interpreting GPT: the logit lens.

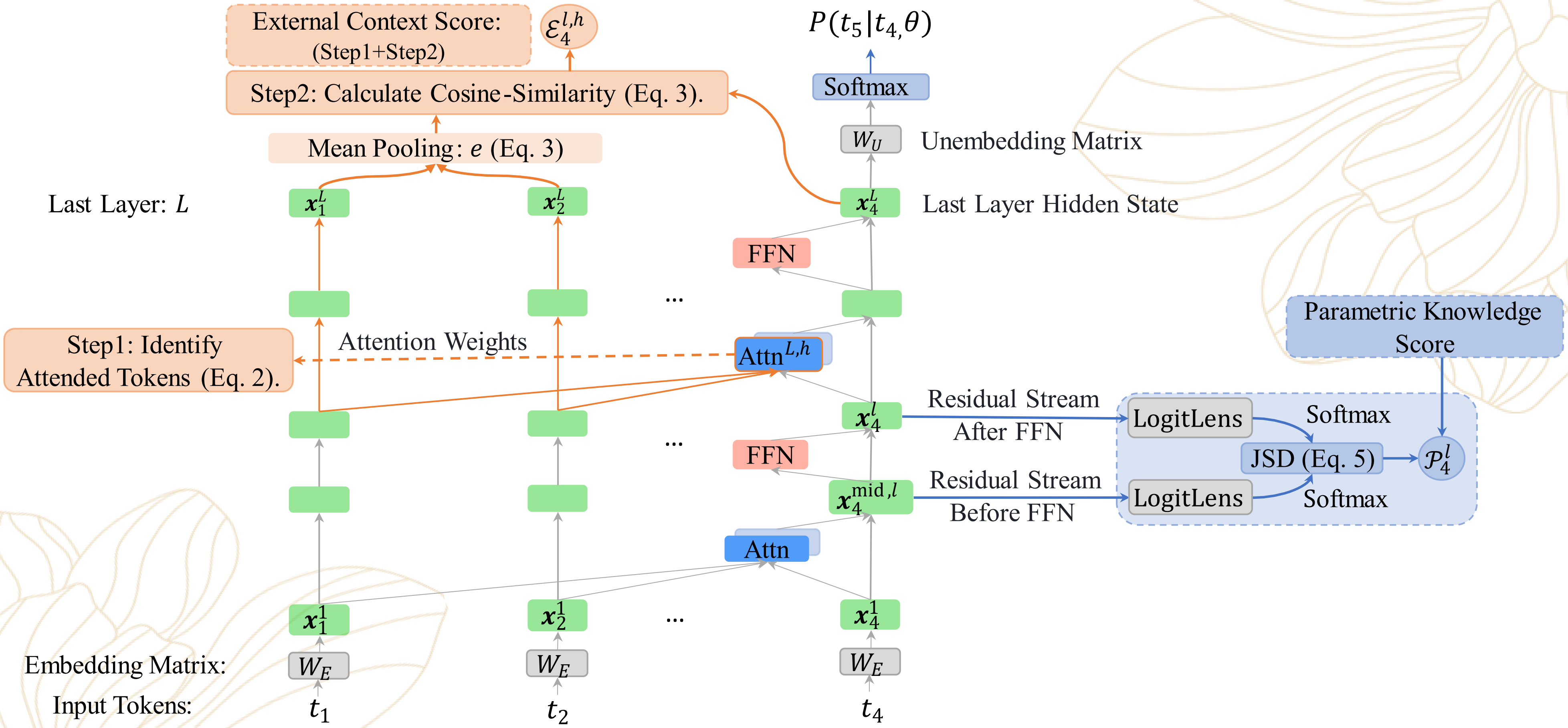
How to Utilize Mechanistic Interpretability to Analyze RAG Hallucination

Definition: ECS



- **Whether attention heads focus on the correct context:**
  - Experiments show that hallucinations often occur despite attention heads correctly attending to the external context.
- **External Context Score:**
  - Whether the LLM effectively retains and utilizes this information from external context during generation.

Definition: PKS



- **Parametric Knowledge Score:**
  - How much Parametric Knowledge dose the LLM utilize during generation.

Empirical Study: Setting

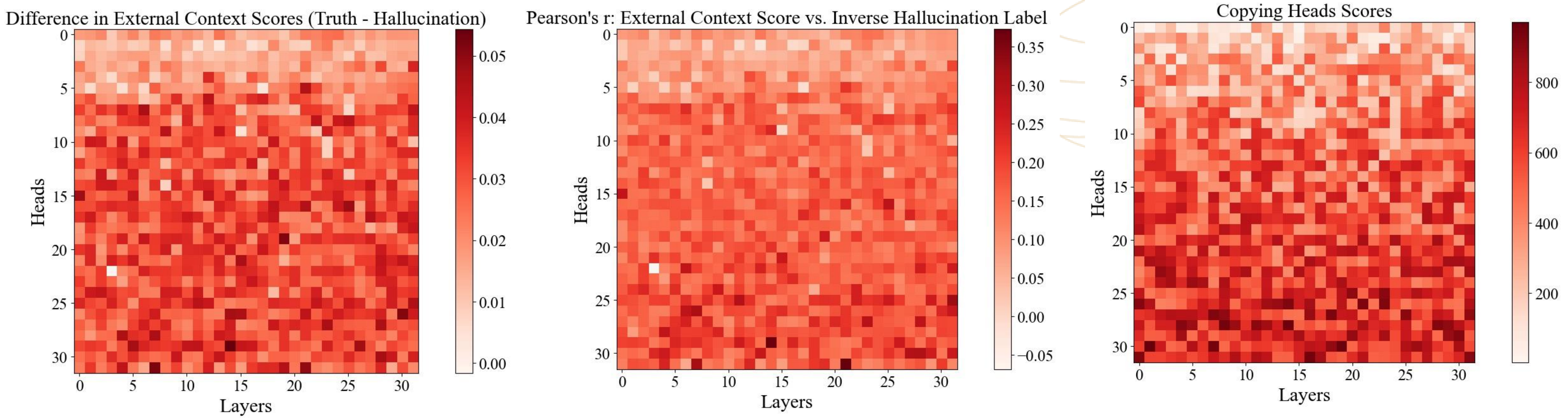
Model: LLaMA2-7B-Chat

Dataset: RAGTruth

Question:

- **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:** How to Analysis Hallucination Behavior Analysis from the Parametric Knowledge Perspective?

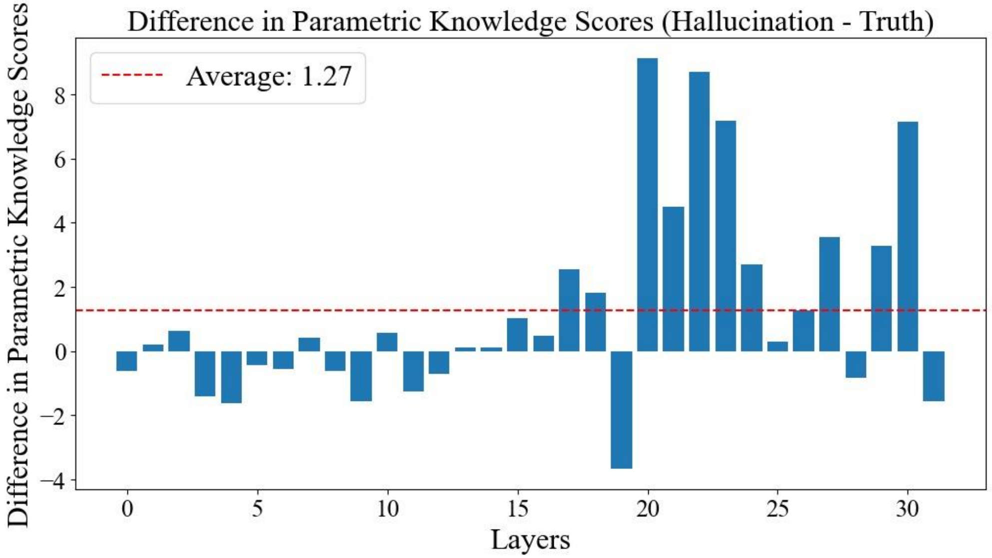
RQ1.1: Relationship Between ECS and Hallucination



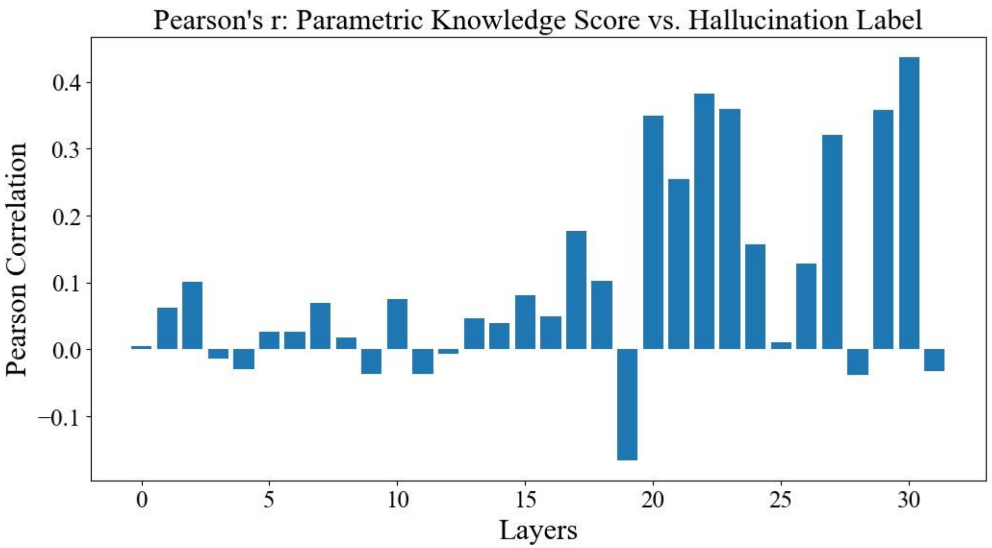
(a) (b) (c)

- **ECS Differences between Truthful and Hallucinated Responses:**
  - LLMs utilize external context information less than truthful responses when generating hallucinations.
- **Correlation between ECS and Hallucination**
  - RAG hallucinations occur when the LLM inadequately leverages external context.
- **Relation between Copying Heads and Hallucination**
  - Attention Heads strongly correlated with hallucination exhibit characteristics of Copying Heads.

RQ1.2: Relationship Between PKS and Hallucination



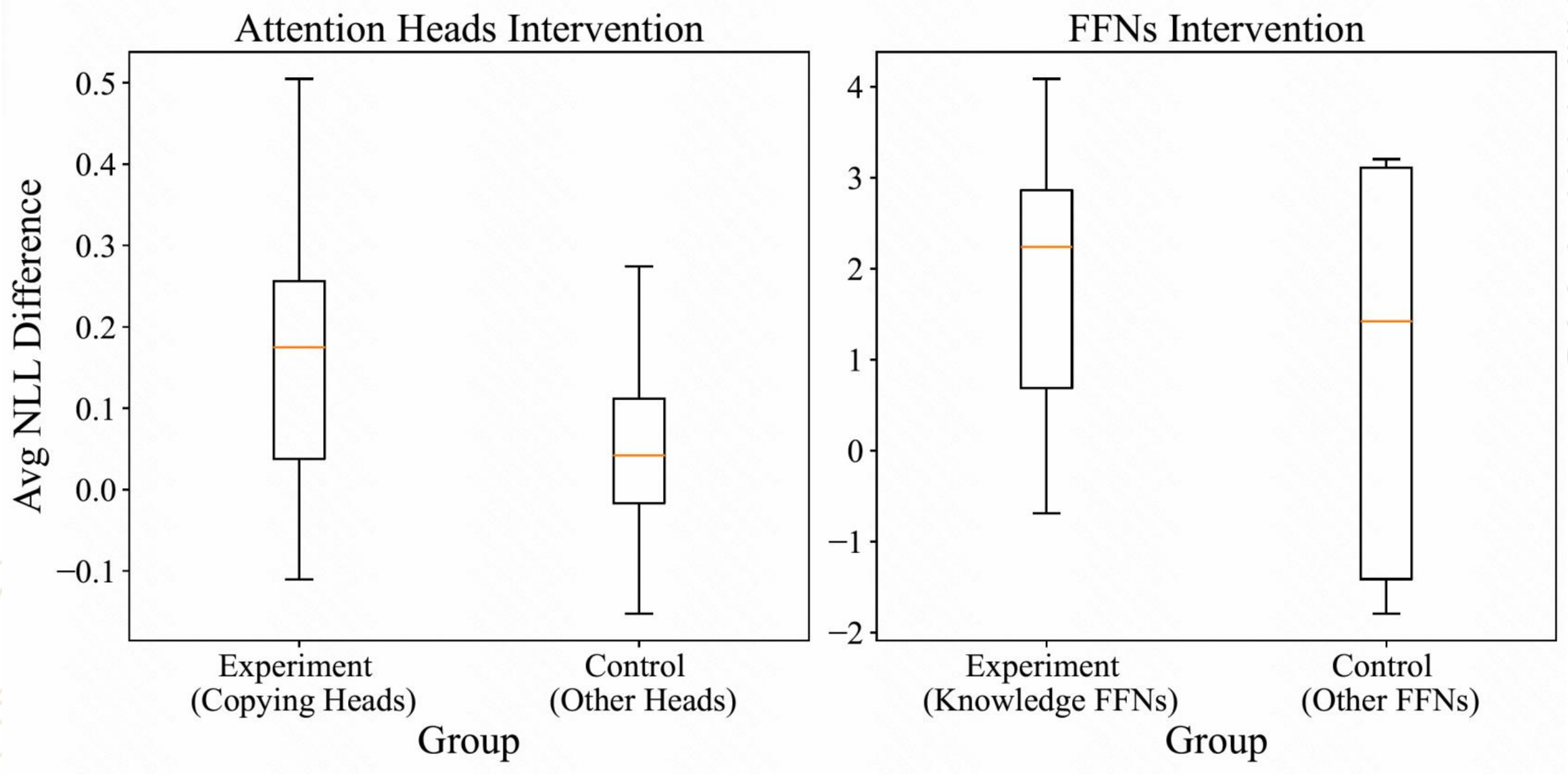
(d)



(e)

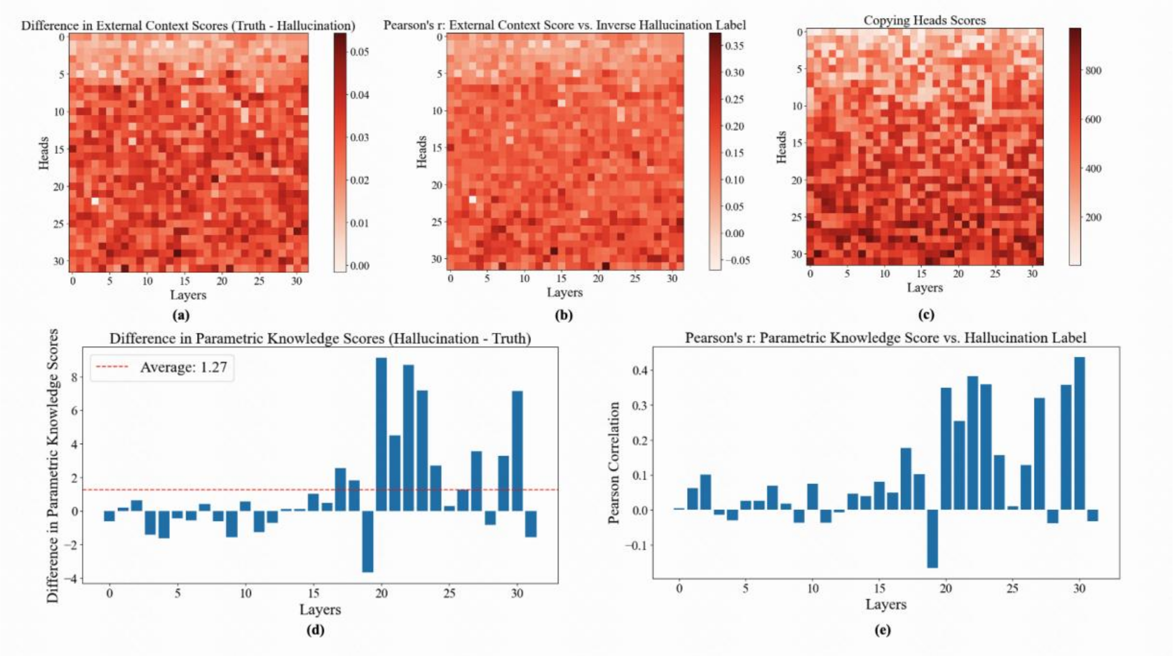
- **PKS Differences between Truth and Hallucination:**
  - Across all layers, hallucination responses exhibit higher parametric knowledge scores than truthful ones..
- **Correlation between PKS and Hallucination:**
  - Parametric knowledge scores in the later layers' FFN modules are positively correlated with the hallucination
- When external context provides sufficient information, shallow layers can generate truthful responses, but over-reliance on parametric knowledge from deeper layers can confuse the model, causing hallucinations.

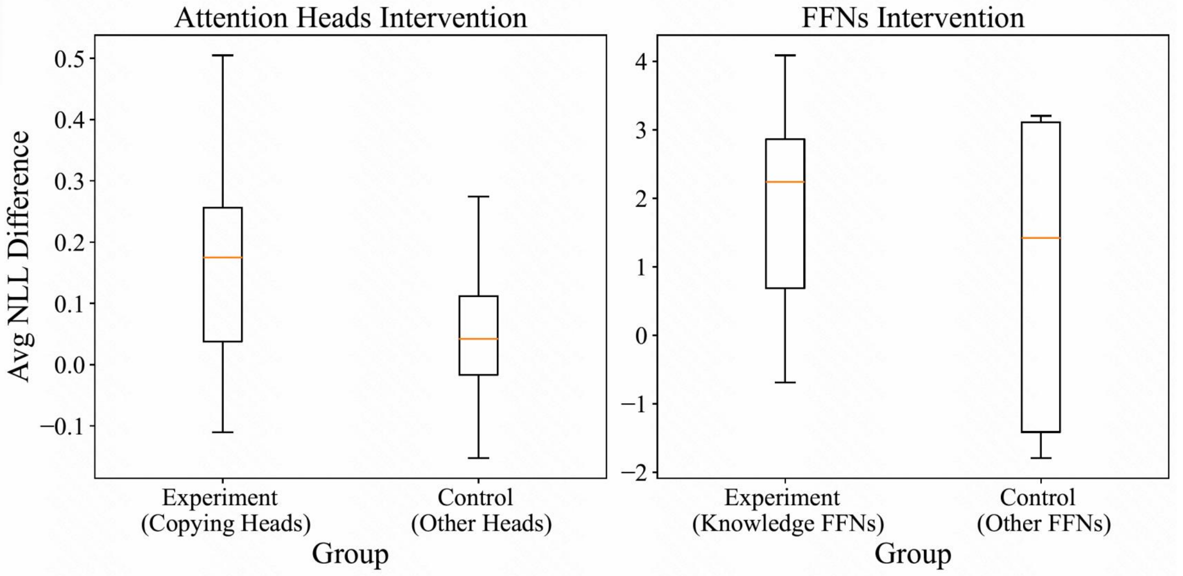
RQ2: Causal Intervention Validation



- **Intervening on attention heads and FFNs**
  - Experimental group's impact on NLL difference was significantly greater than that of the control group for both attention heads and FFN modules.

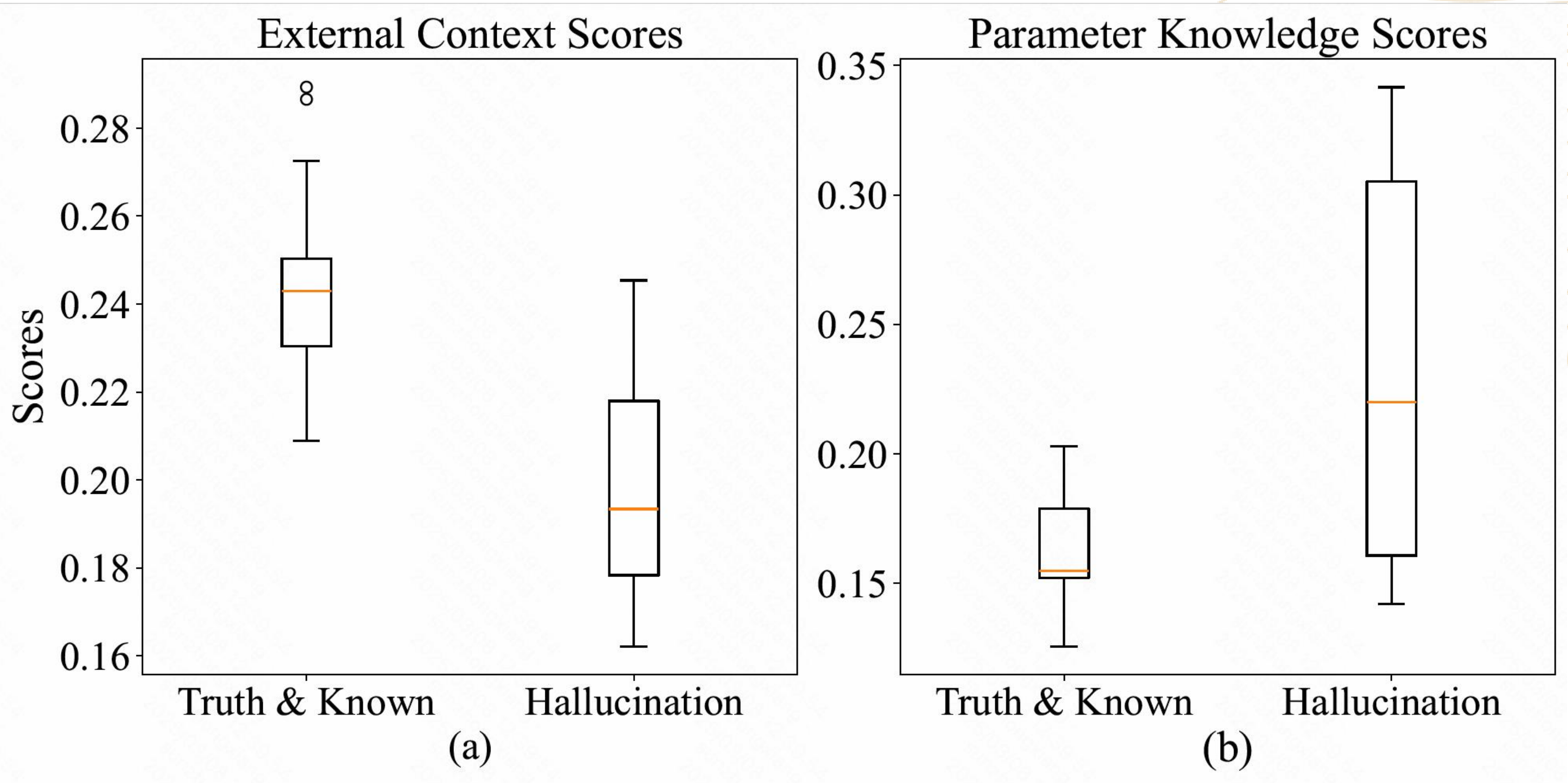
Findings





Finding: The occurrence of RAG hallucinations is causally related to two primary factors: (1) while the Copying Heads may occasionally neglect necessary knowledge from the external context, a more prominent cause is the LLM losing the Copying Heads retrieved information during the generation process (RQ1-1, RQ2, § C), and (2) the Knowledge FFNs within LLM excessively injecting parametric knowledge into the residual stream (RQ1-2, RQ2, § D).

RQ3: Hallucination Behavior Analysis from the Parametric Knowledge Perspective



- When the LLM knows the truthful answer, Copying Heads more accurately capture and utilize external knowledge, and Knowledge FFNs add less parametric knowledge to the residual stream compared to hallucination scenarios.

Leverage these Finding to Design RAG Hallucination Detection Algorithm

Token-Level Hallucination Detection

- Regresses decoupled External Context Score $\mathcal{E}$ and Parametric Knowledge Score $\mathcal{P}$ to predict hallucinations

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

where $\alpha, \beta > 0$ This Linear regression leverages the high Pearson correlation identified in empirical study.

Chunk-Level Hallucination Detection

- As the Token-level Hallucination Detection computes scores for each token, it is computationally expensive and lacks full contextual consideration.

ECS (chunk):

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

PKS (chunk):

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

Chunk-level Hallucination Detection:

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

Truthful RAG Generation

- We propose Add Attention Reduce FFN (AARF) to reduce RAG hallucinations by intervening on attention heads and FFN modules without updating model parameters.

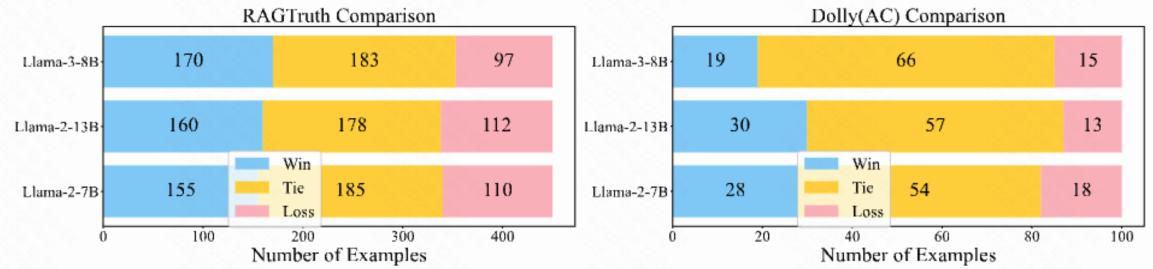
$$f(\mathbf{x}) = \sum_{l=1}^{L} \sum_{h=1}^{H} \widehat{\mathrm{Attn}}^{l,h} \left(\boldsymbol{X}_{\leq n}^{l-1} \right) \boldsymbol{W}_U + \sum_{l=1}^{L} \widehat{\mathrm{FFN}}^{l} \left(\boldsymbol{x}_n^{\mathrm{mid},l} \right) \boldsymbol{W}_U + \boldsymbol{x}_n \boldsymbol{W}_U,$$

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

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.

Experiments

LLMs	Categories	Models	RAGTruth AUC	RAGTruth PCC	RAGTruth Acc.	RAGTruth Rec.	RAGTruth F_1	Dolly (AC) AUC	Dolly (AC) PCC	Dolly (AC) Acc.	Dolly (AC) Rec.	Dolly (AC) F_1
LLaMA2-7B	MPE	SelfCheckGPT	–	–	0.5844	0.3584	0.4642	–	–	0.5300	0.1897	0.3188
		Perplexity	0.5091	-0.0027	0.5333	0.5190	0.6749	0.6825	0.2728	0.6363	0.7719	0.7097
		LN-Entropy	0.5912	0.1262	0.5600	0.5383	0.6655	0.7001	0.2904	0.6162	0.7368	0.6772
		Energy	0.5619	0.1119	0.5088	0.5057	0.6657	0.6074	0.2179	0.5656	0.6316	0.6261
		Focus	0.6233	0.2100	0.5533	0.5309	0.6622	0.6783	0.3174	0.6262	0.5593	0.6534
	ECP	Prompt	–	–	0.6700	0.7200	0.6720	–	–	0.6200	0.3965	0.5476
		Llama2-13B(LR)	–	–	0.6350	0.7078	0.6750	–	–	0.6043	0.5918	0.6616
		LMvLM	–	–	0.5946	0.7389	0.6473	–	–	0.6500	0.7759	0.7200
		ChainPoll	0.6738	0.3563	0.6741	0.7832	0.7066	0.6593	0.3502	0.6200	0.4138	0.5581
		RAGAS	0.7290	0.3865	0.6822	0.6327	0.6667	0.6648	0.2877	0.6500	0.5345	0.6392
		Trulens	0.6510	0.1941	0.6422	0.6814	0.6567	0.7110	0.3198	0.6800	0.5517	0.6667
		RefCheck	0.6912	0.2098	0.6467	0.6280	0.6736	0.6494	0.2494	0.6100	0.3966	0.5412
		P(True)	0.7093	0.2360	0.5466	0.5194	0.5313	0.6011	0.1987	0.5444	0.6350	0.6509
	PCE	EigenScore	0.6045	0.1559	0.5422	0.7469	0.6682	0.6786	0.2428	0.6596	0.7500	0.7241
		SEP	0.7143	0.3355	0.6177	0.7477	0.6627	0.6067	0.2605	0.6060	0.6216	0.7023
		SAPLMA	0.7037	0.3188	0.5155	0.5091	0.6726	0.5365	0.0179	0.5600	0.5714	0.7179
		ITI	0.7161	0.3932	0.5667	0.5416	0.6745	0.5492	0.0442	0.5800	0.5816	0.6281
	Ours	**ReDeEP(token)**	0.7325	0.3979	**0.7067**	0.6770	0.6986	0.6884	0.3266	0.6464	0.8070	0.7244
		ReDeEP(chunk)	**0.7458**	**0.4203**	0.6822	**0.8097**	**0.7190**	**0.7949**	**0.5136**	**0.7373**	**0.8245**	**0.7833**
LLaMA2-13B	MPE	SelfCheckGPT	–	–	0.5844	0.3584	0.4642	–	–	0.5300	0.1897	0.3188
		Perplexity	0.5091	-0.0027	0.5333	0.5190	0.6749	0.6825	0.2728	0.6363	0.7719	0.7097
		LN-Entropy	0.5912	0.1262	0.5600	0.5383	0.6655	0.7001	0.2904	0.6162	0.7368	0.6772
		Energy	0.5619	0.1119	0.5088	0.5057	0.6657	0.6074	0.2179	0.5656	0.6316	0.6261
		Focus	0.7888	0.4444	0.6000	0.6173	0.6977	0.7067	0.1643	0.5900	0.7333	0.6168
	ECP	Prompt	–	–	0.7300	0.7000	0.6899	–	–	0.6700	0.4182	0.5823
		Llama2-13B(LR)	–	–	0.7034	0.6839	0.7123	–	–	0.5545	0.6319	0.6664
		LMvLM	–	–	0.5956	**0.8357**	0.6553	–	–	0.6300	0.7273	0.6838
		ChainPoll	0.7414	0.4820	0.7378	0.7874	0.7342	0.7070	0.4758	0.6800	0.4364	0.6000
		RAGAS	0.7541	0.4249	0.7000	0.6763	0.6747	0.6412	0.2840	0.6200	0.4182	0.5476
		Trulens	0.7073	0.2791	0.6756	0.7729	0.6867	0.6521	0.2565	0.5700	0.3818	0.4941
		RefCheck	0.7857	0.4104	0.7200	0.6800	0.7023	0.6626	0.2869	0.5700	0.2545	0.3944
		P(True)	0.7998	0.3493	0.6266	0.5980	0.7032	0.6396	0.2009	0.5600	0.6180	0.5739
	PCE	EigenScore	0.6640	0.2672	0.5267	0.6715	0.6637	0.7214	0.2948	0.6211	0.8181	0.7200
		SEP	0.8089	0.5276	0.7288	0.6580	0.7159	0.7098	0.2823	0.6800	0.6545	0.6923
		SAPLMA	0.8029	0.3956	0.5488	0.5053	0.6529	0.6053	0.2006	0.6000	0.6000	0.6923
		ITI	0.8051	0.4771	0.6177	0.5519	0.6838	0.5511	0.0646	0.5200	0.5385	0.6712
	Ours	**ReDeEP(token)**	0.8181	0.5478	0.7711	0.7440	0.7494	0.7226	0.3776	0.6465	0.8148	0.7154
		ReDeEP(chunk)	**0.8244**	**0.5566**	**0.7889**	0.7198	**0.7587**	**0.8420**	**0.5902**	**0.7070**	**0.8518**	**0.7603**



- ReDeEP consistently improves performance across two datasets, various backbone methods, and different metrics, validating its effectiveness in detecting RAG hallucinations.
- AARF can reduce hallucinations to a certain extent compared to the baseline model.

Can We Mitigate RAG Hallucinations Without Regeneration or Modifying LLMs?

LargePiG: Your Large Language Model is Secretly a Pointer Generator

Zhongxiang Sun* **Zihua Si**
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
{sunzhongxiang, zihua_si}@ruc.edu.cn

Xiaoxue Zang **Kai Zheng**
Kuaishou Technology Co., Ltd.
Beijing, China

Yang Song
Kuaishou Technology Co., Ltd.
Beijing, China
ys@sonyis.me

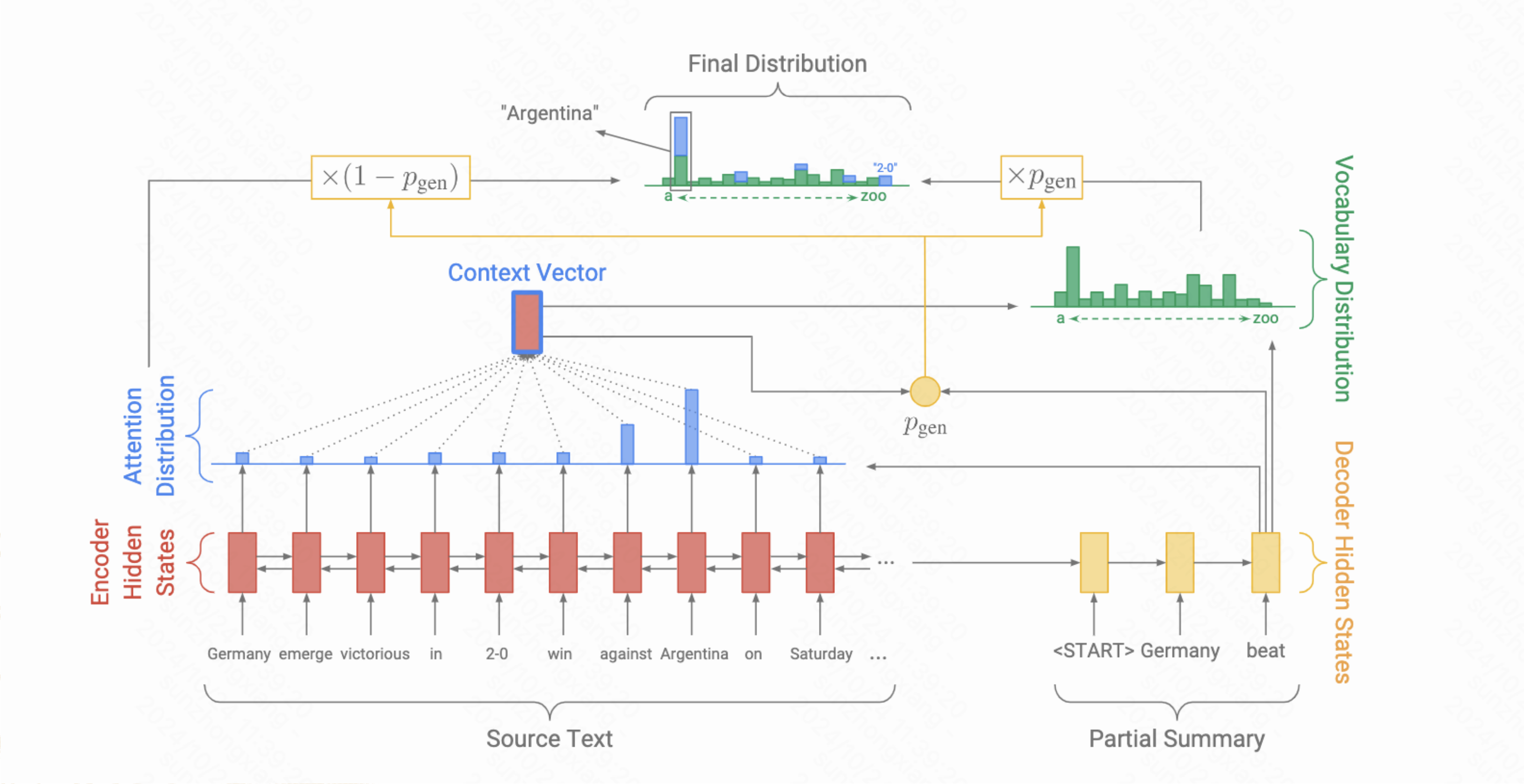
Xiao Zhang **Jun Xu**†
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
{zhangx89, junxu}@ruc.edu.cn

Accepted by The Web Conference 2025

Paper: https://arxiv.org/pdf/2410.11366

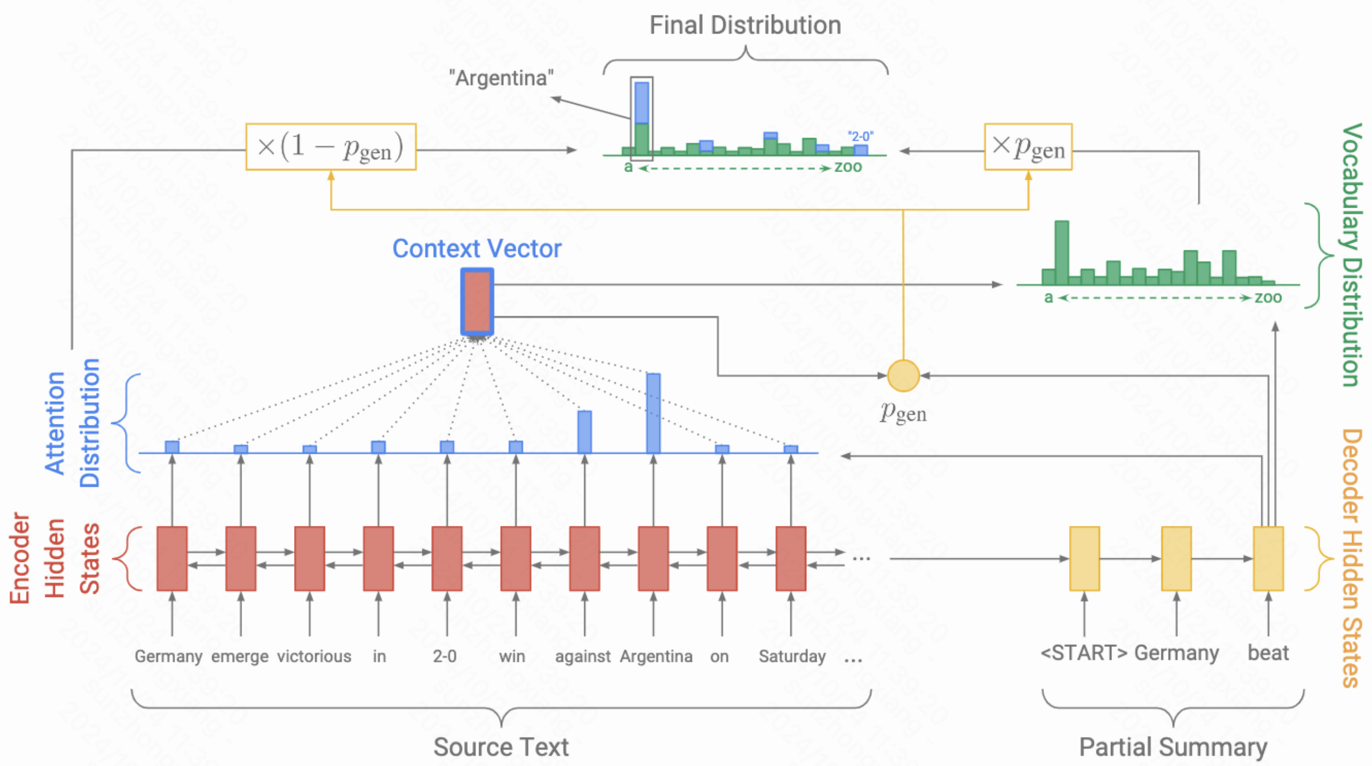
Pointer Generator

Get To The Point: Summarization with Pointer-Generator Networks



- Copying Factual Information from External Context
- Generating Syntactic and Other Information Using LLMs

Problem: Need Additional Training



Applying traditional Pointer Generator mechanisms to LLMs requires substantial computational resources and may disrupt the original representations of the LLM, potentially degrading its representational capacity.

Key Observation

Attention modules are more ‘truthful’ than other modules in LLMs (e.g., FFN modules).

- Knowledge is mainly stored in the FFN module of the transformer layer in pre-trained language model [1].
- Even if the self-attention module correctly focuses on the relevant token, the FFN module may still produce factuality hallucinations due to insufficient pre-training [2].

LLMs generate different types of words (function words and factual knowledge words) with distinct patterns.

Input: *Who was the first Nigerian to win the Nobel Prize, in which year?*
Output: *Wole Soyinka was the first Nigerian to win the Nobel Prize, in 1986.*

i-th early layer	_W	ole	_So	y	ink	a	_was	_the	_first	_Niger	ian	_to	_win	_the	_Nobel	_Prize	,	_in	_	1	9	8	6	.
30	1.9	0.0	0.03	1.76	0.0	0.0	6.45	0.29	0.07	0.6	0.01	0.48	0.13	0.1	0.02	0.11	2.97	1.84	0.12	0.0	0.0	0.0	7.56	0.23
28	4.78	0.04	0.42	10.5	0.05	0.07	3.65	0.21	0.02	0.63	0.0	0.29	0.17	0.02	0.04	0.02	4.77	1.89	6.13	9.76	12.4	15.16	16.86	0.16
26	11.41	3.15	7.15	12.67	5.28	3.5	1.22	0.08	0.02	0.75	0.0	0.18	0.15	0.12	0.05	0.04	3.77	1.19	4.58	16.56	19.31	18.66	19.67	0.13
24	13.21	8.6	10.01	14.28	8.99	8.44	0.8	0.26	0.02	0.44	0.0	2.51	0.08	7.37	0.06	0.04	2.08	0.71	6.68	18.72	23.84	21.68	21.31	0.1
22	14.26	18.81	11.61	15.7	12.34	9.29	0.75	4.57	0.03	0.24	0.0	2.4	0.09	6.57	0.05	0.02	2.03	0.38	8.27	17.82	22.89	22.98	21.46	2.07
20	10.18	15.95	12.99	16.32	13.52	11.07	1.85	9.78	0.03	0.06	0.04	0.39	0.73	6.28	0.02	0.03	11.41	4.36	9.19	16.84	19.57	20.38	19.45	10.26
18	7.75	15.97	12.59	16.46	14.52	12.25	7.76	8.33	5.15	6.47	2.48	5.73	10.67	7.41	1.29	8.92	13.57	10.99	12.59	14.02	19.57	16.98	15.63	12.9
16	8.99	16.05	12.81	17.45	15.47	13.52	9.8	11.18	10.73	10.97	12.1	11.4	14.52	13.09	10.34	11.86	14.34	12.16	13.7	13.73	19.44	17.05	15.85	13.47
14	9.06	16.14	13.33	17.83	16.24	14.0	10.63	13.03	12.78	12.66	15.07	13.2	16.06	14.71	13.61	13.61	14.09	12.04	14.19	14.4	19.76	17.17	16.24	12.87
12	9.75	16.3	13.47	17.92	16.45	14.94	11.52	13.95	14.11	13.92	15.82	14.23	16.76	15.6	14.81	14.42	14.47	13.48	14.47	15.02	19.44	17.4	16.45	13.57
10	10.22	16.4	13.63	18.1	16.24	15.52	12.4	14.54	14.71	14.2	16.34	14.85	16.78	15.66	15.02	15.06	14.53	13.8	14.13	14.96	19.63	17.7	16.62	13.42
8	10.66	16.57	14.04	18.24	16.2	16.21	12.66	14.42	15.09	14.09	16.82	14.71	16.88	15.57	15.2	15.31	14.44	13.89	14.47	15.15	19.93	17.93	16.81	13.9
6	10.68	16.49	14.2	18.38	16.3	16.62	13.18	14.53	15.4	14.27	17.81	15.44	16.98	15.82	15.43	15.8	14.27	14.16	14.65	15.54	19.79	18.2	17.14	13.92
4	10.65	16.59	14.31	18.53	16.38	16.77	13.43	15.02	15.99	14.53	18.29	15.5	17.29	16.33	15.9	16.14	14.31	14.53	14.69	15.81	19.93	18.38	17.4	14.25
2	10.8	16.69	14.29	18.64	16.74	16.9	13.36	15.23	15.97	14.76	18.68	15.45	17.31	16.71	16.05	16.46	14.58	14.51	14.84	16.02	20.13	18.6	17.67	14.44
0	11.0	16.69	14.51	18.78	16.82	17.09	13.54	15.6	16.47	14.88	19.12	15.88	17.45	16.98	16.26	16.87	14.85	15.34	15.16	16.34	20.46	18.79	17.83	14.95

[1] Dai D, Dong L, Hao Y, et al. Knowledge neurons in pretrained transformers[J]. 2021.
[2] Lv A, Zhang K, Chen Y, et al. Interpreting Key Mechanisms of Factual Recall in Transformer-Based Language Models[J]. 2024.

Mitigating RAG Hallucination from Decoding Side

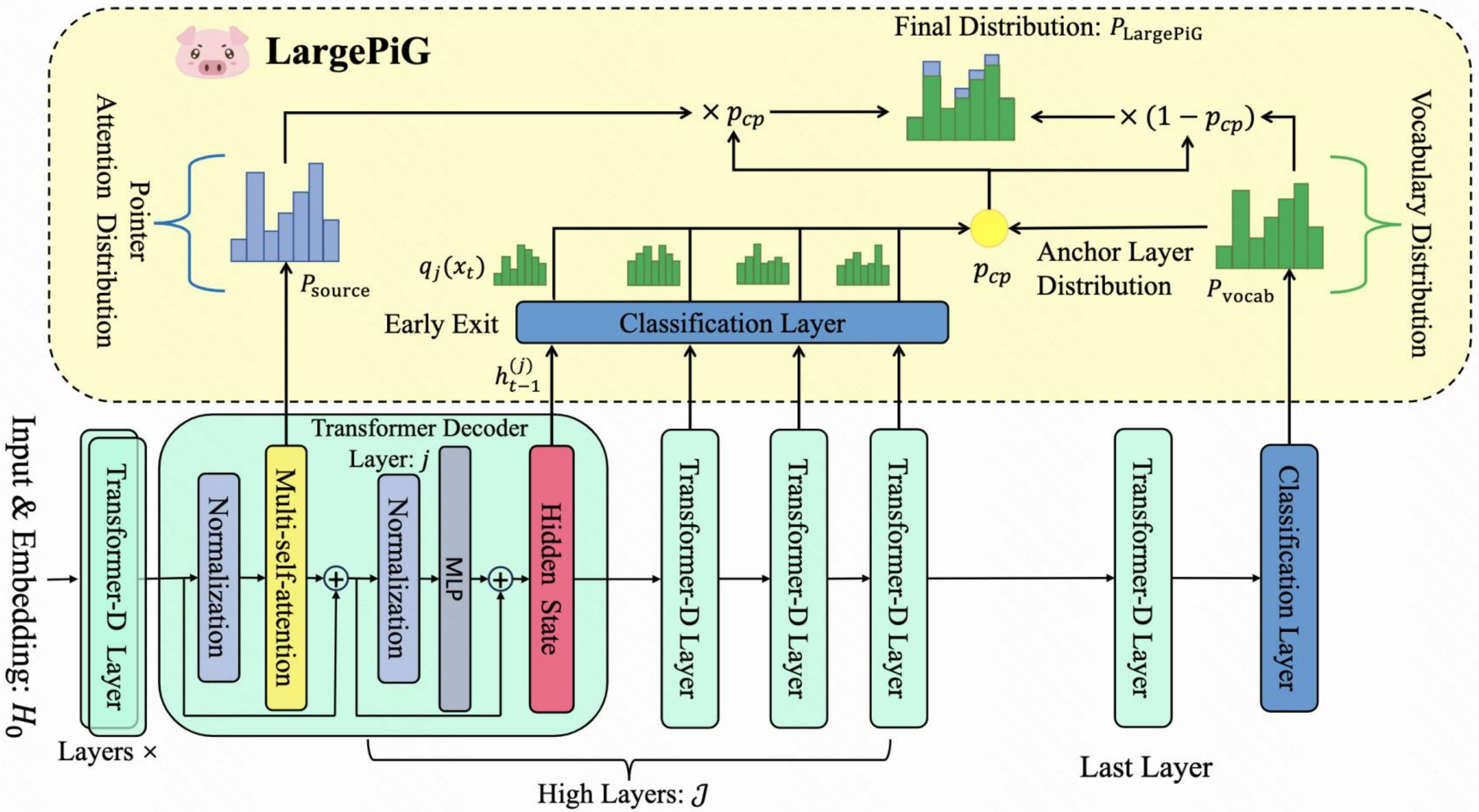


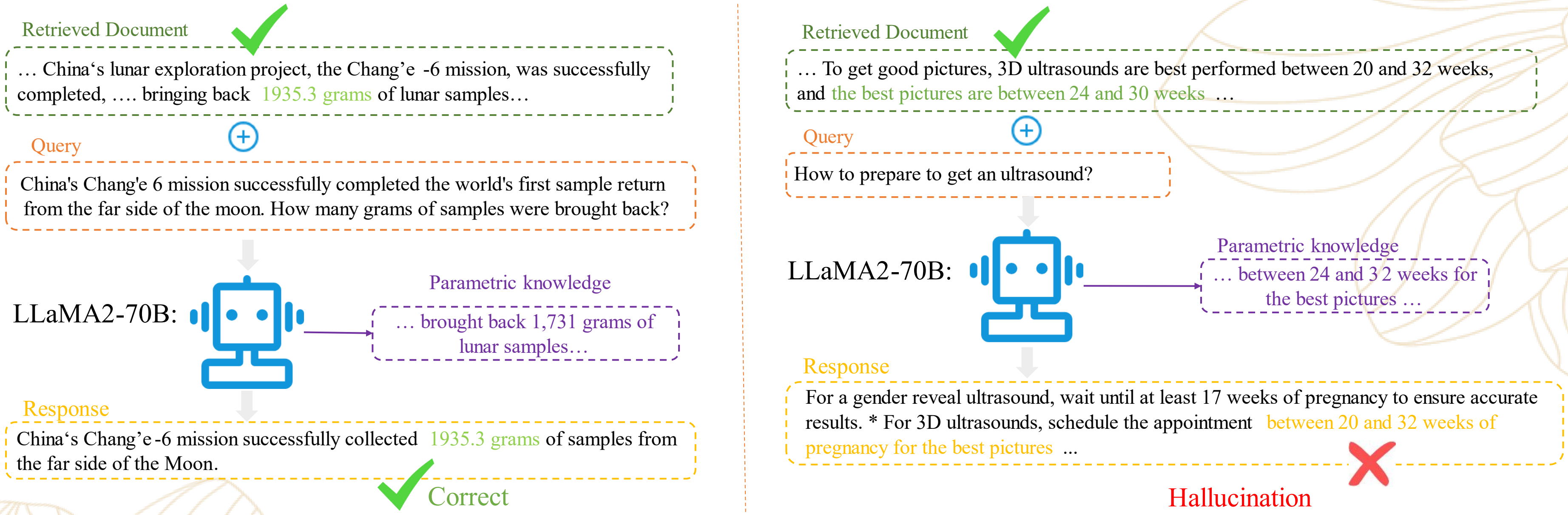
Figure 1: The architecture of the proposed plug-in and training-free method LargePiG. Pointer Attention Distribution (§ 2.1) from the LLM's self-attention weights, Vocabulary Distribution (§ 2.2) from the output of the original LLM, Copy Probability (§ 2.3) from the difference between the vocabulary distribution of the model's high layers and the last layer.

Experiment Results

Table 1: Performance comparisons between LargePiG and the baselines. The boldface represents the best performance. '†' means improvements are significant (paired t-test at p-value < 0.05).

Model	Qwen1.5 7B Chat						LLaMA2 7B Chat					
	TruthfulVQG			TruthfulDQG			TruthfulVQG			TruthfulDQG		
	MC1	MC2	MC3	MC1	MC2	MC3	MC1	MC2	MC3	MC1	MC2	MC3
Base	40.35	66.97	37.70	27.34	85.77	39.83	52.94	75.12	46.01	33.72	**71.61**	34.29
+ DoLa	37.97	64.73	35.68	23.52	85.05	37.09	52.79	75.25	46.10	35.09	69.97	33.19
+ LargePiG	**41.49**†	**68.12**†	**38.92**†	**29.91**†	**89.33**†	**42.18**†	**54.56**†	**76.15**†	**47.20**†	**37.23**†	70.95	**36.93**†
PQGR	43.61	70.08	41.26	25.86	77.23	36.86	52.22	74.21	45.60	32.28	65.74	31.41
+ DoLa	40.13	66.50	38.24	23.79	76.51	35.67	51.83	73.69	44.54	31.92	64.41	31.52
+ LargePiG	**45.52**†	**70.79**†	**42.54**†	**27.12**†	**79.20**†	**38.35**†	**52.87**†	**74.87**†	**46.27**†	**34.66**†	**68.34**†	**34.21**†
InPars	44.35	70.77	41.56	26.09	78.82	37.37	52.53	74.53	45.85	30.66	64.43	30.32
+ DoLa	40.35	66.90	38.48	24.48	77.57	36.96	51.59	74.33	44.86	29.87	63.97	29.52
+ LargePiG	**46.26**†	**71.51**†	**42.82**†	**27.34**†	**81.17**†	**38.53**†	**53.03**†	74.74	**46.20**†	**33.70**†	**67.30**†	**33.36**†
AQG	40.50	67.26	37.85	27.41	85.86	39.93	54.00	75.92	46.87	34.82	**71.62**	34.42
+ DoLa	37.99	64.65	35.62	25.59	85.28	39.21	52.79	75.25	46.10	33.02	70.96	33.17
+ LargePiG	**41.56**†	**68.13**†	**39.06**†	**29.99**†	**89.58**†	**42.35**†	**54.84**†	**76.73**†	**47.76**†	**37.09**†	71.04	**36.82**†

Conclusion: Detection



RAG Hallucination

Finding: The occurrence of RAG hallucinations is causally related to two primary factors: (1) while the Copying Heads may occasionally neglect necessary knowledge from the external context, a more prominent cause is the LLM losing the Copying Heads retrieved information during the generation process (RQ1-1, RQ2, § C), and (2) the Knowledge FFNs within LLM excessively injecting parametric knowledge into the residual stream (RQ1-2, RQ2, § D).

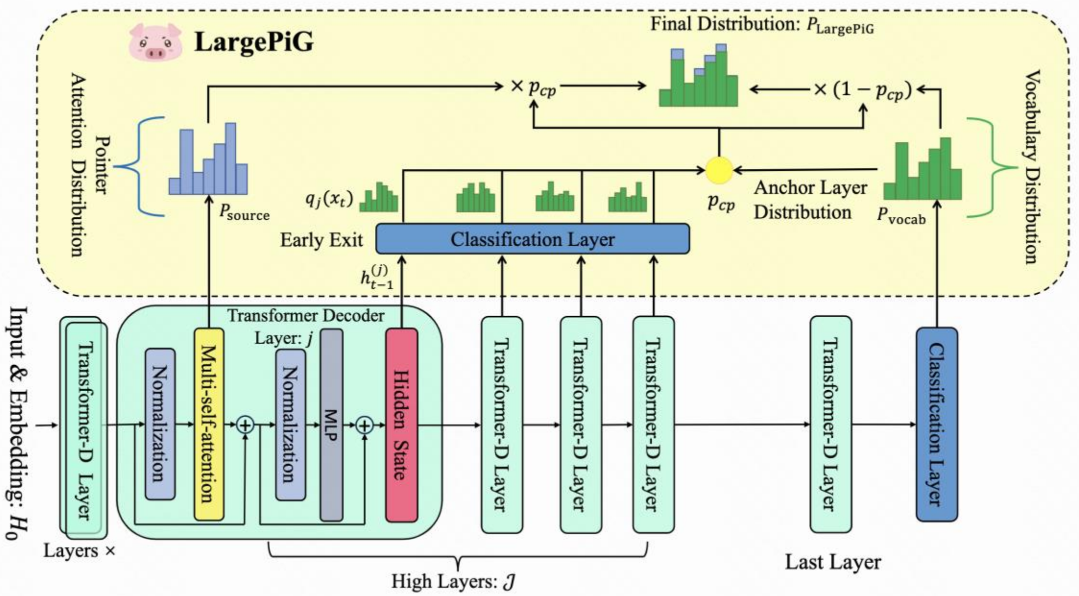
Conclusion: Mitigation

Model Side

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

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

Decoding Side



Discussion

Email: sunzhongxiang@ruc.edu.cn