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

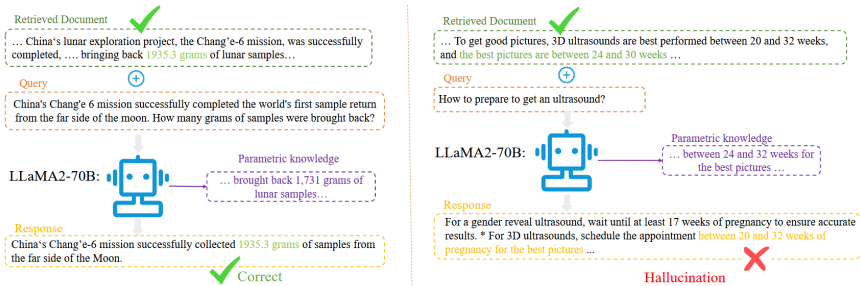
Zhongxiang Sun, Xiaoxue Zang, et al.

Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

Kuaishou Technology Co., Ltd., Beijing, China

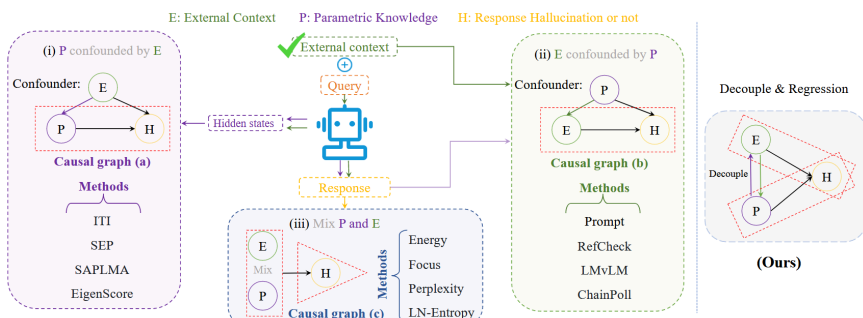
Accepted at the ICLR 2025 main conference (Spotlight)

Motivation

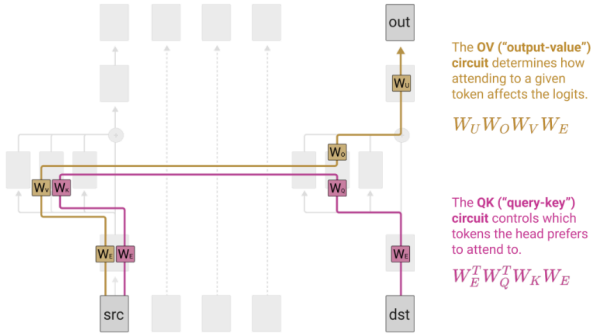


- The information retrieved by RAG could be an updated version of the model's internal knowledge.
- When the knowledge retrieved conflicts with the internal knowledge, it may lead to hallucinations in the response.

- Related work of detecting hallucinations
  - from hidden layer states
  - from input and response
  - by probability metrics
- This paper: mechanistic interpretability
  - measure the impact of external and internal knowledge on hallucinations separately
  - metrics are decoupled



BackGround



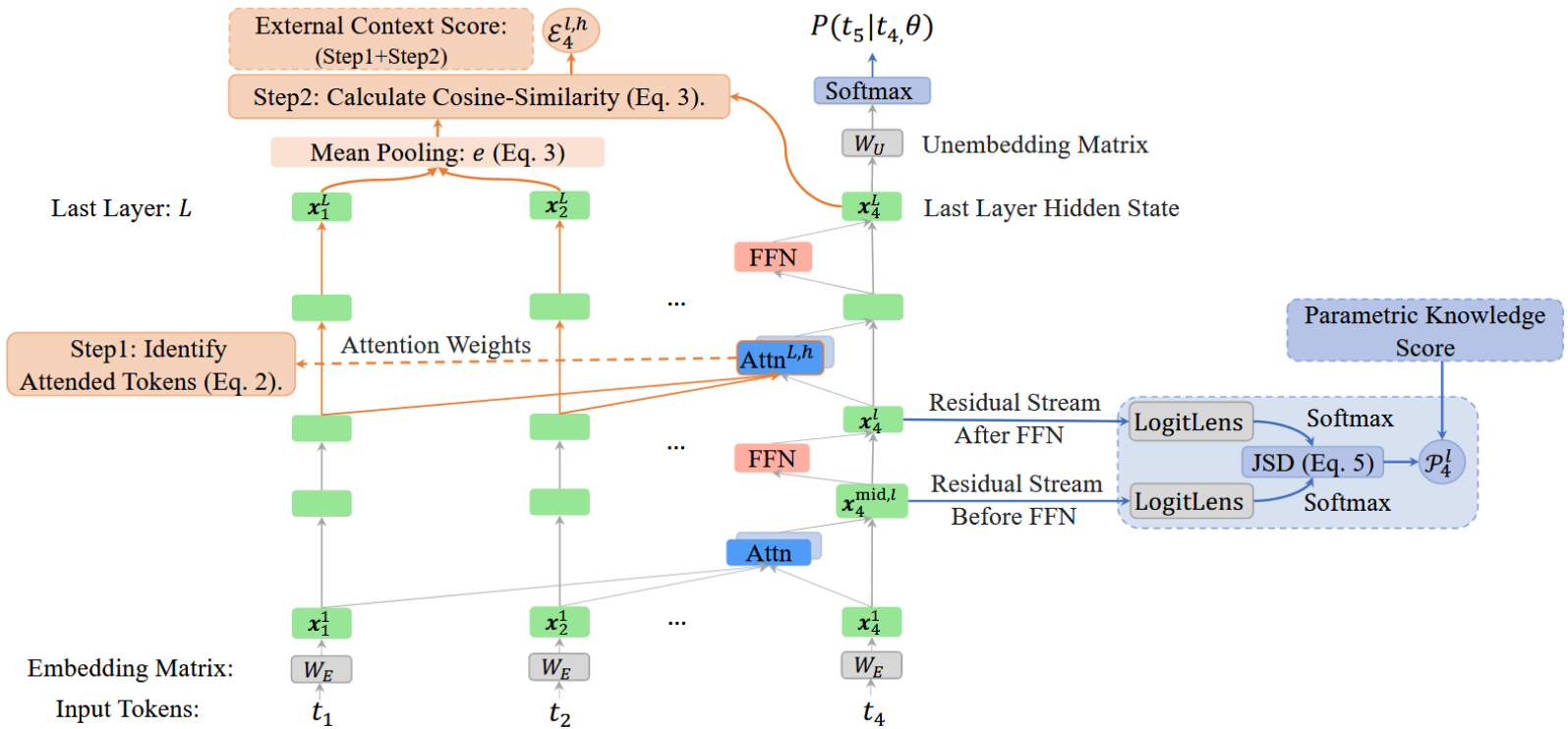
Some examples of large entries QK/OV circuit

Source Token	Destination Token	Out Token	Example Skip Tri-grams
" perfect"	" are", " looks", " is", " provides"	" perfect", " super", " absolute", " pure"	" perfect... are perfect", " perfect... looks super"
" large"	" contains", " using", " specify", " contain"	" large", " small", " very", " huge"	" large... using large", " large... contains small"
" two"	" One", "\n ", " has", "\r\n ", "One"	" two", " three", " four", " five", " one"	" two... One two", " two... has three"
"lambda"	" $\\", "}{\\", "+\\", "(\\", " ${\\"	"lambda", "sorted", " lambda", "operator"	"lambda... $\\lambda", "lambda... +\\lambda"
"nbsp"	"&", " \"&", "}&", ">&", "=&"	"nbsp", "01", "gt", "00012", "nbs", "quot"	"nbsp... ", "nbsp... > "
"Great"	"The", " The", " the", " contains", " /"	" Great", " great", " poor", " Every"	"Great... The Great", "Great... the great"

- Copying Heads
  - some attn heads copy information through the OV matrix
  - copy heads can be identified by analyzing the positive eigenvalues of the OV matrix
  - used to measure the utilization of **external information**
- FFN
  - stores the parameterized knowledge of the model
  - used to measure the utilization of **parameterized knowledge**

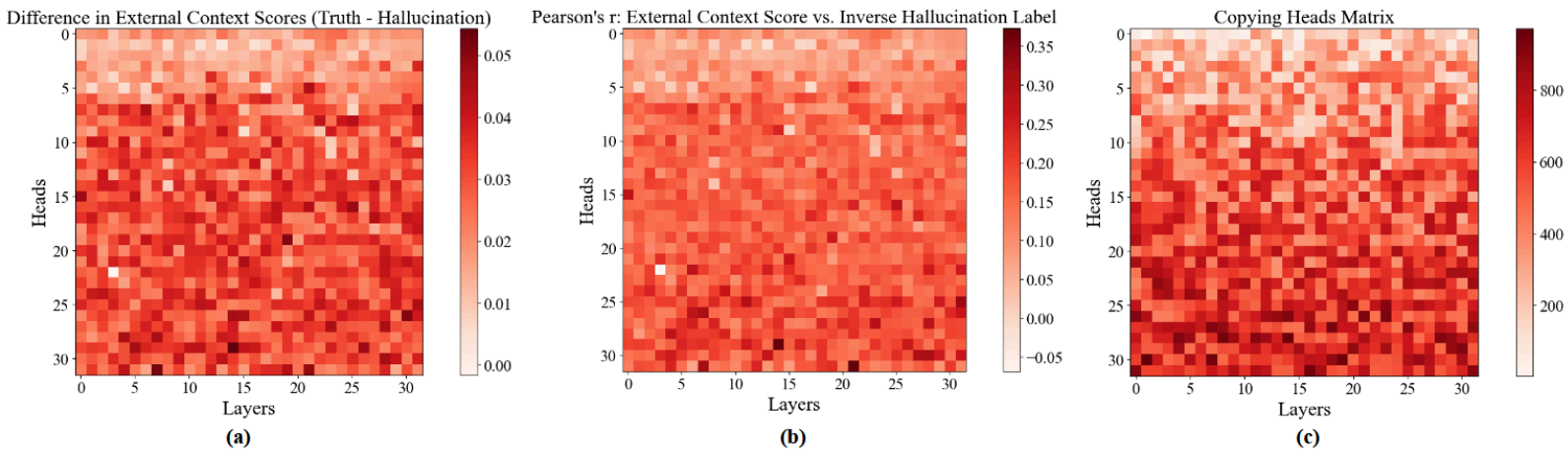
Method - ReDeEP

- External Context Score
- Parametric Knowledge Score



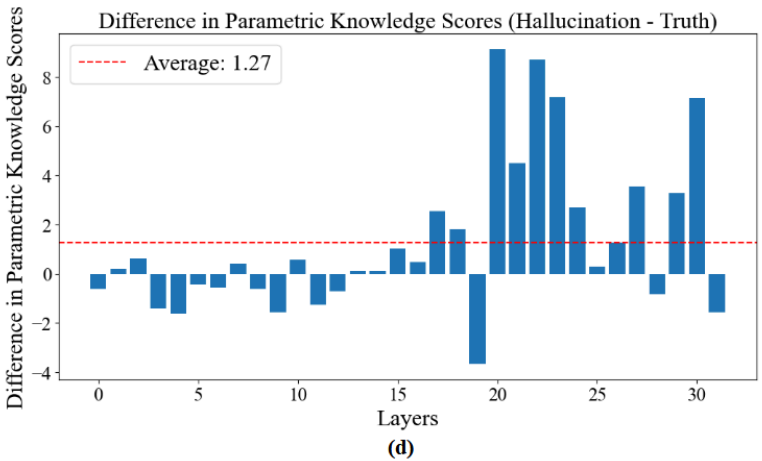
Empirical study - ECS

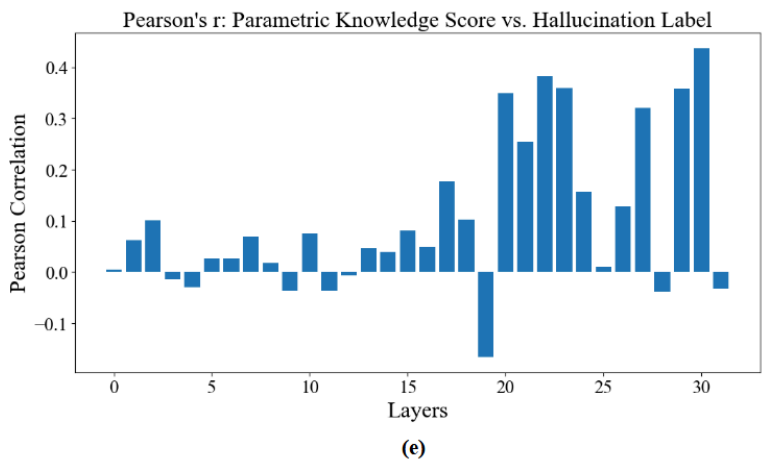
- a) LLMs utilize **less** external context information when generating hallucinations
- b) ECS shows **negative correlation** with the occurrence of hallucinations
- c) Attention heads associated with hallucinations are often **Copying Heads**



Empirical study - PKS

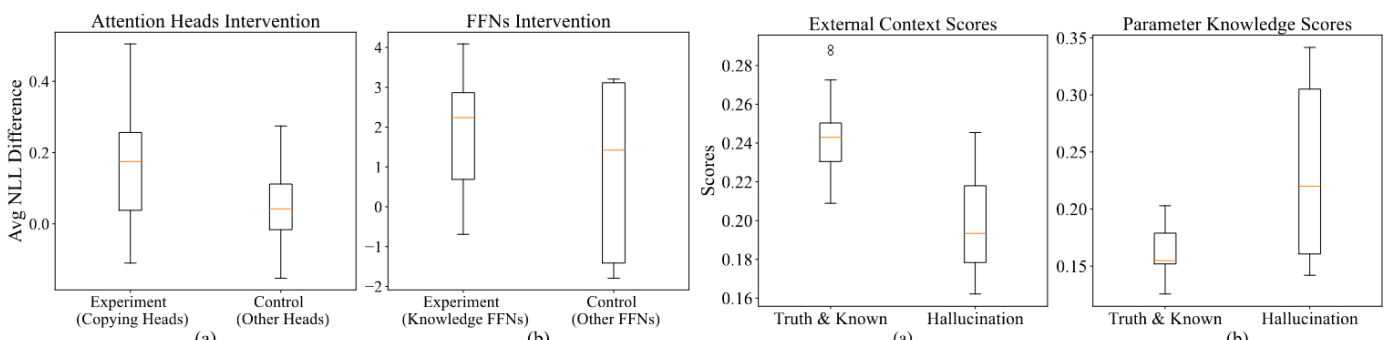
- d) PKS in the **later layers**' FFN modules are significantly **higher** in the hallucination
- e) PKS in the **later layers**' FFN modules are **positively correlated** with the hallucination





Empirical study - Causal Intervention

- Intervention Method
  - applied noise to the attention scores
  - amplified the contributions of FFN modules to the residual stream
- a,b) NLL difference was significantly **greater** than that of the control group for both attention heads and FFN modules
- c,d) when the LLM knows the truthful answer, Copying Heads **more accurately** capture and knowledge FFNs add **less** knowledge



Method - ReDeEP

- Token Level
  - linear regression leverages the high Pearson correlation

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

- Chunk Level ECS
  - compute the ECS for each chunk pair

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

- Chunk Level PKS
  - sum the token-level PKS for each chunk

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

Method - AARF

- Add Attention Reduce FFN

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

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

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 - Settings

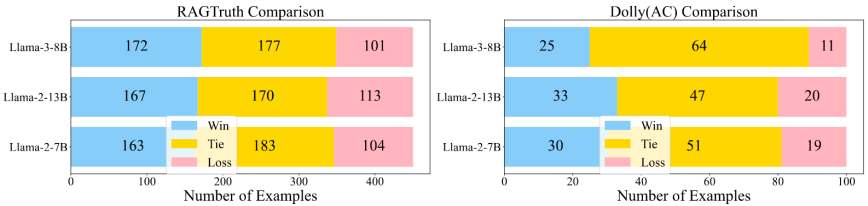
- Datasets
  - RAGTruth (ACL2024): high-quality, manually annotated RAG hallucination dataset
  - Dolly (AC): multiple tasks with accurate context
- Baselines
  - Parametric Confounded by External
    - EigenScore: semantic consistency in the embedding space
    - SEP: Semantic Entropy Probe
    - SAPLMA: trains a classifier on LLM activation values
    - ITI: trains a classifier on attention head activations
  - External Confounded by Parametric
    - Prompt: evaluate whether the LLM-generated responses are hallucinations using GPT-4-o-mini
    - LMvLM: multi-turn interaction between two language models to discover inconsistencies
    - P(True): the uncertainty of the generated claim by querying the LLM itself on the truthfulness of its generated response

Experiments - Results

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**

- The ReDeEP method performs the best
- The black-box method outperforms the white-box method

- Comparison between LLMs+AARF vs LLMs judged by GPT-4o



- using only PKS or only ECS does not achieve the same performance as the Full ReDeEP model

Table 3: Ablation Study of ReDeEP.

RAGTruth						
ReDeEP (Token)		AUC	PCC	**ReDeEP (Chunk)**		AUC	PCC
LLaMA2-7B	Only PKS	0.6950	0.3327	**LLaMA2-7B**	Only PKS	0.6180	0.2103
	Only ECS	0.7234	0.3779		Only ECS	0.7098	0.3944
	Full	0.7325	0.3979		Full	0.7458	0.4203
LLaMA2-13B	Only PKS	0.7214	0.3682	**LLaMA2-13B**	Only PKS	0.6614	0.2566
	Only ECS	0.8040	0.5201		Only ECS	0.7231	0.3922
	Full	0.8181	0.5478		Full	0.8244	0.5566
LLaMA3-8B	Only PKS	0.6102	0.1085	**LLaMA3-8B**	Only PKS	0.6082	0.1695
	Only ECS	0.7336	0.4312		Only ECS	0.6923	0.3102
	Full	0.7522	0.4493		Full	0.7285	0.3964