INSIDE: LLMS' INTERNAL STATES RETAIN THE POWER OF HALLUCINATION DETECTION

Атаев Азнаур, БПМИ203

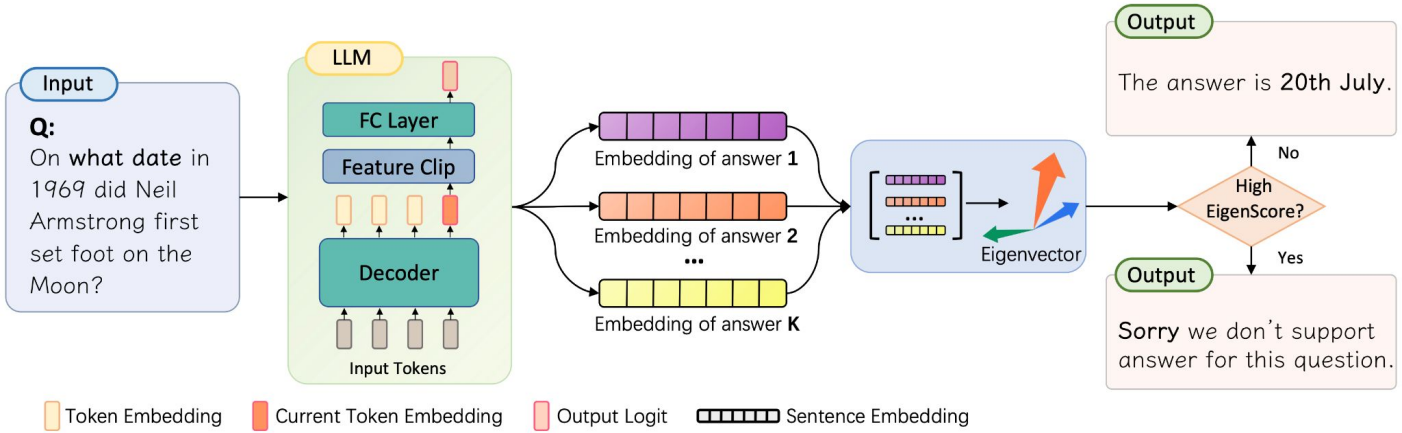
18.03.2024

The problem of hallucinations:

Type	Turn	Example
Input-conflicting Hallucination	*User*	Please summarize the following content: My friend Hill and I love playing basketball together. We spend hours on the court, dribbling, shooting, and competing against each other. Our passion for this game brings us closer as good friends.
	Model	Lucas and I love playing basketball and this makes us good friends.
Context-conflicting Hallucination	*User*	Tell me something about the NBA Commissioner.
	Model	In the world's most popular basketball league, the NBA, Commissioner Adam Silver has long been regarded as a leader and decision-maker. Firstly, Silver has claimed his dedication to maintaining fairness and competitiveness in the game. Secondly, Silver has repeatedly emphasized his concern for player health and well-being. In this case, Stern's words and actions clash with his stated goal of caring for player welfare.
Fact-conflicting Hallucination	*User*	Who was the mother of Afonso II, the third king of Portugal?
	Model	The mother of Afonso II was Queen Urraca of Castile.

Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models

Proposed hallucination detection pipeline:



1. Hallucination Detection By EigenScore

2. Test Time Feature Clipping

INSIDE: LLMS' INTERNAL STATES RETAIN THE POWER OF HALLUCINATION DETECTION

EigenScore:

Sentence embedding:

$$\boldsymbol{z} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{h}_t$$

$$\boldsymbol{z} = \boldsymbol{h}_T$$

$\boldsymbol{h}_T$ – T-token of sentence of length T

Embedding covariance matrix:

$$\boldsymbol{\Sigma} = \mathbf{Z}^\top \cdot \mathbf{J}_d \cdot \mathbf{Z}$$

$$\boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}$$

$$\mathbf{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_K] \in \mathbb{R}^{d \times K}$$

$$\mathbf{J}_d = \boldsymbol{I}_d - \frac{1}{d}\mathbf{1}_K \mathbf{1}_K^\top$$

INSIDE: LLMS' INTERNAL STATES RETAIN THE POWER OF HALLUCINATION DETECTION

EigenScore:

$$E(\mathcal{Y}|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{K}\log\det(\boldsymbol{\Sigma} + \alpha \cdot \mathbf{I}_K)$$

$$E(\mathcal{Y}|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{K}\log(\prod_i \lambda_i) = \frac{1}{K}\sum_i^K \log(\lambda_i)$$

$$\begin{aligned}\det(A-\lambda I) = p(\lambda) &= (-1)^n(\lambda-\lambda_1)(\lambda-\lambda_2)\cdots(\lambda-\lambda_n)\\ &= (-1)(\lambda-\lambda_1)(-1)(\lambda-\lambda_2)\cdots(-1)(\lambda-\lambda_n)\\ &= (\lambda_1-\lambda)(\lambda_2-\lambda)\cdots(\lambda_n-\lambda)\end{aligned}$$

INSIDE: LLMS' INTERNAL STATES RETAIN THE POWER OF HALLUCINATION DETECTION

EigenScore:

$$H_e(X) = -\sum_X -p(x)\log p(x)$$

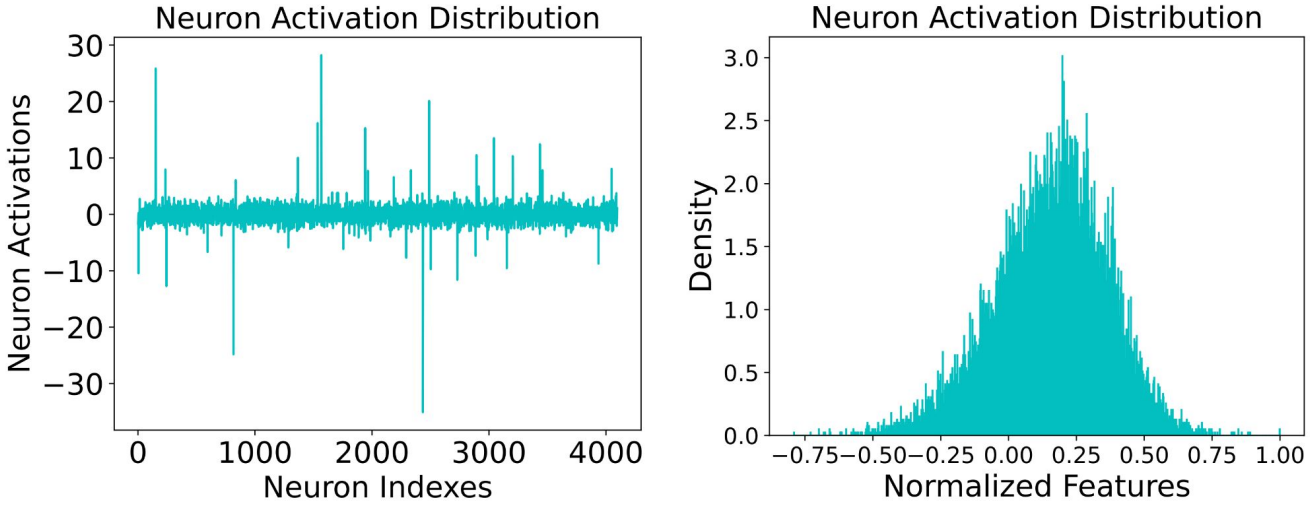
$$H_{de}(X) = -\int_x f(x)\log f(x)dx$$

$$H_{de}(X) = \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{d}{2}(\log 2\pi + 1) = \frac{1}{2}\sum_{i=1}^{d}\log\lambda_i + C$$

$$X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

INSIDE: LLMS' INTERNAL STATES RETAIN THE POWER OF HALLUCINATION DETECTION

Test Time Feature Clipping:



(a) Neuron Activation (b) Feature Distribution

Illustration of activation distributions in the penultimate layer of LLaMA-7B. (a) Activation distribution in the penultimate layer for a randomly sampled token. (b) Activation distribution for a randomly sampled neuron activation of numerous tokens.

INSIDE: LLMS' INTERNAL STATES RETAIN THE POWER OF HALLUCINATION DETECTION

Test Time Feature Clipping:

$$FC(h) = \begin{cases} h_{min}, & h < h_{min} \\ h, & h_{min} \leq h \leq h_{max} \\ h_{max} & h > h_{max} \end{cases}$$

h_{min} 0.2 percentile of feature of hidden embedding

h_{max} 0.8 percentile of feature of hidden embedding

INSIDE: LLMS' INTERNAL STATES RETAIN THE POWER OF HALLUCINATION DETECTION

Experiments:

Perplexity: $$P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = -\frac{1}{T}\log\prod_t p(y_t|y_{<t},\boldsymbol{x}) = -\frac{1}{T}\sum_t \log p(y_t|y_{<t},\boldsymbol{x})$$

Length Normalized Entropy: $$H(\mathcal{Y}|\boldsymbol{x},\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{y}\in\mathcal{Y}}\frac{1}{T_{\boldsymbol{y}}}\sum_t \log p(y_t|y_{<t},\boldsymbol{x})$$

Lexical Similarity: $$S(\mathcal{Y}|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{C}\sum_{i=1}^{K}\sum_{j=i+1}^{K} sim(\boldsymbol{y}^i,\boldsymbol{y}^j)$$

where C = K · (K − 1)/2 and sim(·, ·) is the similarity defined by Rouge-L

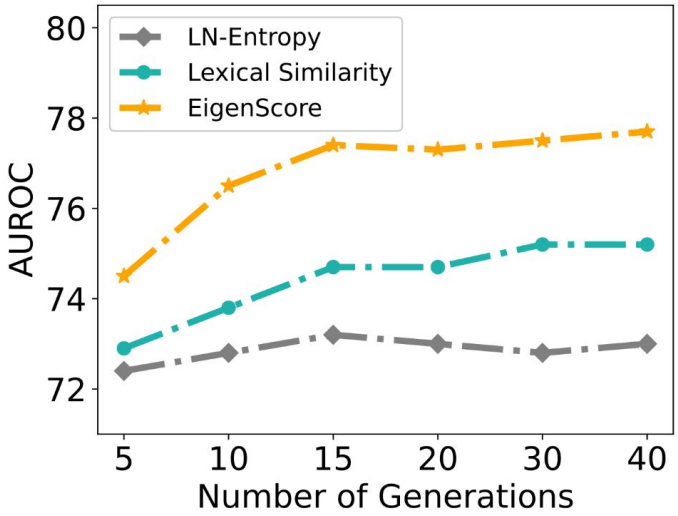
Experiments:

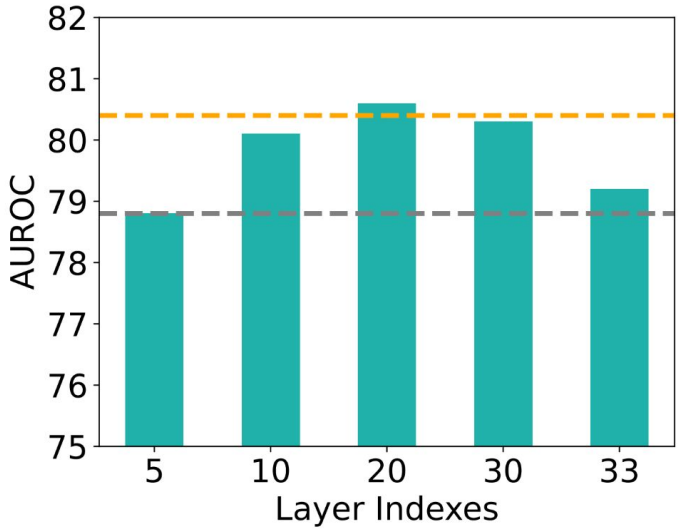
Models	Datasets / Methods	CoQA			SQuAD			NQ			TriviaQA		
		AUC_s	AUC_r	PCC	AUC_s	AUC_r	PCC	AUC_s	AUC_r	PCC	AUC_s	AUC_r	PCC
LLaMA-7B	Perplexity	64.1	68.3	20.4	57.5	60.0	10.2	74.0	74.7	30.1	**83.6**	83.6	54.4
	Energy	51.7	54.7	1.0	45.1	47.6	-10.7	64.3	64.8	18.2	66.8	67.1	29.1
	LN-Entropy	68.7	73.6	30.6	70.1	70.9	30.0	72.8	73.7	29.8	83.4	83.2	54.0
	Lexical Similarity	74.8	77.8	43.5	74.9	76.4	44.0	73.8	75.9	30.6	82.6	**84.0**	55.6
	EigenScore	**80.4**	**80.8**	**50.8**	**81.5**	**81.2**	**53.5**	**76.5**	**77.1**	**38.3**	82.7	82.9	**57.4**
LLaMA-13B	Perplexity	63.2	66.2	20.1	59.1	61.7	14.2	73.5	73.4	36.3	**84.7**	**84.5**	56.5
	Energy	47.5	49.2	-5.9	36.0	39.2	-20.2	59.1	59.8	14.7	71.3	71.5	36.7
	LN-Entropy	68.8	72.9	31.2	72.4	74.0	36.6	74.9	75.2	39.4	83.4	83.1	54.2
	Lexical Similarity	74.8	77.6	44.1	77.4	79.1	48.6	74.9	76.8	40.3	82.9	84.3	57.5
	EigenScore	**79.5**	**80.4**	**50.2**	**83.8**	**83.9**	**57.7**	**78.2**	**78.1**	**49.0**	83.0	83.0	**58.4**
OPT-6.7B	Perplexity	60.9	63.5	11.5	58.4	69.3	8.6	76.4	77.0	32.9	**82.6**	**82.0**	**50.0**
	Energy	45.6	45.9	-14.5	41.6	43.3	-16.4	60.3	58.6	25.6	70.6	68.8	37.3
	LN-Entropy	61.4	65.4	18.0	65.5	66.3	22.0	74.0	76.1	28.4	79.8	80.0	43.0
	Lexical Similarity	71.2	74.0	38.4	72.8	74.0	39.3	71.5	74.3	23.1	78.2	79.7	42.5
	EigenScore	**76.5**	**77.5**	**45.6**	**81.7**	**80.8**	**49.9**	**77.9**	**77.2**	33.5	80.3	80.4	0.485

Experiments:

Model	LLaMA-7B				OPT-6.7B			
Datasets	CoQA		NQ		CoQA		NQ	
Methods	AUC_s	PCC	AUC_s	PCC	AUC_s	PCC	AUC_s	PCC
LN-Entropy	68.7	30.6	72.8	29.8	61.4	18.0	74.0	28.4
LN-Entropy + FC	70.0	33.4	73.4	31.1	62.6	21.4	74.8	30.3
Lexical Similarity	74.8	43.5	73.8	30.6	71.2	38.4	71.5	23.1
Lexical Similarity + FC	76.6	46.3	74.8	32.1	72.6	40.2	72.4	24.2
EigenScore (w/o)	79.3	48.9	75.9	38.3	75.3	43.1	77.1	32.2
EigenScore	80.4	50.8	76.5	38.3	76.5	45.6	77.9	33.5

Experiments:





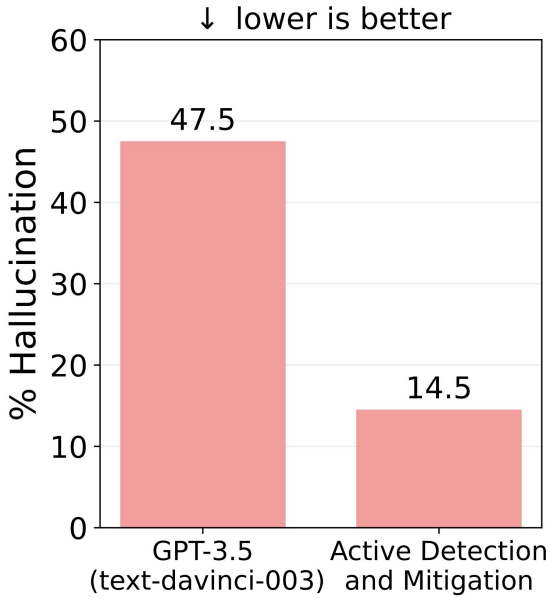
Any questions?

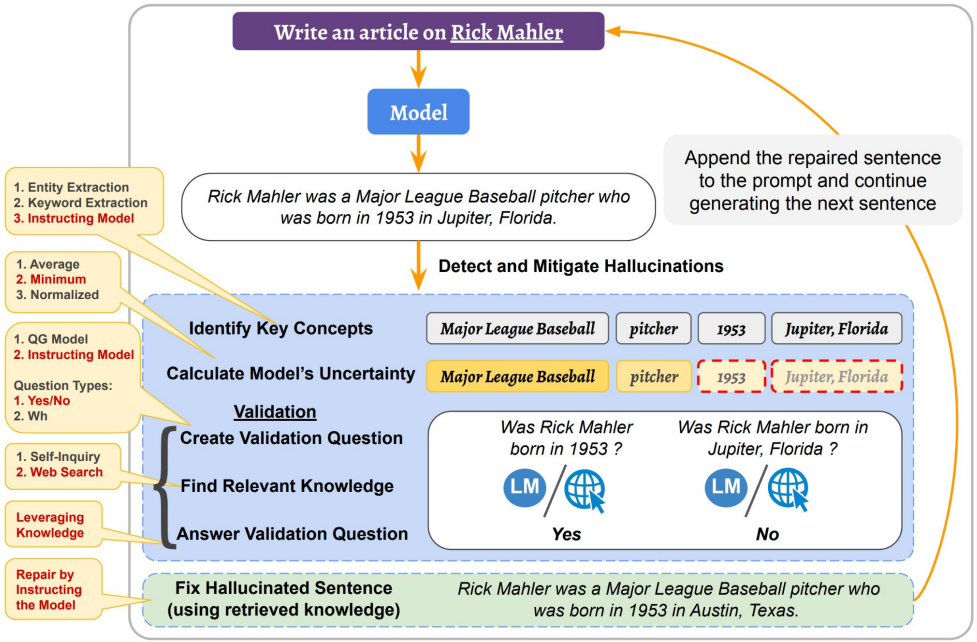
Other types of hallucination detection and mitigation:

1. Embeddings level hallucination detection methods
2. Logit level hallucination detection methods
3. Prompt level hallucination detection methods

здесь должна быть красивая картинка с пайплайном работы LLMки и всякими стрелочками из 1-3 в картинку, но так как здесь этот текст - я не успел :<

A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation





A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

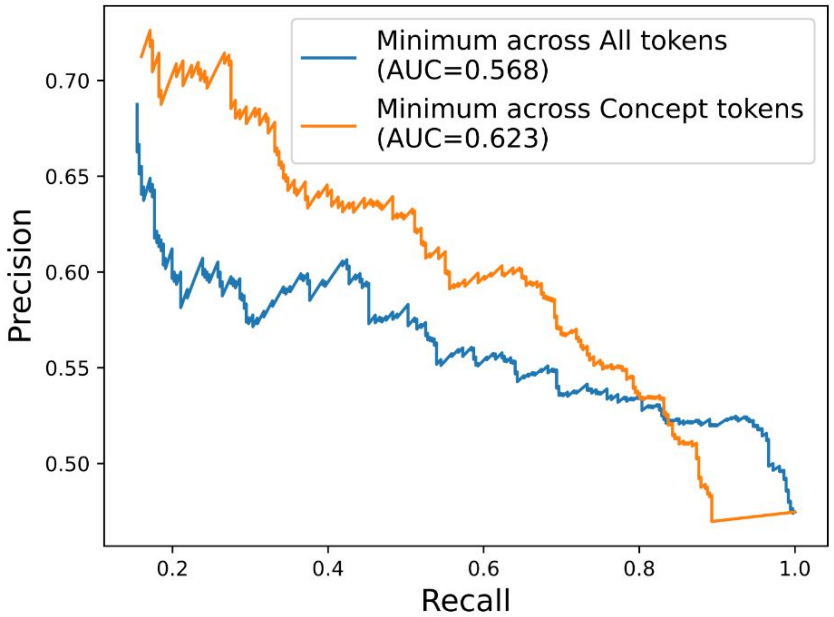
Identify key concepts stage:

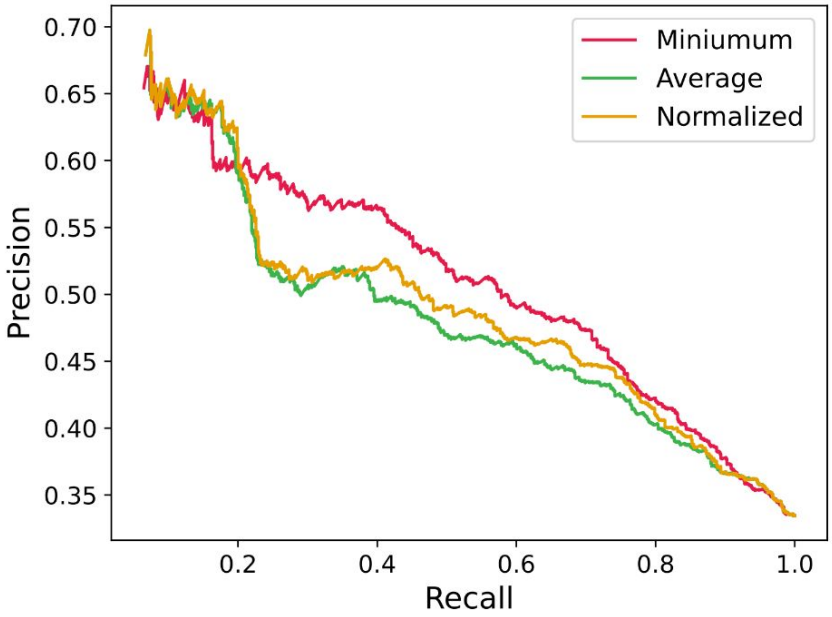
Text	Entity Extraction	Keyword Extraction	Instructing Model
John Russell Reynolds was an English physician and neurologist who made significant contributions to the field of neurology.	John Russell Reynolds, English	John Russell Reynolds, English, physician, neurologist, significant contributions, field, neurology	John Russell Reynolds, English, physician, neurologist, neurology
He was born in London in 1820 and studied medicine at the University of London.	London, 1820, the University of London	born, London, 1820, studied medicine, University, London	London, 1820, medicine, University of London
After college, he worked as a lawyer for the PGA Tour, eventually becoming the Tour's Deputy Commissioner in 1989.	the PGA Tour, Tour, 1989	college, worked, lawyer, PGA, Tour, eventually, Tour, Deputy Commissioner	college, lawyer, PGA Tour, Deputy Commissioner, 1989
He was born in Sydney in 1971 and grew up in the city's western suburbs.	Sydney, 1971	born, Sydney, 1971, grew, city, suburbs	Sydney, 1971, western suburbs

Table 7: Examples of concepts identified by different techniques.

A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

Calculate Model's Uncertainty stage:





A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

Validation Question stage:

Input	Generated Sentence	Concept	Validation Question
Write an article about John Russell Reynolds	Reynolds was born in **London** in **1820** and studied **medicine** at the **University of London**.	**London**	[Y/N] Was John Russell Reynolds born in London? [Wh] Where was John Russell Reynolds born?
		1820	[Y/N] Was John Russell Reynolds born in 1820? [Wh] What year was John Russell Reynolds born?
		medicine	[Y/N] Did John Russell Reynolds study medicine? [Wh] What did John Russell Reynolds study at the University of London?
		University of London	[Y/N] Did Reynolds study medicine at the University of London? [Wh] What university did John Russell Reynolds study medicine at?

Table 8: Examples of validation questions corresponding to the identified keyphrases generated by Instructing the Model technique.

A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

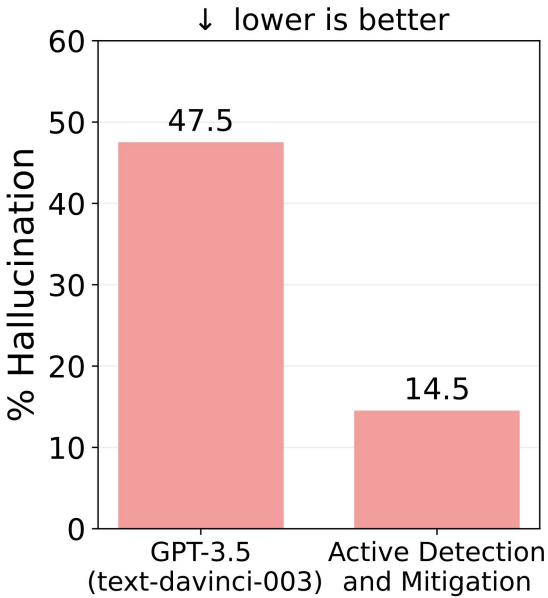
Hallucination mitigation stage:

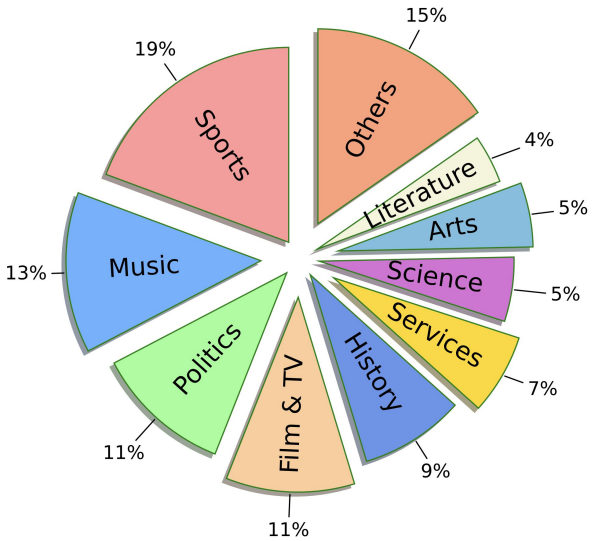
Step	Prompt
Input Prompt	`Write an article about {topic}`
Identify Important Concepts	`Identify all the important keyphrases from the above sentence and return a comma separated list.`
Create Validation Question	`For the above sentence about {topic}, generate a yes/no question that tests the correctness of {concept}.`
Answer Validation Question	`{search results} Answer the below question about topic in Yes or No based on the above context. {validation question}.`
Repair Hallucinated Sentence	`The above sentence has information that can not be verified from the provided evidence, repair that incorrect information and create a new sentence based on the provided evidence.`

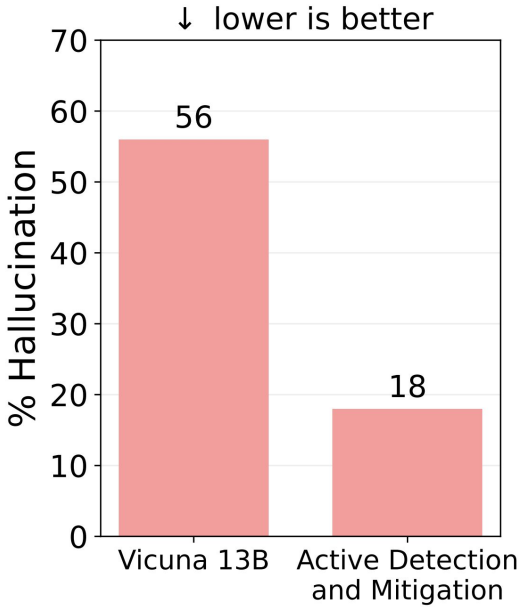
Table 6: Instructional Prompts corresponding to different steps of our approach.

A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

Experiments setup:







A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

Correction with Backtracking Reduces Hallucination in Summarization

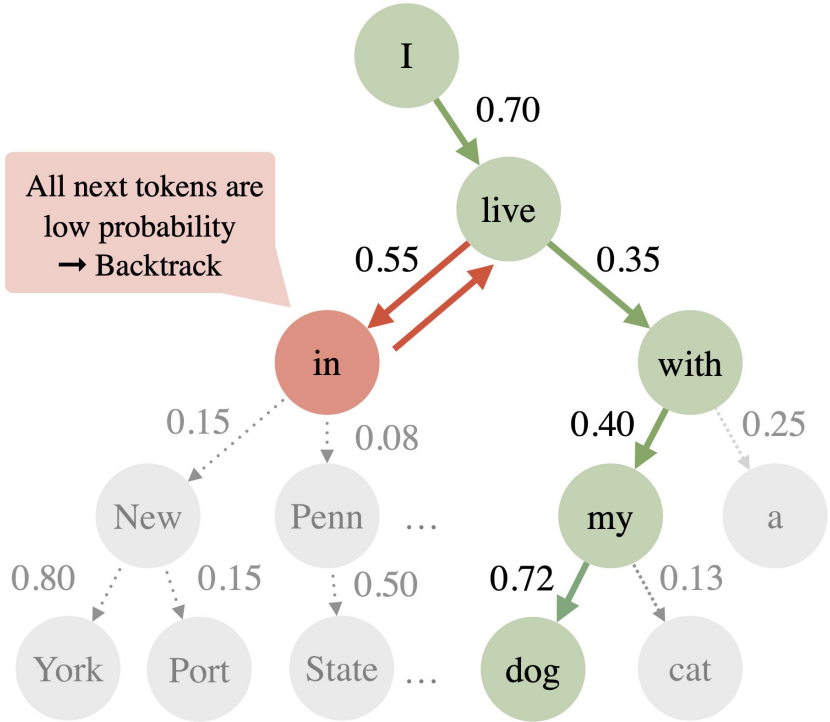
Context Document

I share my home with a loyal and affectionate companion - my dog. Living together has brought joy, companionship, and a unique bond into my life. Each day is marked by our shared adventures, whether it's going for long walks, playing fetch in the park, or simply enjoying quiet moments at home. Their unwavering presence brings comfort and a sense of connection, making every day brighter and more fulfilling.

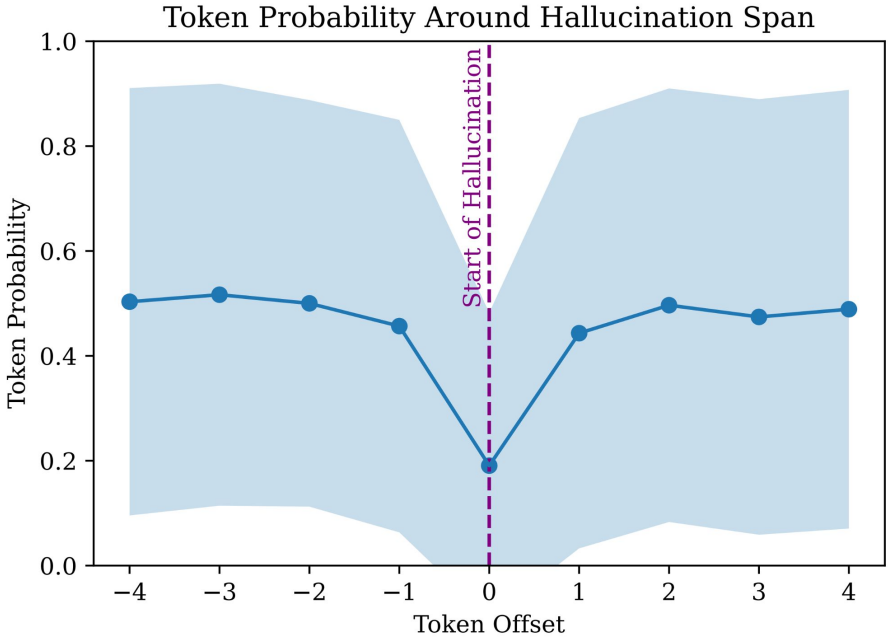
Greedy Decoding: I live in New York . ✗

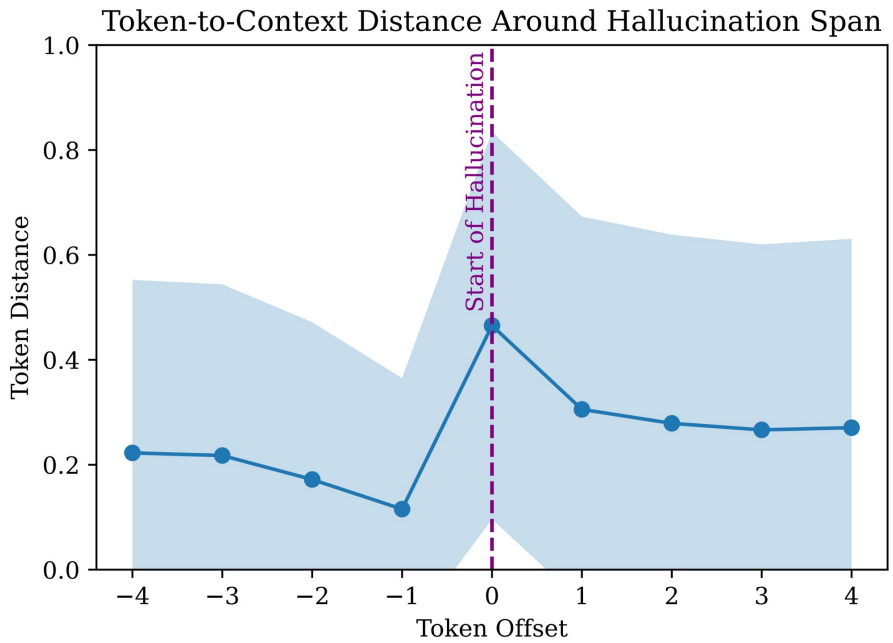
Greedy+Backtrack: I live with my dog. ✓

Summary



Correction with Backtracking Reduces Hallucination in Summarization





Correction with Backtracking Reduces Hallucination in Summarization

Experiments:

	Method	AlignScore↑	FactCC↑	BS-Fact↑	Rouge-L↑
Newsroom	Greedy	0.701	0.321	0.897	0.161
	+ CAD	0.706	0.247	0.910	0.170
	+ CoBa	0.715	0.328	0.906	0.162
	+ CoBa-d	0.729	0.335	0.906	0.164
XSUM	Greedy	0.798	0.406	0.931	0.221
	+ CAD	0.783	0.335	0.931	0.237
	+ CoBa	0.800	0.410	0.932	0.221
	+ CoBa-d	0.805	0.418	0.933	0.223
CNN/DM	Greedy	0.750	0.316	0.900	0.152
	+ CAD	0.740	0.251	0.919	0.176
	+ CoBa	0.753	0.323	0.902	0.153
	+ CoBa-d	0.759	0.327	0.902	0.154