NEURAL INFORMATION PROCESSING SYSTEMS

LLM-Check: Investigating Detection of Hallucinations in LLMs






Gaurang Sriramanan | Siddhant Bharti | Vinu Shankar | Shoumik Saha | Priyatham Kattakinda | Soheil Feizi

UNIVERSITY OF MARYLAND 1856

What are Hallucinations in LLMs?

- Model outputs that are fallacious, incorporating fictional details
- Responses are partly misleading or entirely fabricated
- Seem plausible, appearing tenable before further scrutiny

What are Hallucinations in LLMs?

- Model outputs that are fallacious, incorporating fictional details
- Responses are partly misleading or entirely fabricated
- Seem plausible, appearing tenable before further scrutiny

- Present a significant challenge towards the deployment of LLMs
- Mitigation is seen to be very challenging practically

What are Hallucinations in LLMs?

Let’s look at some real examples:

Prompt: Who are some notable alumni from Monta Vista High School in Cupertino?

Claude-3 Sonnet: Monta Vista High School in Cupertino, California has had some notable alumni, though I don't have fully comprehensive information about all of their accomplished former students. A few that stand out are:

What are Hallucinations in LLMs?

Let’s look at some real examples:

Prompt: Who are some notable alumni from Monta Vista High School in Cupertino?

Claude-3 Sonnet: Monta Vista High School in Cupertino, California has had some notable alumni, though I don't have fully comprehensive information about all of their accomplished former students. A few that stand out are:

Steve Wozniak - Co-founder of Apple Computer. He attended Monta Vista and graduated in 1971.

Kamala Harris - Current Vice President of the United States. She graduated from Monta Vista in 1981.

Jeremy Lin - Professional basketball player who had a stint in the NBA and gained fame during 'Linsanity' with the New York Knicks in 2012. He graduated from Monta Vista in 2006.

Amanda Crew - Actress known for films like Silicon Valley and Haunting of Molly Hartley.

What are Hallucinations in LLMs?

Most popular models hallucinate!

Prompt: Who are some notable alumni from Monta Vista High School in Cupertino?

- ❌ **Claude-3 Sonnet**
- ✅ **Gemini**
- ❌ **GPT-4**
- ❌ **GPT-4o**
- ❌ **ChatGPT**

Why can’t LLMs help themselves from fabricating details?

- LLMs are really well trained - but not optimally so!
- Appreciable degree of world-knowledge

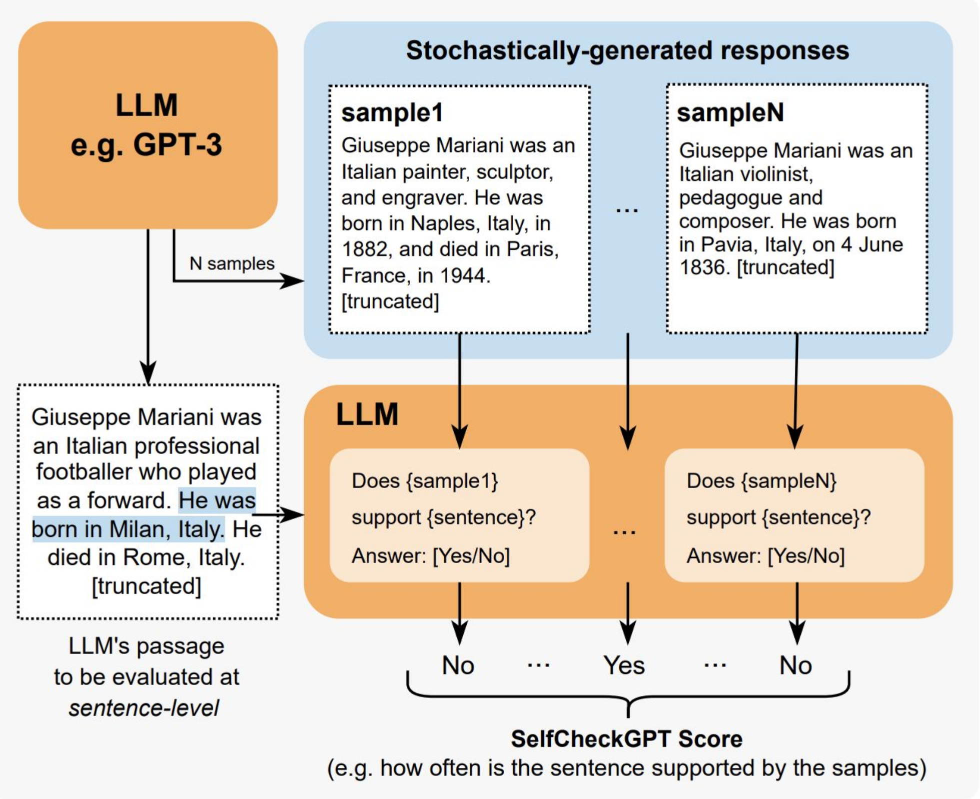
Why can’t LLMs help themselves from fabricating details?

- LLMs are really well trained - but not optimally so!
- Appreciable degree of world-knowledge
- Autoregressive generation: once a token is sampled, it’s fixed!
- LLM attempts to nonetheless maximize likelihood of overall response

Why can't LLMs help themselves from fabricating details?

- LLMs are really well trained - but not optimally so!
- Appreciable degree of world-knowledge
- Autoregressive generation: once a token is sampled, it's fixed!
- LLM attempts to nonetheless maximize likelihood of overall response
- Hallucinations are absent in some of the repeated model generations for the same prompt
- Consistency across different generations can be leveraged

Multi-Response Consistency-Based Detection Methods

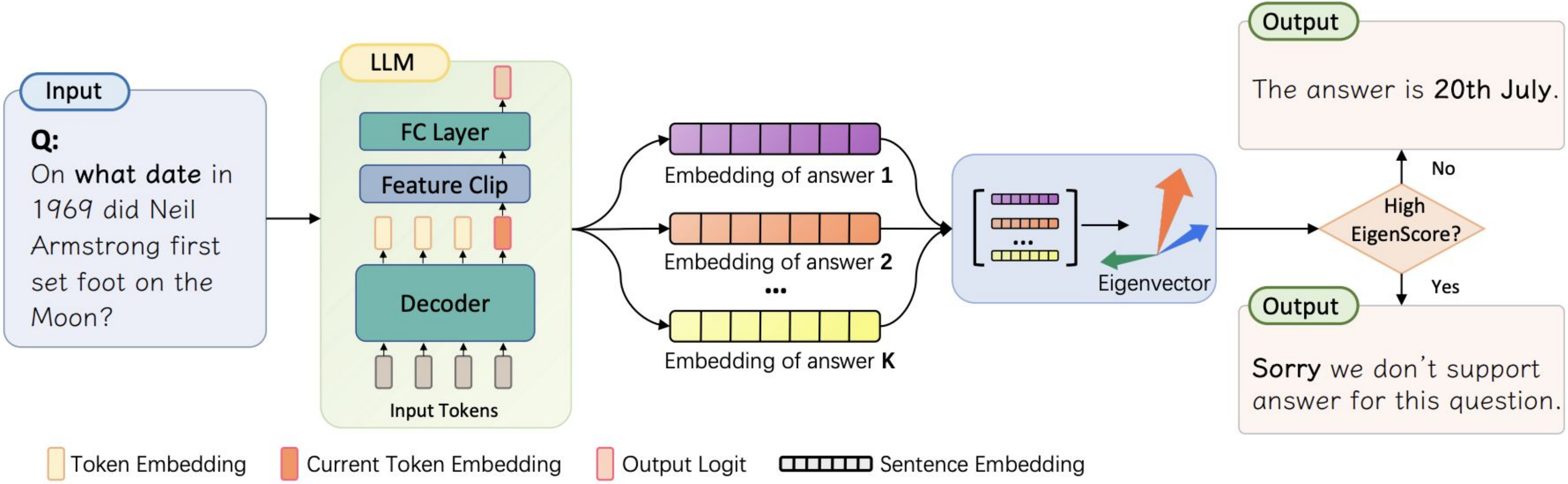


Detection scores with:

1. BERTScore
2. Question Answering
3. N-gram Analysis
4. Natural Language Inference
5. SelfCheckGPT - Prompt

```
------------------------------------------------
Context: {}
Sentence: {}
Is the sentence supported by the context above?
Answer Yes or No:
------------------------------------------------
```

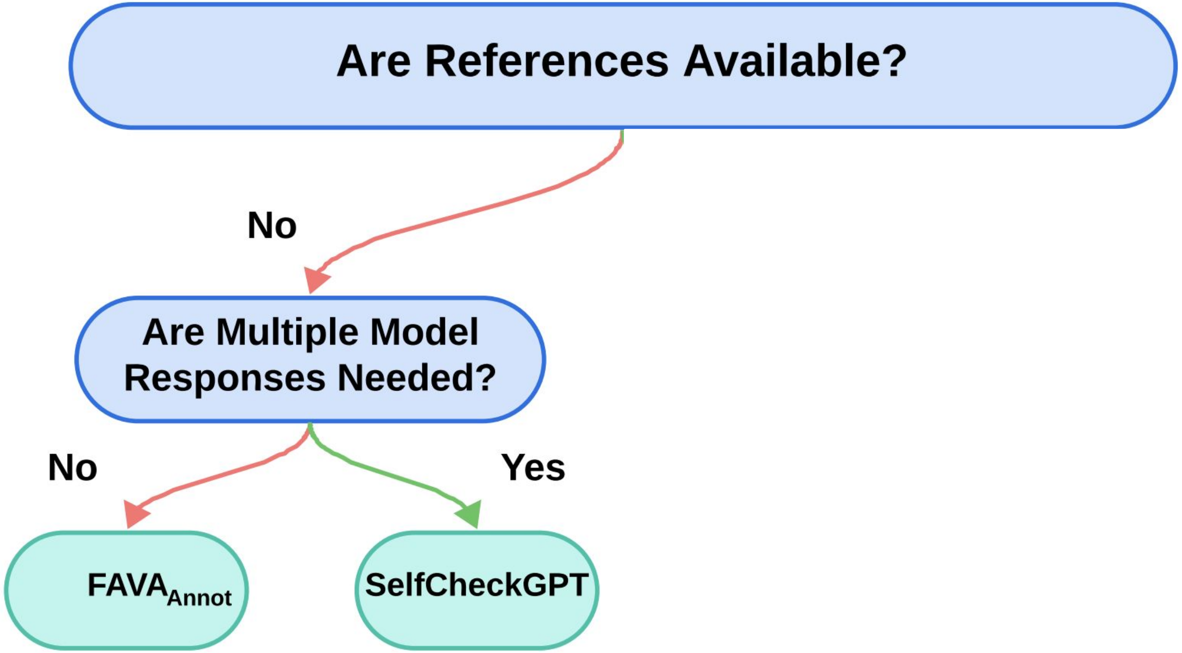
Multi-Response Consistency-Based Detection Methods



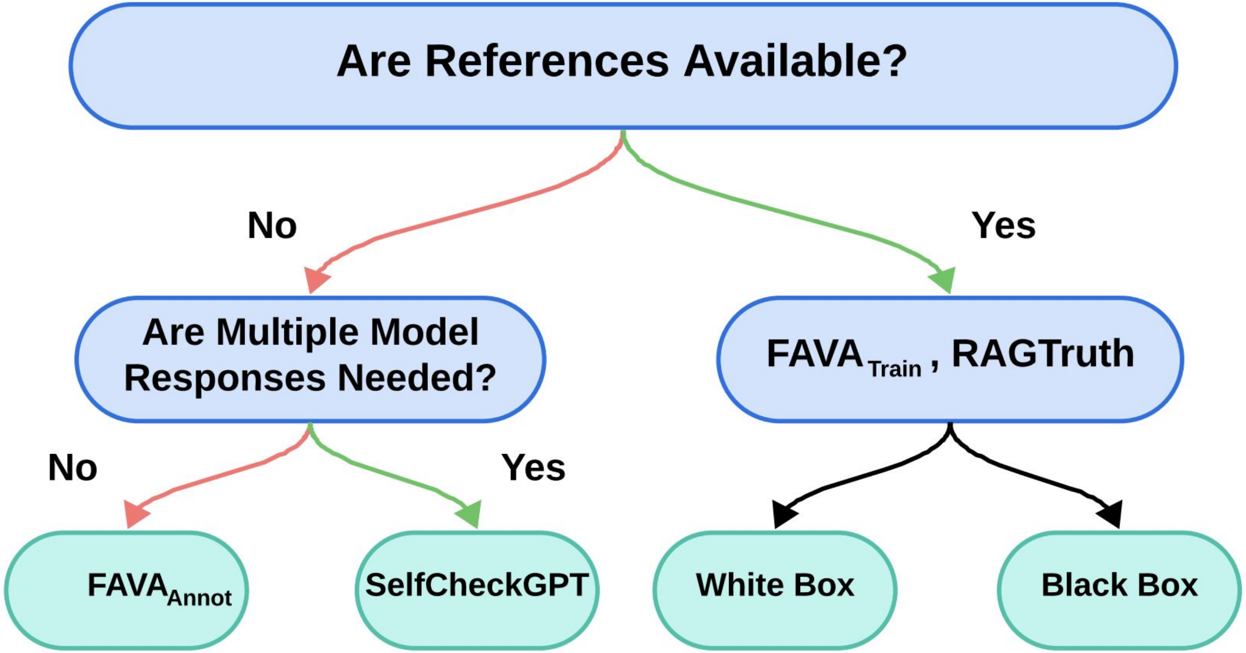
Population-level detection

with INSIDE

Classification of Hallucination Detection Settings



Classification of Hallucination Detection Settings



Detection of Hallucinations in LLMs

- Multiple LLM responses - inference time overheads and expensive
- Retraining a model - train-time overhead and generalization issues

Detection of Hallucinations in LLMs

- Multiple LLM responses - inference time overheads and expensive
- Retraining a model - train-time overhead and generalization issues

- Broad-ranging settings: with/without references, whitebox vs blackbox
- Population vs Single-response analysis
- Without finetuning/retraining or considerable inference time overheads

Detection of Hallucinations in LLMs

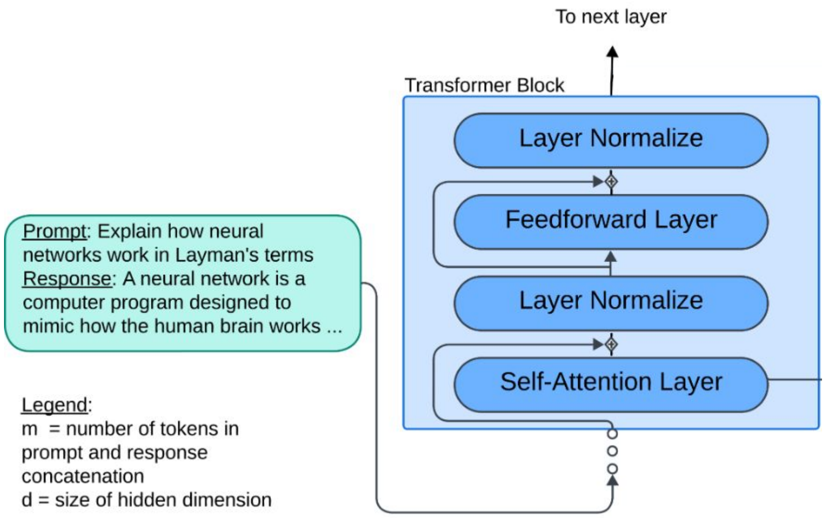
- Multiple LLM responses - inference time overheads and expensive
- Retraining a model - train-time overhead and generalization issues

- Broad-ranging settings: with/without references, whitebox vs blackbox
- Population vs Single-response analysis
- Without finetuning/retraining or considerable inference time overheads

- Can we leverage the rich semantic representations in LLMs?
- Analyze all model-related latent and output observables available with a single forward-pass of an LLM using teacher-forcing

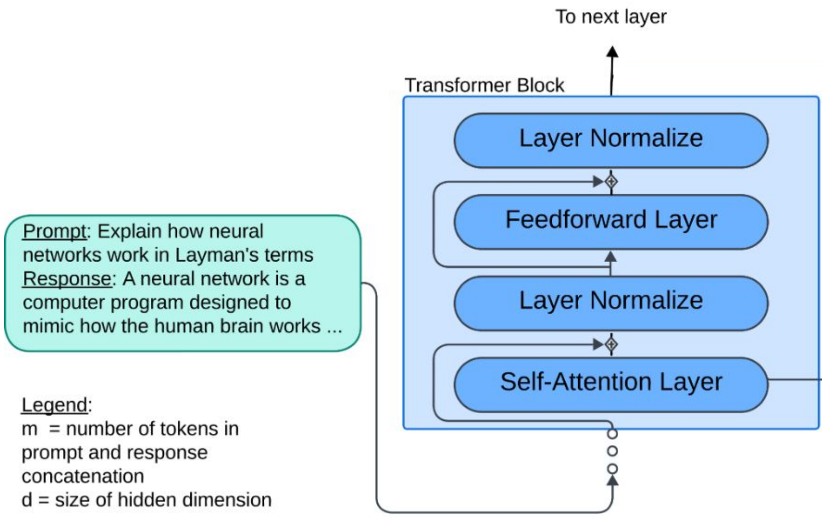
LLM-Check: Detection of Hallucinations in LLMs

- Eigenvalue Analysis of Internal LLM Representations
- Output Token Uncertainty Quantification



LLM-Check: Analyzing model internals

- Prompt $\mathbf{x_p} = (x_1 x_2 \dots x_n)$
- Response $\mathbf{x} = (x_{n+1} \dots x_m)$



LLM-Check: Analyzing model internals

- Prompt $\mathbf{x_p} = (x_1 x_2 \dots x_n)$
- Response $\mathbf{x} = (x_{n+1} \dots x_m)$

- Attention Kernels $\mathbf{A} = Ker(Q, K)V = Softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right)V$

- Hidden Activations $\mathbf{H}_{l+1} = \mathbf{H}_l + \mathbf{A}_{l+1} + \mathbf{M}_{l+1}$

LLM-Check: Analyzing model internals

Based on References [1, 2, 3, 4, 5], write a blog article:

Hallucinated Sample (HS): "The Song of Big Al" is a special episode of the nature documentary series "Walking with Dinosaurs" that focuses on the life story of an Tyrannosaurus specimen called "Big Al". The story is based on a well-preserved fossil of Big Al, which lived during the Early Jurassic period approximately 145 million years ago. The episode was produced by the BBC Natural History Unit and partnered with the National Geographic Channel, ProSieben, and TV Asahi. Rumor has it that the episode was partially shot in Cresswell Craggs, UK. Additionally, a behind-the-scenes episode called "Big Al Uncovered" was aired alongside "The Song of Big Al"

Truthful Sample (TS): "The Ballad of Big Al" is a special episode of the nature documentary series "Walking with Dinosaurs" that focuses on the life story of an Allosaurus specimen called "Big Al". The story is based on a well-preserved fossil of Big Al, which lived during the Late Jurassic period approximately 145 million years ago. The episode was produced by the BBC Natural History Unit and partnered with the Discovery Channel, ProSieben, and TV Asahi. Rumor has it that the episode was partially shot in Cresswell Craggs, UK. Additionally, a behind-the-scenes episode called "Big Al Uncovered" was aired alongside "The Ballad of Big Al"

LLM-Check: Detection of Hallucinations in LLMs

HS Token	The	Song	of	Big	Al	
$\log Ker^{jj}$	-4.99	-4.98	-5.56	-5.88	-5.69	μ = -5.42

TS Token	The	Ball	ad	of	Big	Al	
$\log Ker^{jj}$	-4.99	-5.68	-5.57	-6.72	-6.22	-5.92	μ = -5.85

HS Token	The	National	Geographic	Channel	
$\log Ker^{jj}$	-7.40	-5.61	-4.46	-5.84	μ = -5.83

TS Token	The	Disc	overy	Channel	
$\log Ker^{jj}$	-7.45	-6.57	-5.88	-6.70	μ = -6.60

HS Token	Ty	ran	n	osa	urus	spec	imen	called	"	Big	Al	".	The	
$\log Ker^{jj}$	-5.41	-5.52	-6.60	-5.27	-5.04	-5.63	-5.14	-6.02	-5.85	-6.29	-5.44	-4.81	-6.00	μ = -5.62

TS Token	All	osa	urus	spec	imen	called	"	Big	Al	".	The	story	
$\log Ker^{jj}$	-5.51	-5.35	-5.38	-6.17	-6.06	-6.45	-6.31	-6.34	-5.86	-5.92	-6.32	-5.05	μ = -5.91

LLM-Check: Hidden Score

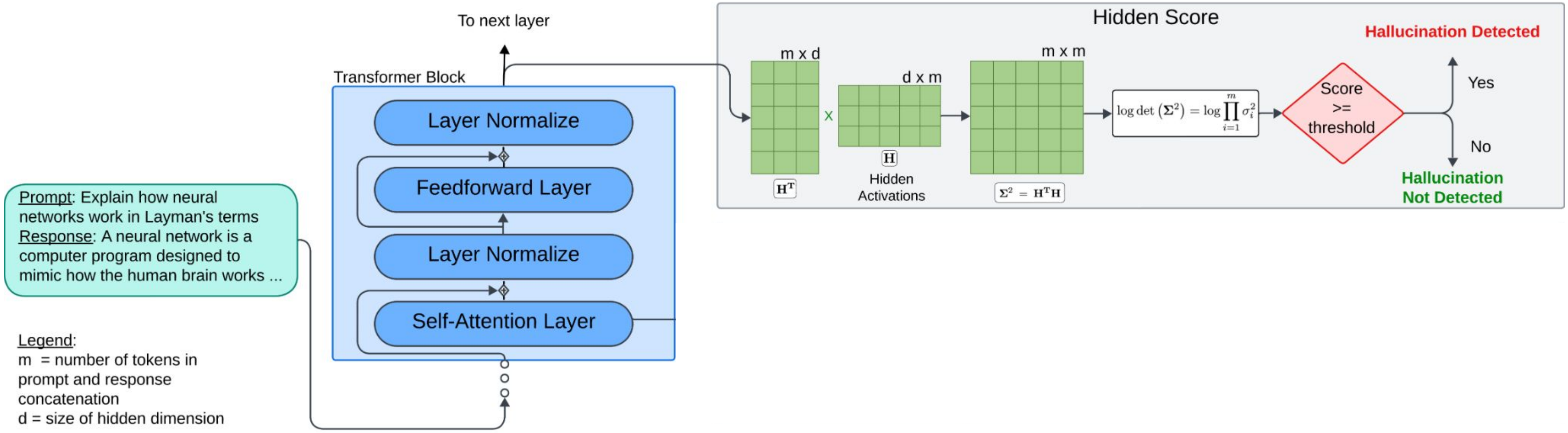
- Distinct changes in model internals within a given hallucinated response
- Quantify this saliency within representations using eigen-analysis

LLM-Check: Hidden Score

- Distinct changes in model internals within a given hallucinated response
- Quantify this saliency within representations using eigen-analysis
- For seq. of m tokens, hidden representations matrix of shape (d × m)
- Compute the mean log-det of its (m x m) covariance matrix:

$$\mathbf{\Sigma}^2 = \mathbf{H}^\mathbf{T}\mathbf{H} \quad , \quad \log\det\left(\mathbf{\Sigma}^2\right) = \log\prod_{i=1}^{m}\sigma_i^2 = \sum_{i=1}^{m} log\ \sigma_i^2 = 2\sum_{i=1}^{m} log\ \sigma_i$$

LLM-Check: Hidden Score

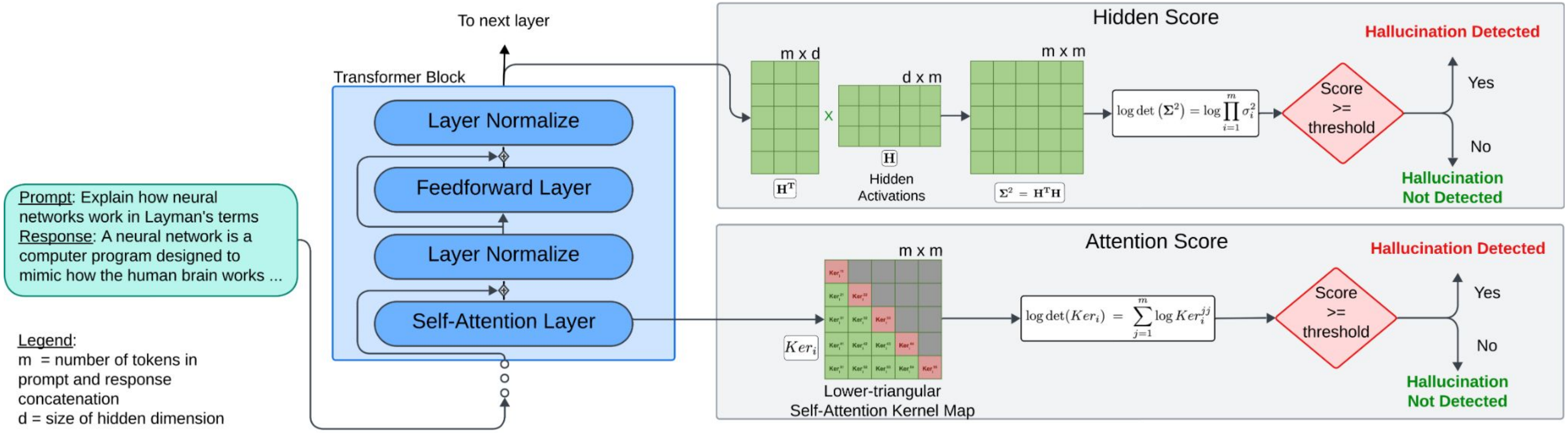


LLM-Check: Attention Score

- Sensitivity to hallucinations acutely reflected in attention mechanism
- Attention kernels are tensors of the shape (a × m × m)
- For each attention head, Ker_i is lower-triangular square matrix of size (m × m)
- Capture distribution shift using Log-determinant, which easily reduces as:

$$\log \det(Ker_i) = \sum_{j=1}^{m} \log Ker_i^{jj}$$

LLM-Check: Detection of Hallucinations in LLMs



LLM-Check: Output Uncertainty Quantification

- Perplexity $\mathrm{PPL}(\mathbf{x}) = \exp\left(-\frac{1}{m-n+1}\sum_{i=n}^{m}\log p_f\left(x_i|\mathbf{x_p}\oplus\mathbf{x}_{<i}\right)\right)$

- Logit Entropy

$$\mathrm{LogitEnt}(\mathbf{x}, k) = -\frac{1}{m-n+1}\sum_{i=n}^{m}\sum_{j=1}^{k} p_f(x_i^j|\mathbf{x_p}\oplus\mathbf{x}_{<i})\log p_f(x_i^j|\mathbf{x_p}\oplus\mathbf{x}_{<i})$$

- Windowed Logit Entropy score

LLM-Check: Comparing with Prior Works

Method	Train Indep.	Single Response	Efficient	Sample Specific	Retrieval Indep.
FAVA	×	√	√	√	√
SelfCheckGPT	√	×	×	√	√
INSIDE	√	×	√	×	√
RAGTruth	×	√	×	√	×
LLM-Check (ours)	√	√	√	√	√

Results on FAVA-Annot (Single Response, No References)

Model	Measure	AUROC	Accuracy	TPR @ 5% FPR	F1 Score
Llama-2-7B	Self-Prompt	50.30	50.30	-	66.53
Llama-2-7B	FAVA Model	53.29	53.29	-	43.88
Llama-2-7B	SelfCheckGPT-Prompt	50.08	54.19	-	67.24
Llama-2-7B	INSIDE	59.03	57.98	13.17	39.66
LLM-Check (Ours)					
	PPL Score	53.22	58.68	3.59	68.33
	Window Entropy	56.90	56.59	2.99	42.52
	Logit Entropy	53.80	55.99	2.99	56.73
Llama-2-7B	Hidden Score (LY 20)	58.44	58.08	11.98	59.66
	Attn Score (LY 21)	**72.34**	**67.96**	**14.97**	**69.27**

Results on FAVA-Annot (Single Response, No References)

Model	Measure	AUROC	Accuracy	TPR @ 5% FPR	F1 Score
Llama-2-7B	Self-Prompt	50.30	50.30	-	66.53
Llama-2-7B	FAVA Model	53.29	53.29	-	43.88
Llama-2-7B	SelfCheckGPT-Prompt	50.08	54.19	-	67.24
Llama-2-7B	INSIDE	59.03	57.98	13.17	39.66
LLM-Check (Ours)					
Llama-2-7B	PPL Score	53.22	58.68	3.59	68.33
	Window Entropy	56.90	56.59	2.99	42.52
	Logit Entropy	53.80	55.99	2.99	56.73
	Hidden Score (LY 20)	58.44	58.08	11.98	59.66
	Attn Score (LY 21)	**72.34**	**67.96**	**14.97**	**69.27**
Vicuna-7B	PPL Score	53.96	56.89	3.59	64.20
	Window Entropy	55.24	58.38	5.99	66.02
	Logit Entropy	52.29	55.69	1.80	57.31
	Hidden Score (LY 15)	58.22	59.28	10.18	**66.99**
	Attn Score (LY 19)	**71.69**	**66.47**	**24.55**	62.00
Llama-3-8B	PPL Score	53.22	58.68	3.59	67.40
	Window Entropy	56.90	56.59	2.99	55.52
	Logit Entropy	53.80	55.99	2.99	56.27
	Hidden Score (LY 15)	57.10	57.78	10.78	65.38
	Attn Score (LY 23)	**68.19**	**65.87**	**15.57**	**70.53**

Results on SelfCheckGPT Dataset (Multi-responses, No Refs.)

Model	Method	AUC-PR	Accuracy	TPR @ 5% FPR
Llama-2	SelfCheck	72.84	51.44	4.81
Llama-3	SelfCheck	75.06	54.84	5.10
		LLM-Check (Ours)		
Llama-2	Attn Score	**80.04**	58.91	9.41
Llama-2	Prompt	79.46	**61.21**	8.76
Llama-3	Attn Score	79.96	58.92	**9.48**
Llama-3	Prompt	78.49	58.54	7.11

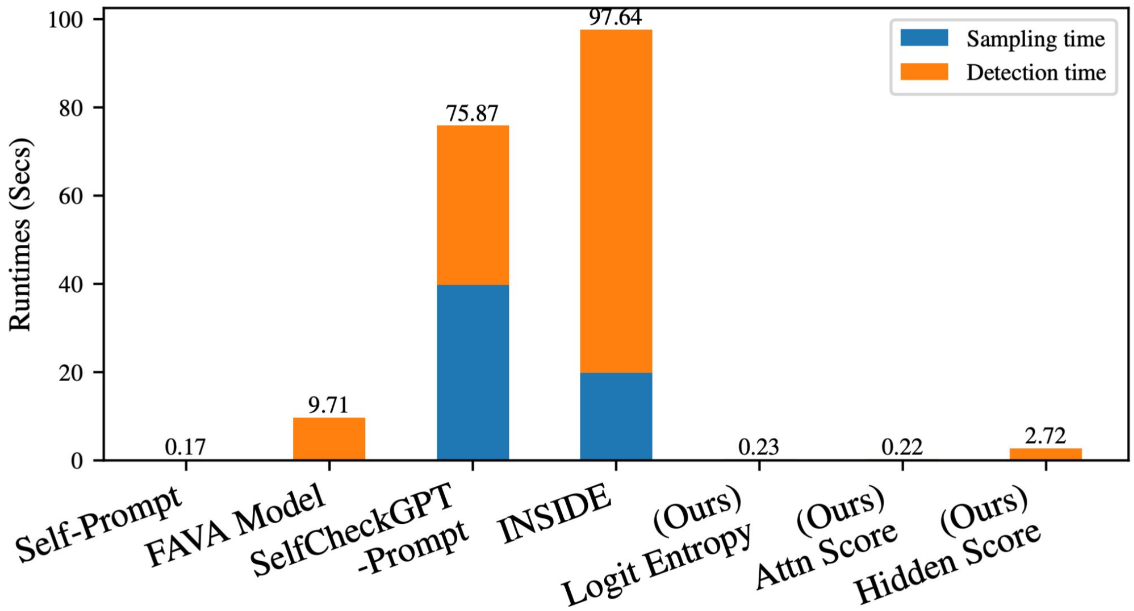
Results on Synthetic Hallucinations on FAVA Train Split

Model	Measure	AUROC	Accuracy	TPR @ 5% FPR
Llama-2	PPL Score	74.20	70.00	26.00
	Window Entropy	**77.00**	**72.00**	**34.00**
	Logit Entropy	74.36	71.00	26.00
	Hidden Score	51.44	54.00	4.00
	Attn Score	69.57	66.60	11.60
Llama-3	PPL Score	73.48	68.80	13.20
	Window Entropy	78.44	72.00	**28.00**
	Logit Entropy	**79.24**	**73.60**	28.00
	Attn Score	71.91	68.20	19.60

Results on RAGTruth (with External References)

Target Model	Measure	White-box	Black-box				
		Llama-2-7b	Llama-2-13b	Llama-2-70b	GPT-4	Mistral-7b	Overall
Hidden Score	AUROC	54.11	59.67	59.31	61.87	53.68	57.24
	Accuracy	56.33	59.66	58.42	68.52	54.15	57.62
	TPR@5%FPR	8.14	12.41	9.9	3.7	5.18	8.37
	F1 Score	61.51	50.42	66.14	67.86	32.58	47.45
Logit (Perplexity)	AUROC	53.73	52.46	56.97	52.13	52.11	53.27
	Accuracy	54.07	55.17	57.92	59.26	54.66	55.79
	TPR@5%FPR	7.69	8.97	6.93	0.00	4.15	6.01
	F1 Score	58.7	50.57	61.26	61.02	43.23	50.45
Logit (Win Entropy)	AUROC	52.08	55.71	56.38	55.83	52.61	54.58
	Accuracy	53.17	56.9	57.43	59.26	53.89	55.90
	TPR@5%FPR	4.98	15.86	1.98	7.41	10.36	10.08
	F1 Score	53.98	33.68	62.01	54.9	49.29	47.51
Logit (Log Entropy)	AUROC	53.95	51.18	55.14	50.34	50.43	51.68
	Accuracy	55.43	53.79	57.43	57.41	53.89	54.83
	TPR@5%FPR	7.24	9.66	4.95	0.00	6.22	6.65
	F1 Score	53.74	15.09	66.41	60	48.41	42.62
Attention Score	AUROC	54.19	60.05	60.01	63.51	55.37	58.30
	Accuracy	54.52	59.66	60.89	66.67	56.99	59.23
	TPR@5%FPR	5.88	14.48	12.87	7.41	5.18	9.87
	F1 Score	54.5	55.97	55.06	67.8	57.72	57.18

LLM-Check: Compute Efficiency



Upto 45x and

450x Speedup!

Summary

- LLM-Check - suite of simple, effective detection techniques over current LLMs
- Analyses hidden representations, attention kernel maps and logit outputs
- Considerable improvements over prior methods over diverse detection settings
- Applicable with/without RAG, single/multiple responses, white/black box settings
- Extremely compute-efficient: upto 45x and 450x speedup

Thank You!

Poster Session 1: 11am — 2pm PST, Wed Dec 11th

UNIVERSITY OF MARYLAND 1856

NEURAL INFORMATION PROCESSING SYSTEMS