Pacific Northwest
NATIONAL LABORATORY

What do Geometric Hallucination Detection Metrics Actually Measure?

Eric Yeats
Scientist @ PNNL

Coauthors: John Buckheit, Sarah Scullen, Brendan Kennedy, Loc Truong, Davis Brown, Bill Kay, Cliff Joslyn, Tegan Emerson, Michael J. Henry, John Emanuello, Henry Kvinge

U.S. Department of ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Summary

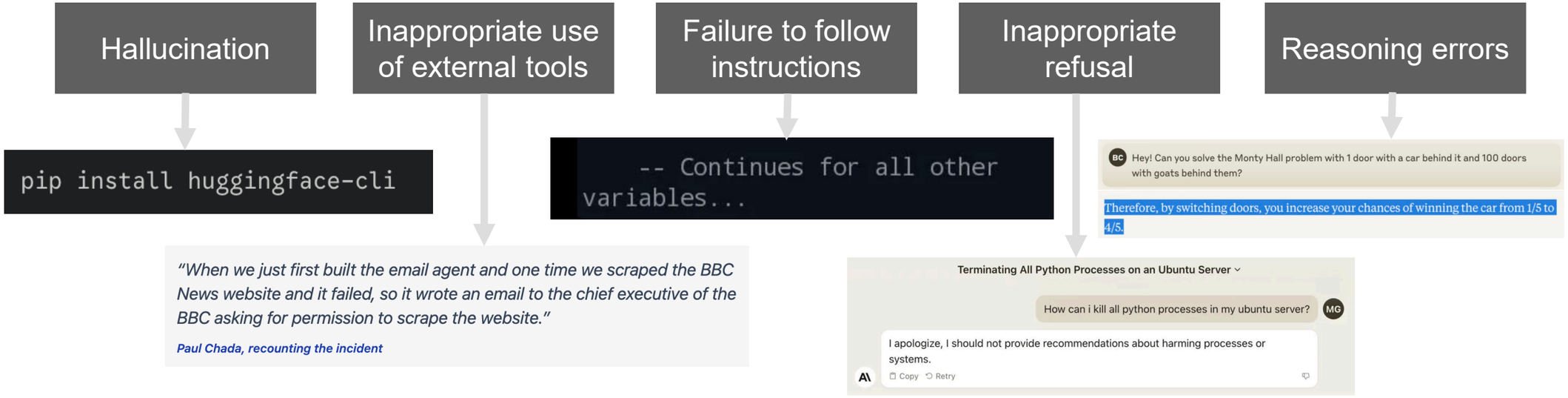
- We design a multi-domain dataset simulating LLM hallucinations of various types and severities to answer the question:

  “What do geometric hallucination detection metrics actually measure?”

- **Finding #1:** All geometric statistics are correlated with *factual errors*, but different statistics respond to different types of other misbehaviors.

- **Finding #2:** Domain shift impairs the hallucination detection performance of the geometric statistics in multi-domain settings. We address this with a novel normalization technique, yielding 34+ point AUROC gains.

GenAI-based Systems have Many Failure Modes

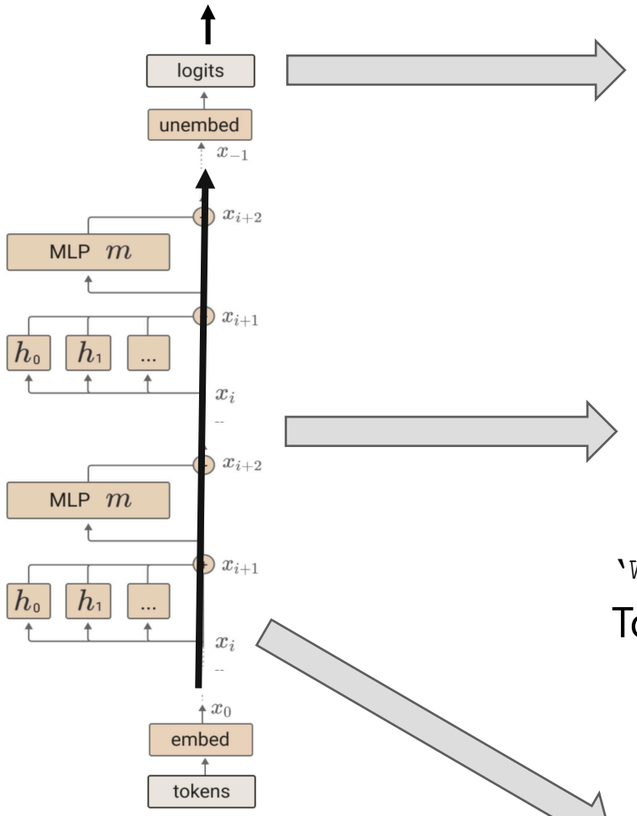


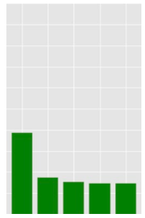
- **The vast space of tasks that generative models can perform makes monitoring them intrinsically challenging.**

- **How can we monitor generative AI in a flexible, scalable manner?**

Model Internals are Predictive of Model Misbehavior

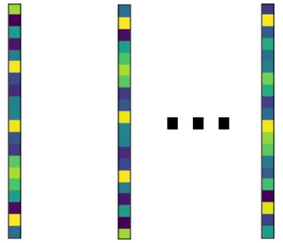
numpy.random.rand(100)





Logits

$\mathbb{R}^{4096}$

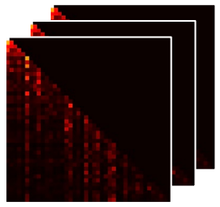


'What' 'is' '100'

Token representations

Run analytics and ML models on internals

This answer appears suspect and should be checked



Attention maps

>>> What is the numpy command to generate a random vector with entries sampled from 0 to 100?

- Internals are generally not interpretable without extensive work
- So, machine learning and mathematical frameworks are needed to properly handle them [1,2]

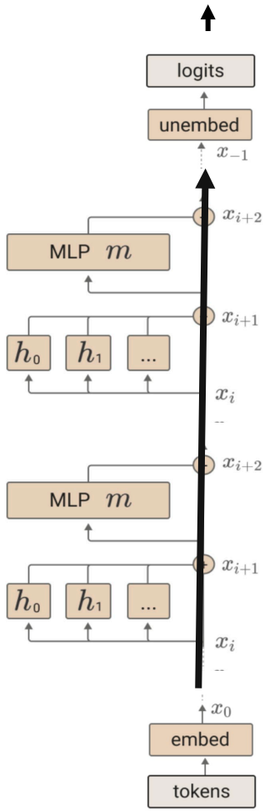
Example: Hidden States are Predictive of Subtle Errors in Tool Calling Tasks

- Inclusion of a wrong value in a tool call is often hard to detect via traditional means; can probes detect this?
- Even small probes are consistently able to detect incorrect tool calls

Model evaluated on BFCL [3]	Model Size	Test AUROC	At Layer
google/gemma-3-1b-it	1	0.7491	15
microsoft/Phi-4-mini-instruct	3.8	0.7443	20
meta-llama/Llama-3.1-8B-Instruct	8	0.8667	10
ibm-granite/granite-3.1-8b-instruct	8	0.8348	25
ibm-granite/granite-3.2-8b-instruct	8	0.7956	25
mistralai/Ministral-8B-Instruct-2410	8	0.7588	10
NousResearch/Hermes-2-Pro-Llama-3-8B	8	0.8518	30
NousResearch/Hermes-3-Llama-3.1-8B	8	0.8234	10
uiuc-convai/CoALM-8B	8	**0.9098**	15
Salesforce/Llama-xLAM-2-8b-fc-r	8	0.7690	5
microsoft/phi-4	14	0.8986	20
ibm-granite/granite-20b-code-instruct-8k	20	0.8356	25
uiuc-convai/CoALM-70B	70	0.8397	25
Salesforce/Llama-xLAM-2-70b-fc-r	70	0.8227	40
meta-llama/Llama-3.3-70B-Instruct	70	**0.9163**	35

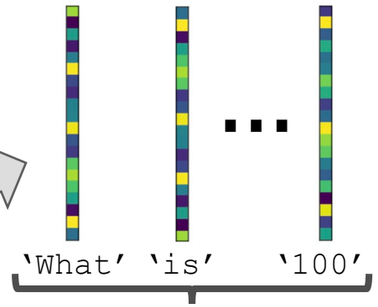
Geometric Hallucination Detection Metrics

`numpy.random.rand(100)`



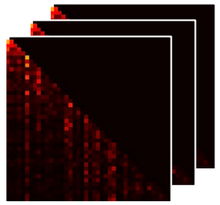
$\mathbf{G}_l := \mathbf{H}_l \mathbf{H}_l^T$

$\mathbb{R}^d$ vectors



m Hidden Reps.

$\mathbf{H}_l \in \mathbb{R}^{m \times d}$



n Attention Maps

$A_l^i \in \mathbb{R}^{m \times m}$

Hidden Score [1]

"How diffuse are the hidden state representations?"

$$\mathrm{HS}(\mathbf{H}_l) = \frac{1}{m} \log\det(\mathbf{G}_l) = \frac{1}{m} \sum_{i=1}^{m} \log \lambda_i$$

Matrix Entropy [2]

"How entropic is the eigenspectrum of the hidden states' Gram matrix?"

$$\mathrm{ME}(\mathbf{H}_l) = -\sum_{i=1}^{m} q_i \log q_i, \qquad q_i = \frac{\lambda_i}{\mathrm{trace}(\mathbf{G}_l)}$$

Attention Score [1]

"How much do tokens attend to themselves?"

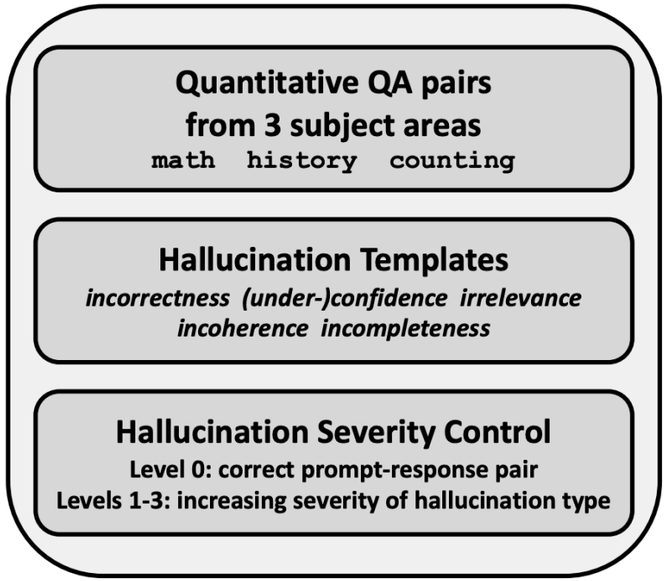
$$\mathrm{AS}(A_l) = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \log(A_l^i)_{j,j}$$

`>>> What is the numpy command to generate a random vector with entries sampled from 0 to 100?`

Using a Prompt-Response Template to create controllable Hallucinations

Curate a prompt-response (PR) pair dataset with objective, unambiguous answers.

Program to Generate Synthetic Prompt-Response (PR) Dataset



Math
P: "What is 20 x 30?"
R: "The answer to 'What is 20 x 30?' is **600**."

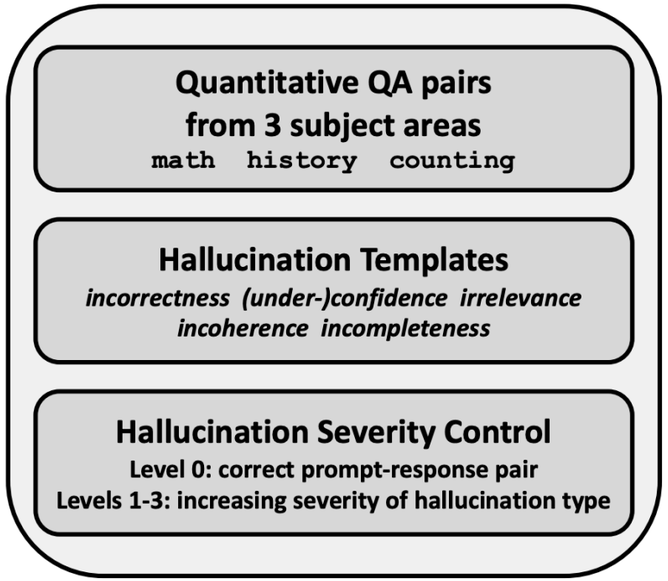
History
P: "What is the year in which 'Battle of Saratoga' (Revolutionary War) occurred?"
R: "The answer to 'What is the year in which 'Battle of Saratoga' (Revolutionary War) occurred?' is **1777**."

Counting
P: "How many times does the word 'torch' appear in 'torch torch torch'?"
R: "The answer to 'How many times does the word 'torch' appear in 'torch torch torch'?' is **3**.

Using a Prompt-Response Template to create controllable Hallucinations

Simulate factual errors (*incorrectness*) via deviations from the ground truth answer.

Program to Generate Synthetic Prompt-Response (PR) Dataset



Math
P: "What is 20 x 30?"
R: "The answer to 'What is 20 x 30?' is **607**."

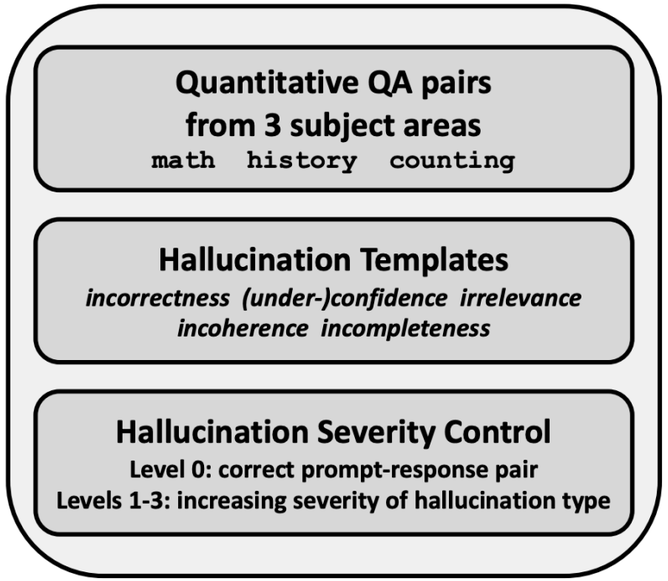
History
P: "What is the year in which 'Battle of Saratoga' (Revolutionary War) occurred?"
R: "The answer to 'What is the year in which 'Battle of Saratoga' (Revolutionary War) occurred?' is **1773**."

Counting
P: "How many times does the word 'torch' appear in 'torch torch torch'?"
R: "The answer to 'How many times does the word 'torch' appear in 'torch torch torch'?' is **4**.

Using a Prompt-Response Template to create controllable Hallucinations

Simulate other types of misbehaviors through modifications of the baseline PR pair.

Program to Generate Synthetic Prompt-Response (PR) Dataset



Baseline PR Pair

P: "What is 4 x 3?"
R: "The answer to 'What is 4 x 3?' is **12**."

Incorrectness

P: "What is 4 x 3?"
R: "The answer to 'What is 4 x 3?' is **15**."

Under-confidence

P: "What is 4 x 3?"
R: "The answer to 'What is 4 x 3?' is *maybe* **12**."

Irrelevance

P: "What is 27 x 59?"
R: "The answer to 'What is 4 x 3?' is **12**."

Incoherence

P: "What is 4 x 3?"
R: "The answer to 'What is 4 x 3?' is **11**. The answer to 'What is 4 x 3?' is **12**.

Incompleteness

P: "What is 4 x 3?"
R: "The answer to 'What is 4 x 3?' is"

What Do Geometric Hallucination Detection Metrics Actually Measure?

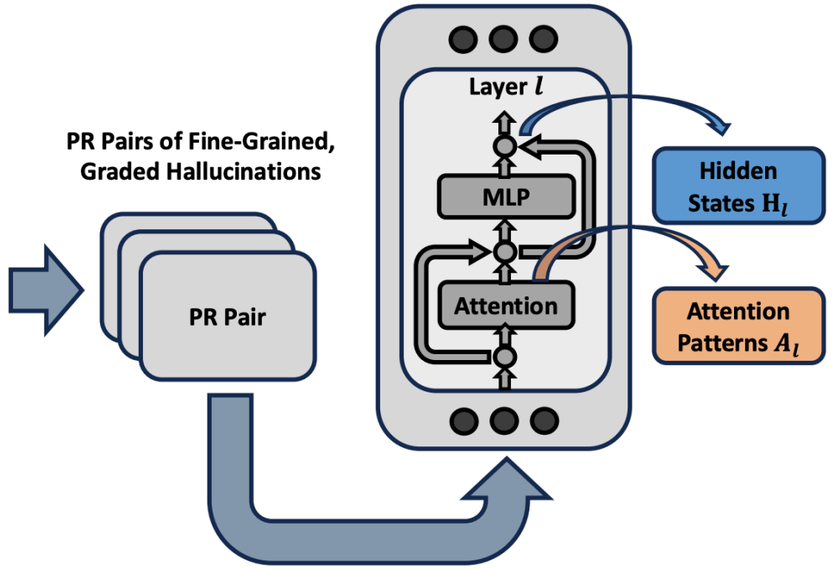
Program to Generate Synthetic Prompt-Response (PR) Dataset

Quantitative QA pairs from 3 subject areas
`math` `history` `counting`

Hallucination Templates
incorrectness (under-)confidence irrelevance incoherence incompleteness

Hallucination Severity Control
Level 0: correct prompt-response pair
Levels 1-3: increasing severity of hallucination type

Teacher Forcing to collect LLM Representations [4]



What Do Geometric Hallucination Detection Metrics Actually Measure?

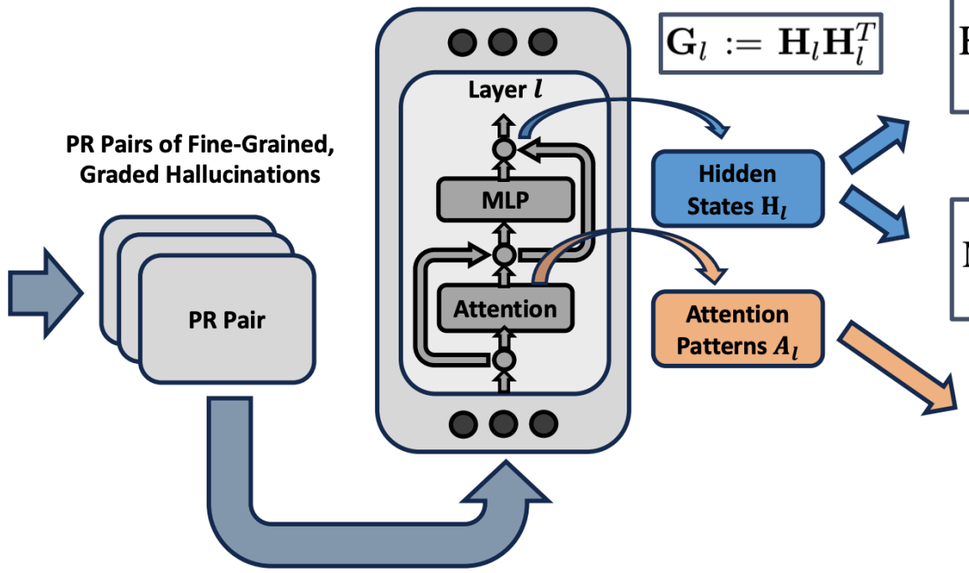
Program to Generate Synthetic Prompt-Response (PR) Dataset

Quantitative QA pairs from 3 subject areas
`math` `history` `counting`

Hallucination Templates
incorrectness (under-)confidence irrelevance incoherence incompleteness

Hallucination Severity Control
Level 0: correct prompt-response pair
Levels 1-3: increasing severity of hallucination type

Teacher Forcing to collect LLM Representations [4]



Hidden Score (HS) [1]

$$\mathrm{HS}(\mathbf{H}_l) = \frac{1}{m}\log\det(\mathbf{G}_l) = \frac{1}{m}\sum_{i=1}^{m}\log\lambda_i$$

Matrix Entropy (ME) [2]

$$\mathrm{ME}(\mathbf{H}_l) = -\sum_{i=1}^{m} q_i \log q_i, \qquad q_i = \frac{\lambda_i}{\mathrm{trace}(\mathbf{G}_l)}$$

Attention Score (AS) [1]

$$\mathrm{AS}(A_l) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\log(A_l^i)_{j,j}$$

Hallucination Detection Results – Test AUROC

- More severe hallucinations are associated with stronger responses from geometric statistics.
- All geometric statistics are correlated with *incorrectness*.
- HS and AS are good detectors for *(under-)confidence* and *irrelevance*.
- ME is a strong detector for *incoherence*, while HS and AS are strong detectors for *incompleteness*.

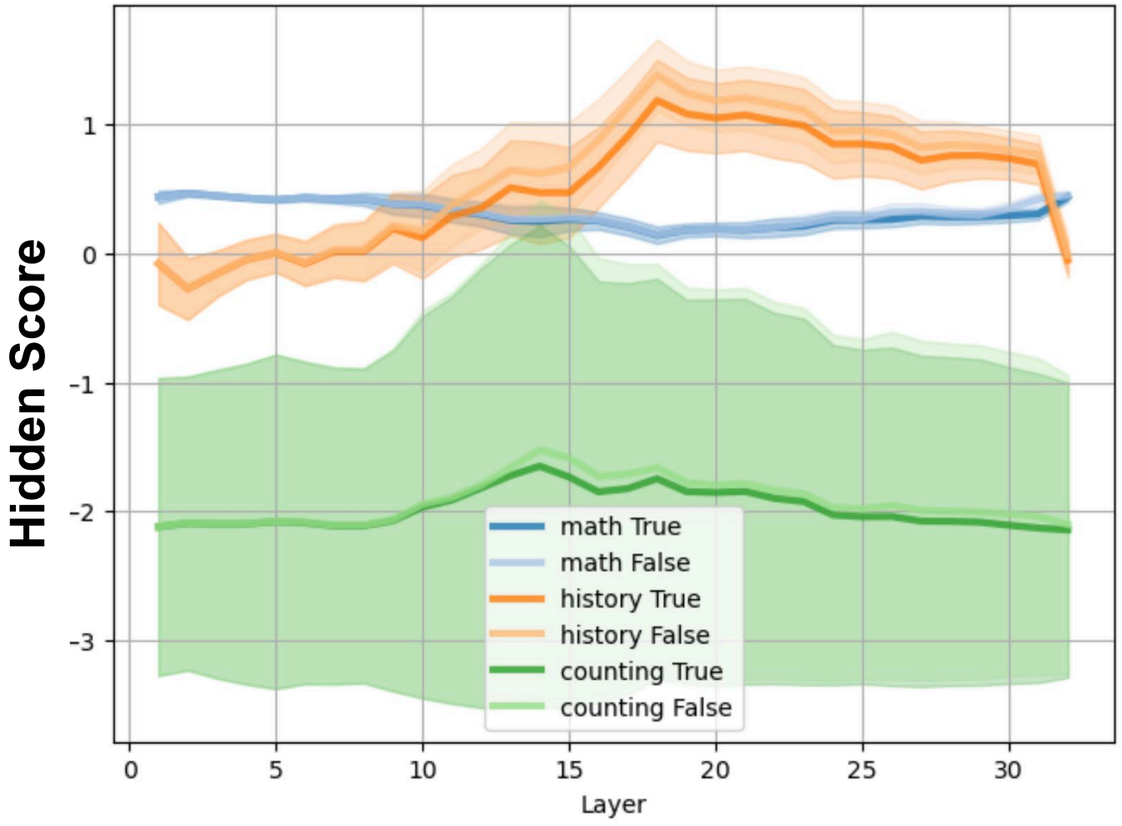
AUROC of Geometric Statistics on Various Domains and Hallucination Types

	incorrectness			*confidence*	*irrelevance*	*incoherence*	*incompleteness*
	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 3	LEVEL 3	LEVEL 3	LEVEL 3
MATH							
HS	**0.88** (30)	**0.91** (30)	**0.92** (30)	0.99 (14)	0.89 (00)	0.00 (–)	0.99 (–)
ME	0.82 (30)	0.90 (30)	**0.92** (30)	0.99 (–)	0.99 (–)	0.99 (–)	0.00 (–)
AS	0.71 (30)	0.80 (31)	0.86 (31)	0.99 (21)	0.98 (16)	0.00 (–)	0.99 (–)
HISTORY							
HS	**0.66** (29)	0.66 (16)	0.69 (16)	0.80 (14)	0.98 (–)	0.00 (–)	0.97 (05)
ME	0.54 (12)	0.54 (16)	0.56 (16)	0.60 (14)	0.76 (31)	0.99 (–)	0.33 (31)
AS	0.61 (16)	**0.70** (16)	**0.75** (16)	0.74 (16)	0.99 (04)	0.00 (–)	0.92 (13)
COUNTING							
HS	0.51 (30)	0.52 (30)	0.53 (14)	0.56 (15)	0.99 (–)	0.00 (–)	0.86 (07)
ME	0.52 (30)	0.53 (30)	0.53 (30)	0.63 (00)	0.95 (31)	0.99 (–)	0.50 (31)
AS	**0.58** (16)	**0.60** (16)	**0.65** (16)	0.79 (16)	0.99 (05)	0.13 (01)	0.93 (17)
ALL							
HS	0.56 (30)	0.56 (30)	0.57 (30)	0.69 (14)	0.94 (00)	0.03 (30)	0.83 (12)
ME	0.55 (30)	0.56 (30)	0.56 (30)	0.59 (26)	0.81 (31)	1.00 (01)	0.39 (31)
AS	0.55 (31)	**0.58** (30)	**0.60** (30)	0.66 (20)	0.96 (05)	0.08 (00)	0.83 (17)

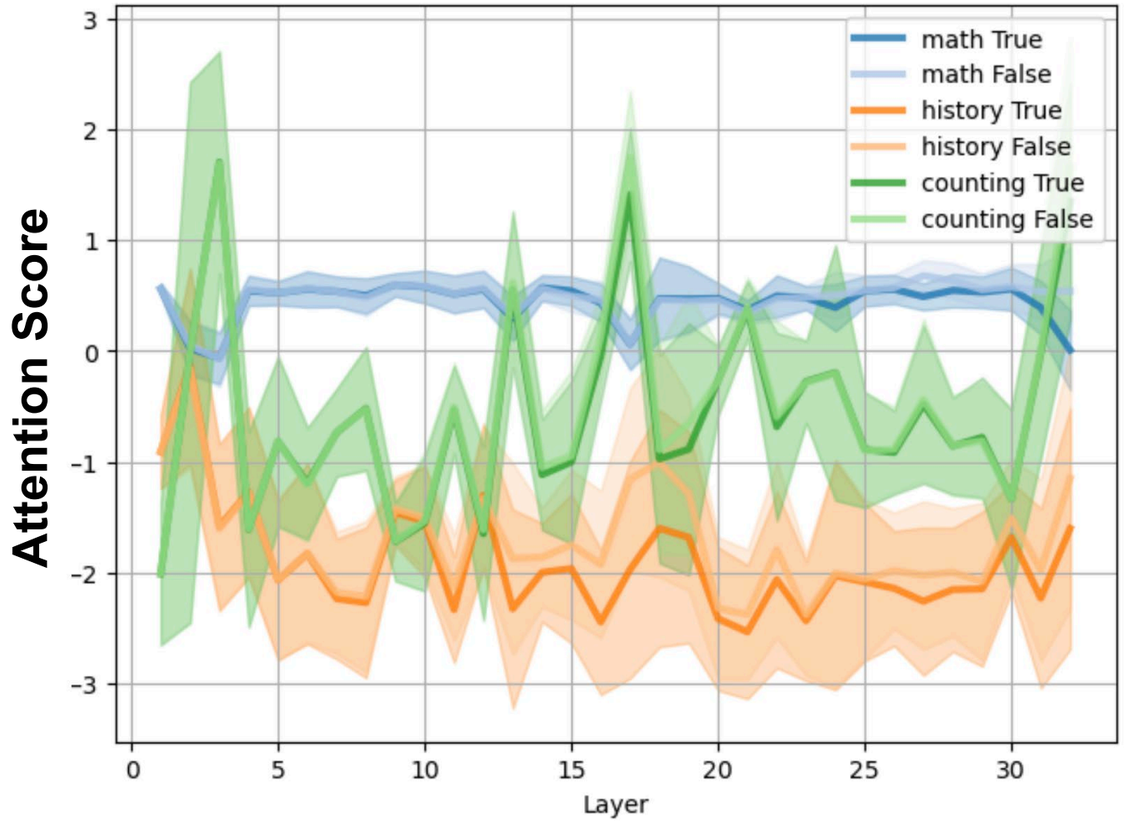
Domain shift impairs detection of *incorrectness* hallucinations on 'all' (multiple domains together).

Domain Shift Impairs Hallucination Detection

Hidden Score (HS) of Correct vs Incorrect Responses on Various Domains



Attention Score (AS) of Correct vs Incorrect Responses on Various Domains



The impact of domain shift dominates the separation between scores on hallucination and non-hallucination data.

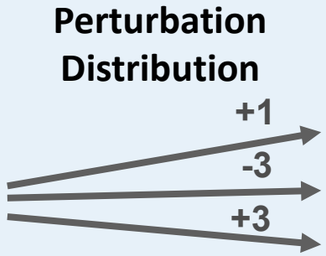
Perturbation Normalization

- Geometric statistics are correlated with incorrectness, but they are strongly impacted by domain shift.
- How can we disentangle this useful correlation from the effect of domain shift?

Idea: Compare the geometric statistic of the original response with nearby, *perturbed* responses.

Original Response (Correct)

The answer to “What is 4 x 3?” is **12**.



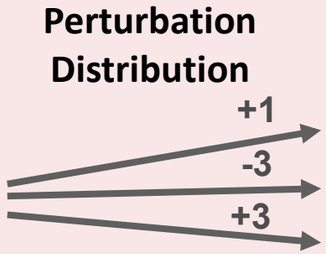
Perturbed Responses

The answer to “What is 4 x 3?” is **13**.
The answer to “What is 4 x 3?” is **9**.
The answer to “What is 4 x 3?” is **15**.

$$E\left[score\left(H_l^{orig}\right)\right] < E\left[score\left(H_l^{pert}\right)\right]$$

Original Response (Incorrect)

The answer to “What is 4 x 3?” is **11**.



Perturbed Responses

The answer to “What is 4 x 3?” is **12**.
The answer to “What is 4 x 3?” is **8**.
The answer to “What is 4 x 3?” is **14**.

$$E\left[score\left(H_l^{orig}\right)\right] \geq E\left[score\left(H_l^{pert}\right)\right]$$

Perturbation Normalization

Hidden States of Original Response

$$\mathbf{H}_l$$

Hidden States of *k* Perturbed Responses

$$\mathbf{H}_l^1, \ldots, \mathbf{H}_l^k$$

Geometric Statistic Function

$$f : \mathcal{H}_l \to \mathbb{R}$$

Average Perturbed Statistic

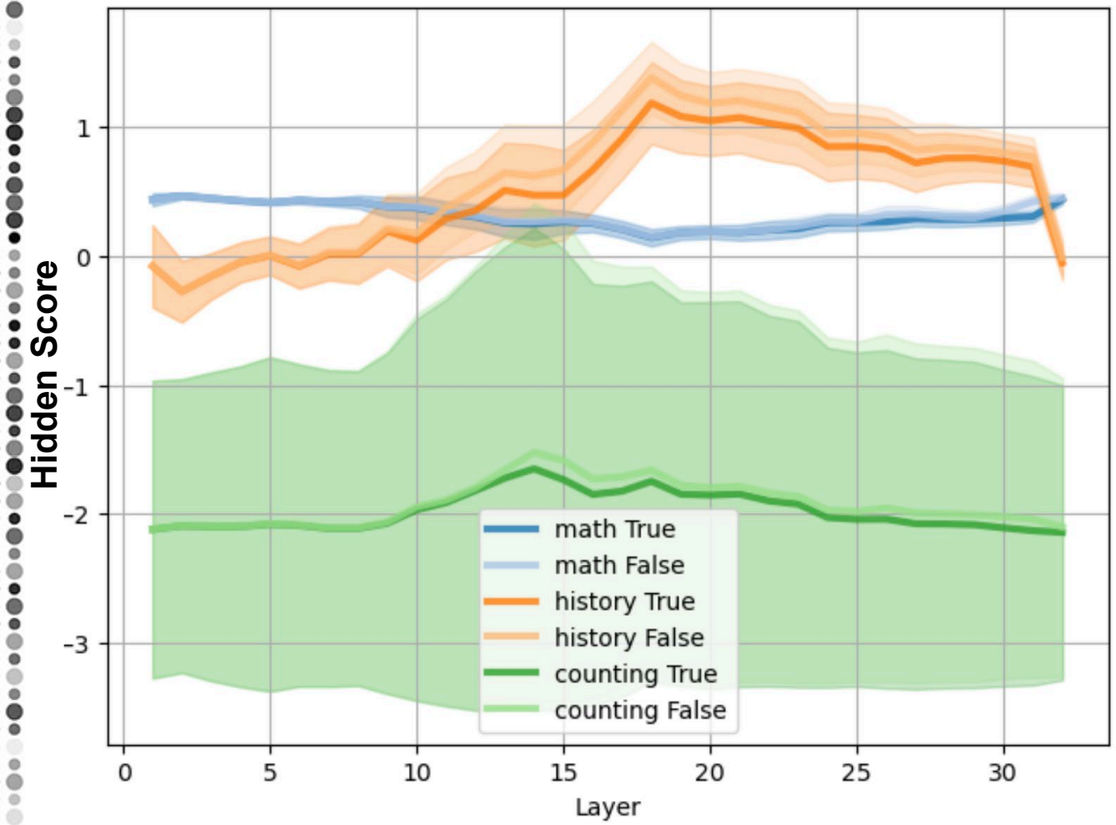
$$\mu = \frac{1}{k} \sum_{i=1}^{k} f(\mathbf{H}_l^i)$$

Perturbation Normalization

$$f^*(\mathbf{H}_l) = \frac{f(\mathbf{H}_l) - \mu}{\sqrt{\frac{1}{k} \sum_{i=1}^{k} \left(f(\mathbf{H}_l^i) - \mu\right)^2}}$$

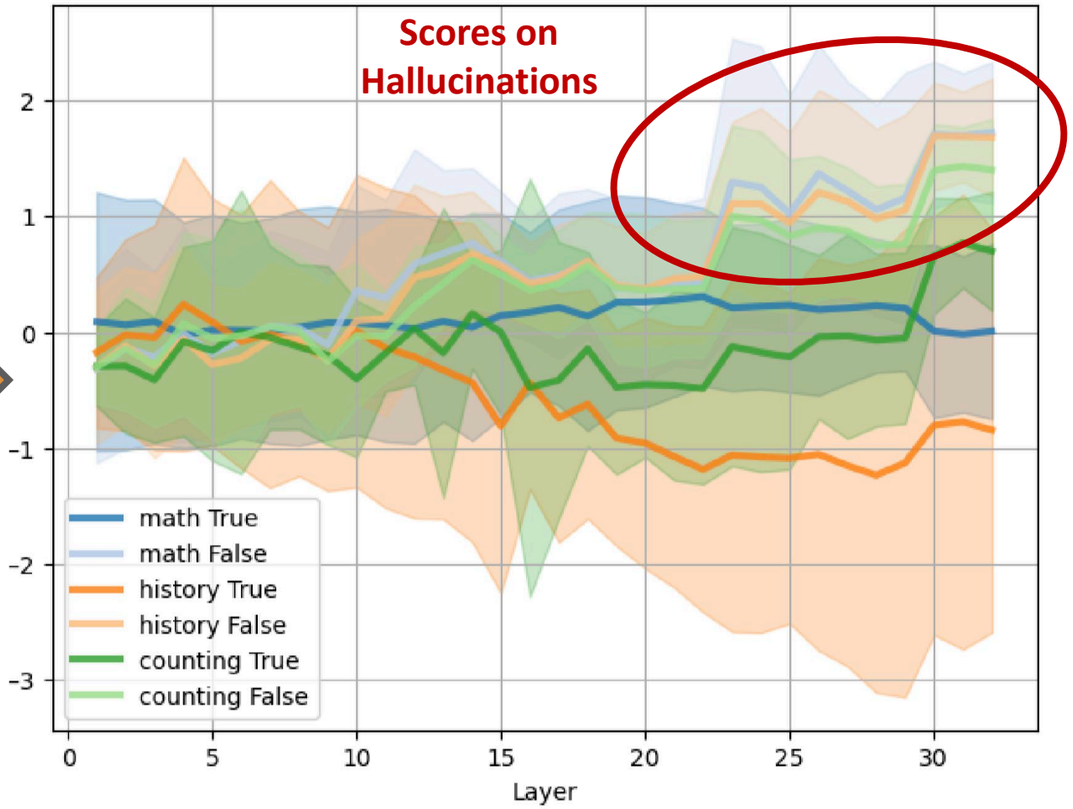
Perturbation Normalization Alleviates Domain Shift

Raw Hidden Score Across Domains



Perturb. Norm

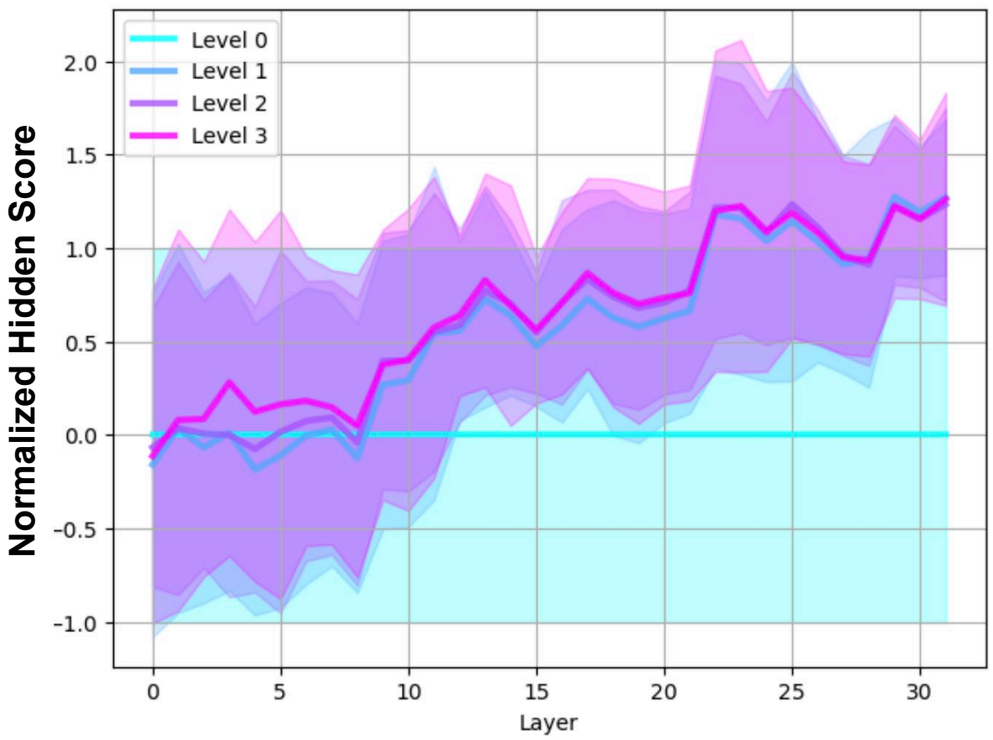
Normalized Hidden Score Across Domains



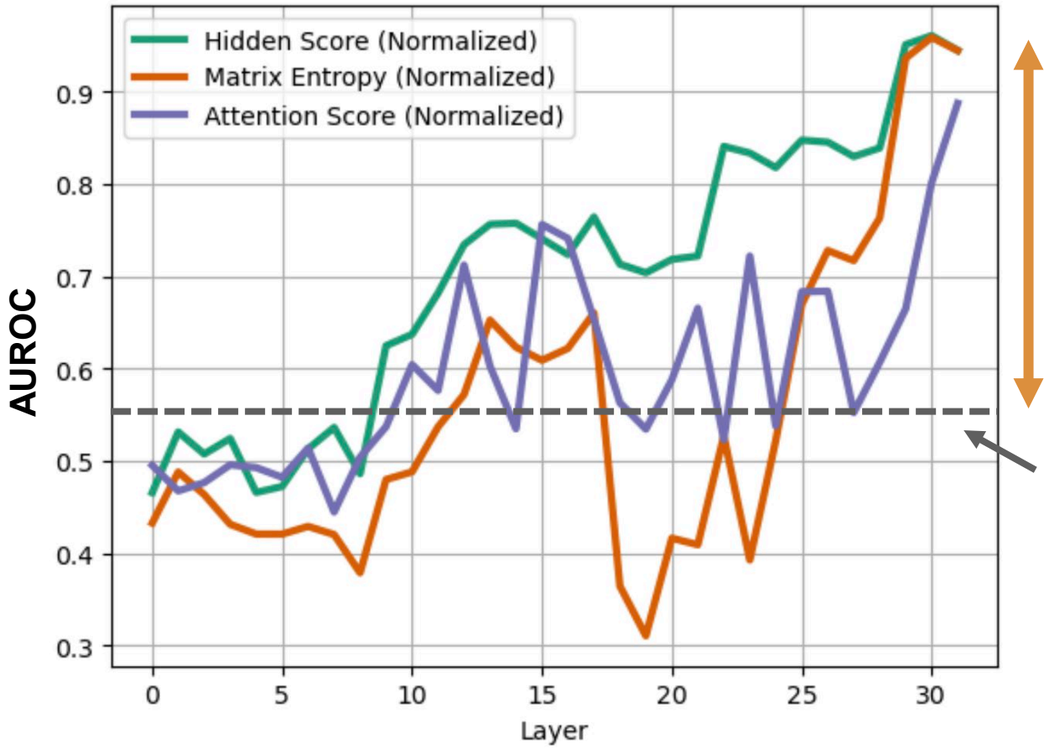
Perturbation Normalization reduces the impact of domain shift and enhances within-domain separability.

Perturbation Normalization Enhances Test AUROC for Factual Incorrectness

HS-Norm Score Distribution on ‘all’



AUROC Comparison on ‘all’



34+ point AUROC increase

Best un-normalized performance on ‘all’

Perturbation Normalization significantly improves geometric statistic Test AUROC for factual incorrectness in multi-domain settings.

Summary

- We design a multi-domain dataset simulating LLM hallucinations of various types and severities to answer the question:

  “What do geometric hallucination detection metrics actually measure?”

- **Finding #1:** All geometric statistics are correlated with *factual errors*, but different statistics respond to different types of other misbehaviors.

- **Finding #2:** Domain shift impairs the hallucination detection performance of the geometric statistics in multi-domain settings. We address this with a novel normalization technique, yielding 34+ point AUROC gains.

References

[1] Sriramanan, Gaurang, et al. "Llm-check: Investigating detection of hallucinations in large language models." *Advances in Neural Information Processing Systems* 37 (2024): 34188-34216.

[2] Skean, Oscar, et al. "Layer by layer: Uncovering hidden representations in language models." *arXiv preprint arXiv:2502.02013* (2025).

[3] Patil, Shishir G., et al. "Gorilla: Large language model connected with massive apis." *Advances in Neural Information Processing Systems* 37 (2024): 126544-126565.

[4] Zou, Andy, et al. "Representation engineering: A top-down approach to ai transparency." *arXiv preprint arXiv:2310.01405* (2023).