# What do Geometric Hallucination Detection Metrics Actually Measure?

**Eric Yeats** [1] **Davis Brown** [1 2] **John Buckheit** [1] **Sarah Scullen** [1] **Brendan Kennedy** [1] **Loc Truong** [1] **Bill Kay** [1]
**Cliff Joslyn** [1] **Tegan Emerson** [1 3 4] **Michael Henry** [1] **John Emanuello** [5] **Henry Kvinge** [1 6]

## Abstract

Hallucination remains a barrier to deploying generative models in high-consequence applications. This is especially true in cases where external ground truth is not readily available to validate model outputs. This situation has motivated the study of geometric signals in the internal state of an LLM that are predictive of hallucination and require limited external knowledge. Given that there are a range of factors that can lead model output to be called a hallucination (e.g., irrelevance vs incoherence), in this paper we ask what specific properties of a hallucination these geometric statistics actually capture. To assess this, we generate a synthetic dataset which varies distinct properties of output associated with hallucination. This includes output *correctness*, *confidence*, *relevance*, *coherence*, and *completeness*. We find that different geometric statistics capture different types of hallucinations. Along the way we show that many existing geometric detection methods have substantial sensitivity to shifts in task domain (e.g., math questions vs. history questions). Motivated by this, we introduce a simple normalization method to mitigate the effect of domain shift on geometric statistics, leading to AUROC gains of +34 points in multi-domain settings.

## 1. Introduction

Developing efficient and effective methods for detecting hallucinations is currently a major need for the larger goal of achieving reliable and responsible generative models (Huang et al., 2025). Proposed approaches come in a range of flavors, from methods that validate large langauge model's (LLM's) output using external data sources (Lewis et al., 2020; Asai et al., 2023), to methods that validate using a set of distinct judge LLMs (Jacovi et al., 2025; Zheng et al., 2023), to methods that use the consistency of a particular output as a signal for factual accuracy (Chen et al., 2024; Manakul et al., 2023). One family of methods that is particularly attractive is those that use signals extracted from model internals such as token representations in the residual stream or attention maps to detect when output is likely to be a hallucination. Though sometimes requiring labeled data upfront (Azaria & Mitchell, 2023; Orgad et al., 2024), these methods are often compute efficient to run and do not require external knowledge at inference time. Since model internals tend to be high-dimensional and not intrinsically interpretable, it has been common to leverage geometric or information-theoretic statistics in detection frameworks (Sriramanan et al., 2024; Du et al., 2024; Yin et al., 2024).

However, hallucinations can come in many forms and be characterized in several ways (Huang et al., 2025). What aspect of a hallucination a particular method is flagging remains mostly unexplored. In this paper we try to answer the question: *what characteristics of a hallucination do popular geometric detection methods actually capture?* Specifically, we programmatically generate user prompts and model responses which exhibit different properties of a hallucination within different domains. We then run these prompts and responses through the model and extract hidden states within a teacher forcing framework (Sutton, 1988). As the model has no 'awareness' that it did not actually generate these responses, we can take these hidden states as a reasonable surrogate for natural model hallucinations.

Surprisingly, we find that different geometric properties of hidden states tend to correlate to varying degrees with different flavors of hallucinations. For instance, hidden score and attention score are sensitive to irrelevant responses while matrix entropy is sensitive to incoherent responses. This suggests that existing taxonomies of hallucination may be describable in the geometry of their representations. In the process of running these experiments, we note another important factor impacting detection effectiveness: the question domain (aligning with prior work on probes in (Liu et al., 2024)). For example, when detector evaluations involve multiple domains, the statistic variance across domains is significantly larger than the detection margin, harm-

[*]Equal contribution [1]Pacific Northwest National Laboratory [2]University of Pennsylvania [3]Colorado State University [4]University of Texas, El Paso [5]Laboratory for Advanced Cybersecurity Research, National Security Agency [6]University of Washington. Correspondence to: Eric Yeats <eric.yeats@pnnl.gov>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models. 

ing performance. We introduce a method to mitigate the impact of domain shift using difference testing, leading to detector AUROC increases of $+34$ points in multi-domain settings. In summary, our work provides the following contributions: (1) We introduce a dataset to study the response of geometric statistics to hallucinated responses of various types: *(factual) incorrectness*, *(verbal) confidence*, *irrelevance*, *incoherence*, and *incompleteness*. (2) We analyze the responses of geometric statistics to hallucinations of various types and severities across three domains. In the process of doing this we identify domain shift as a key challenge for these metrics. (3) To address this we propose a simple normalization method to mitigate the effect of domain shift on geometric statistics, leading to a $+34$ point increases in AUROC when detecting hallucinations across domains.

## 2. Geometric Hallucination Detection Metrics

Hallucination detectors that utilize the internals of an LLM tend to either operate on the sequence of token representations found in the residual stream or on the attention maps extracted from multihead attention. Here we review the three such methods that we use in the experiments in Section 3.

**Hidden Score:** Let $\mathbf{H}_l$ be the $m \times d$ matrix formed by the sequence of $d$-dimensional hidden representations of $m$ tokens following layer $l$ in the LLM residual stream. Let $\mathbf{G}_l := \mathbf{H}_l\mathbf{H}_l^T$ be the corresponding Gram matrix for $\mathbf{H}_l$. (Sriramanan et al., 2024) defined the hidden score at layer $l$ to be the sequence-normalized log determinant of $\mathbf{G}_l$:

$$\text{HS}_l(x) = \frac{1}{m}\log\det\left(\mathbf{G}_l\right) = \frac{1}{m}\sum_{i=1}^{m}\log\lambda_i, \qquad (1)$$

where $\lambda_i$ is the $i$-th eigenvalue of $\mathbf{G}_l$. The hidden score can be interpreted as the log volume of the parallelpiped formed by columns of $\mathbf{H}_l$. Hallucinations are associated with larger volumes, indicating a more diffuse representation.

Though it has not been used specifically to detect hallucinations, the matrix entropy (ME) of $\mathbf{G}_l$ was used in (Skean et al., 2025) to study the information content of LLM internal representations and is a natural statistic to characterize a sequence of hidden activation vectors. Matrix entropy is parameterized by a non-negative real-number $\alpha$. We use the version obtained by letting $\alpha \to 1$ which is equivalent to Shannon entropy[1]:

$$\text{ME}_l(x) = -\sum_{i=1}^{m} q_i \log q_i, \qquad q_i = \frac{\lambda_i}{\text{trace}(\mathbf{G}_l)}. \qquad (2)$$

The matrix entropy provides another way of measuring how diffuse a sequence of token representations is.

Auto-regressive LLMs often employ causal self-attention. The corresponding attention maps are lower triangular and non-negative. Hence, the log determinant of the $i$-th attention map at layer $l$, $A_l^i$, is simply the sum of log entries along the diagonal. Introduced by (Sriramanan et al., 2024), the attention score is the sequence-normalized log determinant for each of the $n$ attention heads:

$$\text{AS}_l = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\log(A_l^i)_{j,j}. \qquad (3)$$

Informally, the attention score measures how much each token attends to itself. Experimentally, hallucinations are associated with higher attention scores.

## 3. What do Geometric Statistics Measure?

We design an experiment to study the responses of geometric statistics to various hallucination types and levels of severity.

**Dataset Design** Our dataset contains correct question and (numeric) answer (QA) pairs belonging to three domains: `math` (multiplication of positive integers), `history` (year of event, CE), and `counting` (number of times a word appears in a sequence). We use the QA pairs to generate prompt and response (PR) pairs. The PR pairs follow a template consisting of a `prompt_question`, `response_question`, `conf_mod` (confidence modifier), `answer`, and `answer_offset`. The template is:

**P:** "{`prompt_question`}" "What is $46 \times 53$?"

**R:** "The answer to '{`response_question`}' is {`conf_mod`} {`answer` + `answer_offset`}."

"The answer to 'What is $46 \times 53$?' is 2438."

Given a QA pair (a `response_question` and an `answer`), the baseline, correct PR sequence consists of a `prompt_question` which is the `response_question`, a `conf_mod` set to "", and an `answer_offset` set to 0. We craft hallucinations by modifying the template in the following ways:

*(Factual)* ***Incorrectness***: Assign a non-zero integer to `answer_offset`, causing the response to be incorrect. Larger magnitude integers are used to simulate more severe hallucination.

[1]Matrix entropy uses the $\alpha$-Renyi entropy which generalizes several information-theoretic statistics including Shannon entropy. For instance, when $\alpha = 2$, $\alpha$-Renyi entropy corresponds to von Neumann entropy.

*(Under-)* ***Confidence*****:** Insert a verbal confidence modifier in place of `conf_mod` e.g., "*probably*" or "*maybe*".

***Irrelevance*****:** Sample a `prompt_question` which is different from `response_question`. A cross-domain `prompt_question` (as opposed to intra-domain) simulates a more severe hallucination.

***Incoherence*****:** Repeat the response (**R**) multiple times while changing `answer_offset`. More repetitions with inconsistent responses simulate stronger hallucinations.

***Incompleteness*****:** Terminate the response (**R**) early with an end-of-text token. Earlier termination simulates more severe hallucination.

**Data Recording** We collect the LLM's intermediate representations from each of the question and answer pairs using "teacher forcing" (Sutton, 1988). The LLM ingests the PR pair, and the hidden states and attention maps are recorded for each token and LLM layer. We calculate and analyze HS, ME, and AS for each of these. We use Llama 3.1-8B-Instruct in our experiments (Grattafiori et al., 2024) which were run on a single Nvidia H100 GPU. In all cases, the task is detection of hallucinations (of any type and severity) among all responses, including correct ones.

### 3.1. Experimental results and analysis

**Single-Domain Performance on *Factual Incorrectness*** Table 1 depicts the area under the receiver operating characteristic (AUROC) hallucination detector scores for all statistics, domains, and hallucination types. We observe three primary trends regarding statistic responses to *factual incorrectness*. **First**, more severe hallucinations (levels 1 to 3) are associated with better detection scores for all statistics. **Second**, different statistics are better detectors for different domains. Hidden Score (HS) and Matrix Entropy (ME) (those which process the Gram matrix of a hidden states sequence) are better detector scores for `math`, with AUROCS of 0.92 on level 3 hallucinations. However, Attention Score (AS) is the best detector score for `history` and `counting`, with AUROCs of 0.75 and 0.65 on level 3 hallucinations, respectively. **Third**, we observe that the domain impacts optimal recording layer more than the choice of metric. Layer indices 30-31 are the best for detecting hallucinations on the `math` dataset. However, layer indices 14-16 yield the highest AUROC for the three statistics on `history` and `counting`.

**Sensitivity to Alternative Hallucination Types** The right-hand side of Table 1 reports AUROCs for the statistics as they respond to *(under-)confidence*, *irrelevance*, *incoherence*, and *incompleteness*. We observe that in general, the statistics are highly responsive to these types of hallucinations. For example, HS and AS achieve AUROCs of 0.94 and 0.96 for detecting *irrelevant* hallucinations on `all`. Moreover, they yield AUROCs of 0.83 for *incomplete* hallucinations on `all`. However, HS and AS do not respond to *incoherent* responses as hallucinations (they have *lower* scores than the baseline), leading to AUROCs far below random guessing for this type of hallucination. On the other hand, ME is highly receptive to *incoherent* hallucinations, achieving near perfect AUROC for this category. Each of the statistics are somewhat responsive to *under-confidence*, yielding AUROCs of 0.59-0.69 on `all`.

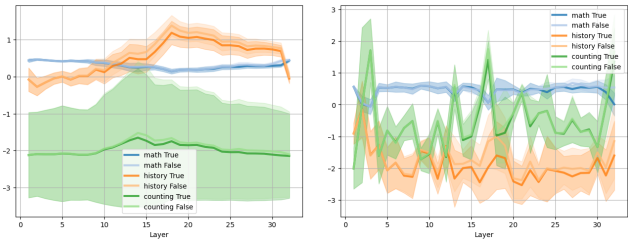



*Figure 1.* Distributions of HS (left) and AS (right) for *correct* and *incorrect* (level 3) responses for each domain. Solid lines: distribution means. Shaded areas: one standard deviation.

**Domain Shift Harms Detection of *Factual Incorrectness*** For *factual incorrectness* questions and responses both HS and AS achieve strong AUROC on the individual domains (with the exception of HS on `counting`). However, when the domains are combined in `all`, the performance of each statistic drops significantly, with HS and AS yielding AUROCs of 0.57 and 0.60 on level 3 hallucinations. Figure 1 depicts the distributions of HS and AS across domains. This poses issues in cases where one might want a detection system capable of catching hallucinations across domains (e.g., a general purpose chatbot).

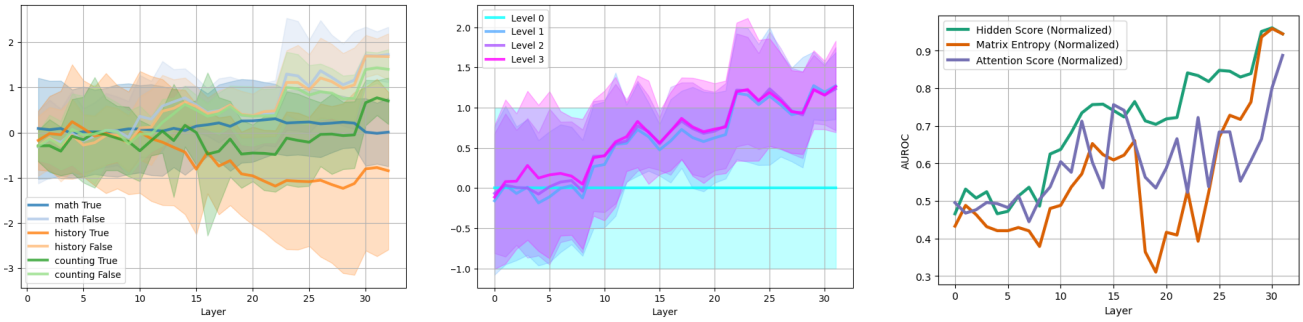
## 4. Mitigating Domain Shift

We propose a simple normalization method to mitigate domain shift, leading to improved detection performance. First, we locate the exact token(s) of the answer to be verified. This is simple in our experiments, but in real-world settings this task could be achieved by an auxiliary LLM. Second, we collect the internal representations of the original response $\mathbf{H}_l \in \mathcal{H}$ as well as $k$ versions of the same response $\mathbf{H}_l^1, ..., \mathbf{H}_l^k$ with *perturbed* answers. This is done using a perturbation set. In our setting we added $-5, -2, -1, 1, 2, 5$ to the answer. Third, for a given statistic function $f : \mathcal{H}_l \to \mathbb{R}$ and $\mu = \frac{1}{k}\sum_{i=1}^{k} f(\mathbf{H}_l^i)$, we collect the following normalization score:

$$f^*(\mathbf{H}_l) = \frac{f(\mathbf{H}_l) - \mu}{\sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(f(\mathbf{H}_l^i) - \mu\right)^2}}. \tag{4}$$

*Table 1.* Hallucination detection AUROC. Each reported AUROC is the highest score computed across all layers. The layer index $[0, 32)$ corresponding to the maximum AUROC is in parentheses $(\cdot)$. We report (–) if multiple layers have the same AUROC.

| | *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) |




(a) HS-Norm - Domain Alignment (b) HS-Norm - score distribution on `all` (c) AUROC Comparison on `all`

*Figure 2.* Our *perturbation normalization* significantly reduces domain shift, leading to boosted detection performance for *incorrectness*.

An equivalent definition holds when attention maps $A_l^1, \ldots, A_l^n$ are used. Intuitively, $f^*(\mathbf{H}_l)$ measures how much of an outlier $f(\mathbf{H}_l)$ is compared with perturbed versions of itself. One should expect a *correct* response to have lower statistics than its perturbed neighbors, while *incorrect* responses should not. We call this *perturbation normalization* of a geometric statistic.

**Results** We augment each of the statistics with $f^*(\cdot)$ and record their scores on `all`. Figure 2a depicts the distributions of HS-Norm scores for baseline responses and for level-1 *incorrectness* hallucinations (this should be compared with Figure 1). The *correct* responses for each of the domains are aligned on the bottom half of the chart at layer 30, while the *incorrect* response distributions are aligned on the top half of the chart at layer 30. Figure 2b depicts the normalized distributions of HS-Norm scores on `all` for *correct* responses and *incorrect* responses (levels 1-3). We observe that HS-Norm separates *incorrect* from *correct* well in the later stages of the network. Figure 2c depicts the AUROC for each statistic on `all` (level 1 hallucinations). The aligned domains lead to significantly improved detector performance. HS-Norm and ME-Norm achieve AUROCs of 0.96 **(40 point gain)** at layer 30, while AS-Norm achieves an AUROC of 0.89 **(34 point gain)** at layer 31.

## 5. Conclusion

We design a multi-domain dataset of hallucinations with various properties and severities to try to answer the question "What do geometric hallucination detection metrics actually measure?" We find that all geometric statistics are correlated with *incorrectness*, but different statistics respond to different hallucination characteristics. Additionally, we find that domain shift impairs the detection performance of the statistics on *incorrectness*. We mitigate domain shift

with a simple difference testing technique, leading to 34 to 40 point AUROC gains for each statistic in multi-domain hallucination detection scenarios.

## References


Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2023.

Azaria, A. and Mitchell, T. The internal state of an LLM knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., and Ye, J. Inside: LLMs' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*, 2024.

Du, X., Xiao, C., and Li, S. Haloscope: Harnessing unlabeled LLM generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37: 102948–102972, 2024.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Jacovi, A., Wang, A., Alberti, C., Tao, C., Lipovetz, J., Olszewska, K., Haas, L., Liu, M., Keating, N., Bloniarz, A., et al. The facts grounding leaderboard: Benchmarking LLMs' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*, 2025.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Liu, J., Chen, S., Cheng, Y., and He, J. On the universal truthfulness hyperplane inside LLMs. *arXiv preprint arXiv:2407.08582*, 2024.

Manakul, P., Liusie, A., and Gales, M. J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. *arXiv preprint arXiv:2410.02707*, 2024.

Skean, O., Arefin, M. R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., and Shwartz-Ziv, R. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

Sriramanan, G., Bharti, S., Sadasivan, V. S., Saha, S., Kattakinda, P., and Feizi, S. LLM-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37: 34188–34216, 2024.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988.

Yin, F., Srinivasa, J., and Chang, K.-W. Characterizing truthfulness in large language model generations with local intrinsic dimension. *arXiv preprint arXiv:2402.18048*, 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.

## A. Related Work: LLM Hallucination Detection

In this section we briefly review some of the important families of hallucination detection methods with a special emphasis on those that use the internal states of a model.

**Retrieval-Augmented Generation:** RAG approaches address hallucinations preemptively by grounding LLM outputs in external knowledge. Lewis et al. (2020) pioneered this technique by combining dense retrieval with sequence generation. Recent advances include adaptive retrieval (Asai et al., 2023), where systems dynamically determine when to retrieve information. These systems often require ground truth source data, specialized LLM training, significant infrastructure for ground truth data retrieval, and additional inference costs.

**Consistency Methods:** Consistency-based approaches detect contradictions as hallucination indicators. Self-consistency methods such as SelfCheckGPT (Manakul et al., 2023) generate multiple responses and identify inconsistencies among them. More recently, Chen et al. (2024) leverage semantic information from the internal states of LLMs to boost consistency methods.

**Logit-based Methods:** Several hallucination detection methods leverage the logits during token generation to detect hallucinations. Farquhar et al. (2024) proposed to evaluate the truthfulness of LLM generations using *semantic entropy*. That is, to deduplicate semantically equivalent generations and calculate the entropy of the grouped response. Kadavath et al. (2022) propose P(True), where LLMs assess the veracity of their own generations using the output logit for "True".

**External Validation (Judge LLM):** External validation employs separate models or systems to assess LLM outputs. Judge LLM approaches (Zheng et al., 2023) use specialized models trained to evaluate factuality of other models' outputs. Jacovi et al. (2025) employ an ensemble of powerful judge LLMs with sophisticated prompting techniques to check the veracity of LLM responses using ground-truth sources.

**Detection with Internal States:** There are several approaches that exploit LLM internal representations to detect hallucinations. Azaria & Mitchell (2023) showed that lightweight probes can be trained with labeled data to detect hallucinations. Similar to our work, (Liu et al., 2024) showed that these probes sometimes fail to generalize between domains. (Liu et al., 2024) showed that more diverse training sets are key to probes that generalize. More recently, Sriramanan et al. (2024) employ geometric statistics (Hidden Score and Attention Score) as detector scores for zero-resource hallucination detection. These statistics do not require ground truth and are extremely efficient to compute, but we show that they do not generalize well across domains. Du et al. (2024) identify hallucinations as outliers when projecting hidden states to a lower-dimensional subspace. They use this geometric indicator to assign pseudo-labels to generated responses and subsequently train a lightweight probe on the pseudo-labeled hidden states.

## B. Dataset Details

Our dataset has $550$ baseline correct QA pairs which are augmented with the PR template to generate hallucinations of various types. *(Under-)confidence* is simulated by setting `conf_mod` to "probably" (level 1), "maybe" (level 2), or "not" (level 3). *Incoherence* is simulated by repeating the response $n$ times with inconsistent, incorrect answers until the last repetition in which the answer is correct. The response is repeated 2 times for level 1, 3 times for level 2, and 4 times for level 3. *Incompleteness* is simulated by truncating the response early with an end-of-text token. The response is terminated after $90\%$ for level 1, $80\%$ for level 2, or $70\%$ for level 3.

**`math`** Our `math` domain contains $400$ integer multiplication QA pairs in the format $a \times b =$ `answer`, with $a, b \in [40, 60)$. *Incorrectness* is simulated by setting `answer_offset` to a random integer between $\pm[1, 9]$ (level 1), $\pm[10, 99]$ (level 2), or $\pm[100, 999]$ (level 3). Given a `response_question` with $a$ and $b$, *irrelevance* is simulated by sampling a new $b' \in \mathbb{Z}$ where $b' \neq b$ such that $|b' - b| < 5$ (level 1) or $5 < |b' - b| < 20$ (level 2) and inserting it into `prompt_question`. Level 3 *irrelevance* is simulated by sampling a question from another domain and using it as the `prompt_question`.

**`history`** Our `history` domain contains 70 QA pairs consisting of famous events and their year for three geographical regions. Each geographical region has at least two sets of ten questions each corresponding to specific timeframes. *Incorrectness* is simulated by adding a random integer between $\pm[1, 5]$ (level 1), $\pm[6, 20]$ (level 2), or $\pm[21, 50]$ (level 3).

*Irrelevance* is simulated by resampling a question from the same timeframe and region (level 1) or by resampling a question from a different timeframe (level 2). Level 3 *irrelevance* is simulated by sampling a question from another domain and using it as the `prompt_question`.

**`counting`** Our `counting` domain contains 80 QA pairs of word sequences and word counts (e.g., "*word word*, *word*" and 3). The sequences are generated by iterating through a set of eight words and repeating the word for $c \in [3, 12]$ times. *Incorrectness* is simulated by adding a random integer: $\pm 1$ (level 1), $\pm 2$ (level 2), or $\pm 3$ (level 3). *Irrelevance* is simulated by resampling a new sequence of the same word which has a count difference $|c - c'| \leq 3$ (level 1) or $|c - c'| > 3$ (level 2). Level 3 *irrelevance* is simulated by sampling a question from another domain and using it as the `prompt_question`.

**`all`** We include 75 QA pairs from `math`, 70 QA pairs from `history`, and 80 QA pairs from `counting` into our `all` set. This ensures that the representation of the domains is relatively balanced.

## C. Additional Results

**Results on Alternative Hallucination Types** Figure 3 depicts the distributions of responses of the geometric statistics to *(under-)confidence*, *irrelevance*, *incoherence*, and *incompleteness* for each of the 32 layers of Llama-3.1-8B-Instruct. Each row corresponds to the responses of a single statistic (e.g., HS) to a different hallucination type on all three domains (`math`, `history`, `counting`) together. In each figure, the distribution of responses at each layer are normalized by that of the correct, base PR sequence.

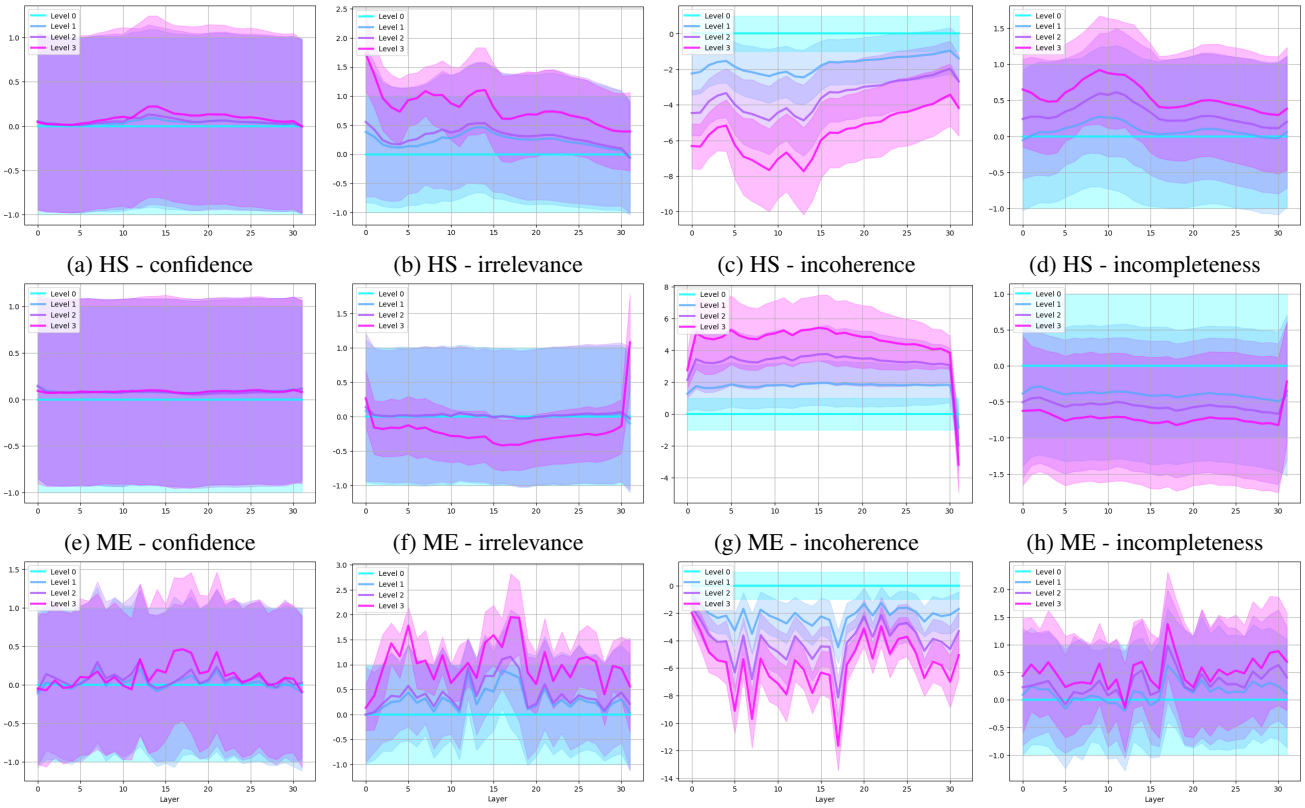
Figures 3a, 3e and 3i show that geometric statistics are not strongly correlated with *(verbal) confidence*. Figures 3b, 3f and 3j indicate that HS and AS respond to *irrelevant* responses as hallucinations, while ME does not. However, Figures 3c, 3g and 3k indicate that ME responds to *incoherent* responses as hallucination, while HS and AS do not. We note that HS and AS are sensitive to incoherence in that this property induces significantly lower scores, but this trend is the opposite when compared to *incorrectness* (hallucinations of *incorrectness* are associated with higher scores). Hence, if HS or AS are used as detector scores for *incorrectness* with a "greater than or equal to" threshold, they will not be able to detect *incoherence*. Figures 3d, 3h and 3l show that HS and AS respond to *incompleteness* as hallucination, while ME does not. Overall, these results indicate that geometric statistics are receptive to most alternative hallucination types, and that more severe hallucinations are associated with stronger responses from the statistics.

**Results on Hallucinations of Facts** Figure 4 depicts the distributions of geometric statistics for varying *(factual) incorrectness* on the individual domains. Each row of sub-figures corresponds to a single statistic (e.g., AS).

For `math` (Figures 4a, 4f and 4k), all statistics are most separable at layer 30, near the output of the LLM. Moreover, statistics which process hidden state vectors (HS and ME) perform the best. For `history` (Figures 4b, 4g and 4l), all statistics are most separable just after the midpoint of the LLM, at layer 16. HS and AS separate *factual* from *non-factual* responses relatively well for `history`. For `counting`, only AS is responsive to non-factual responses. Its most separable layer for `counting` is also 16.

Figures 4d, 4i and 4n depict the impact of domain shift on each of the statistics. The domain-specific distributions for each statistic are shifted and run parallel through the layers of the network. HS and ME are further impacted by the scale of the hidden states - their variance is much larger on `counting` than on `math`. These differences across domains impact detection performance on `all` subsets negatively. Figures 4e and 4j depict statistic responses across `all` domains. HS and ME show little to no separation on `all`. Of the three statistics, AS is least impacted by domain shift, leading to better performance on `all`. AS shows a small amount of separation at layers 16 and 31 (due to its performance on `history` and `math`, respectively).

Overall, the findings reveal that geometric statistics are impacted by domain shift, hindering their performance in more practical settings with variable domains. AS appears to be the least impacted by domain shift. We conjecture that this is due to the normalization of attention maps through the softmax function.

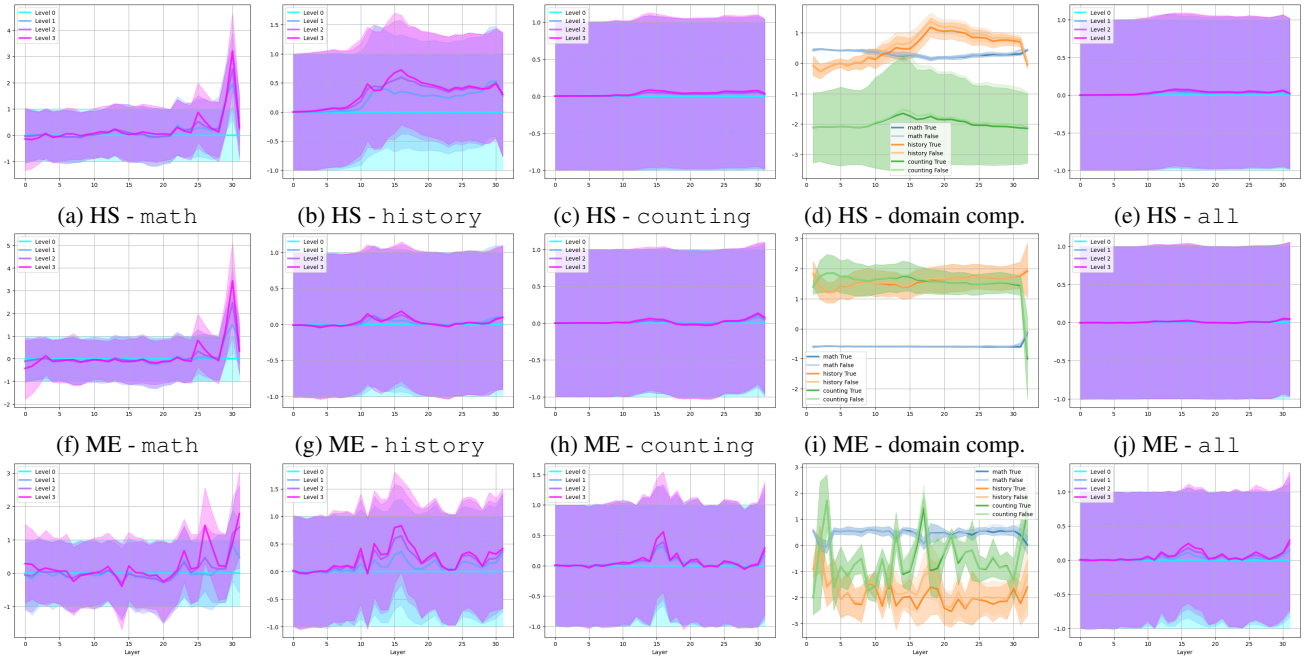



(a) HS - confidence (b) HS - irrelevance (c) HS - incoherence (d) HS - incompleteness

(e) ME - confidence (f) ME - irrelevance (g) ME - incoherence (h) ME - incompleteness

(i) AS - confidence (j) AS - irrelevance (k) AS - incoherence (l) AS - incompleteness

*Figure 3.* The response of geometric statistics (each row) are plotted for each alternative hallucination type (each column).



(a) HS - `math` (b) HS - `history` (c) HS - `counting` (d) HS - domain comp. (e) HS - `all`

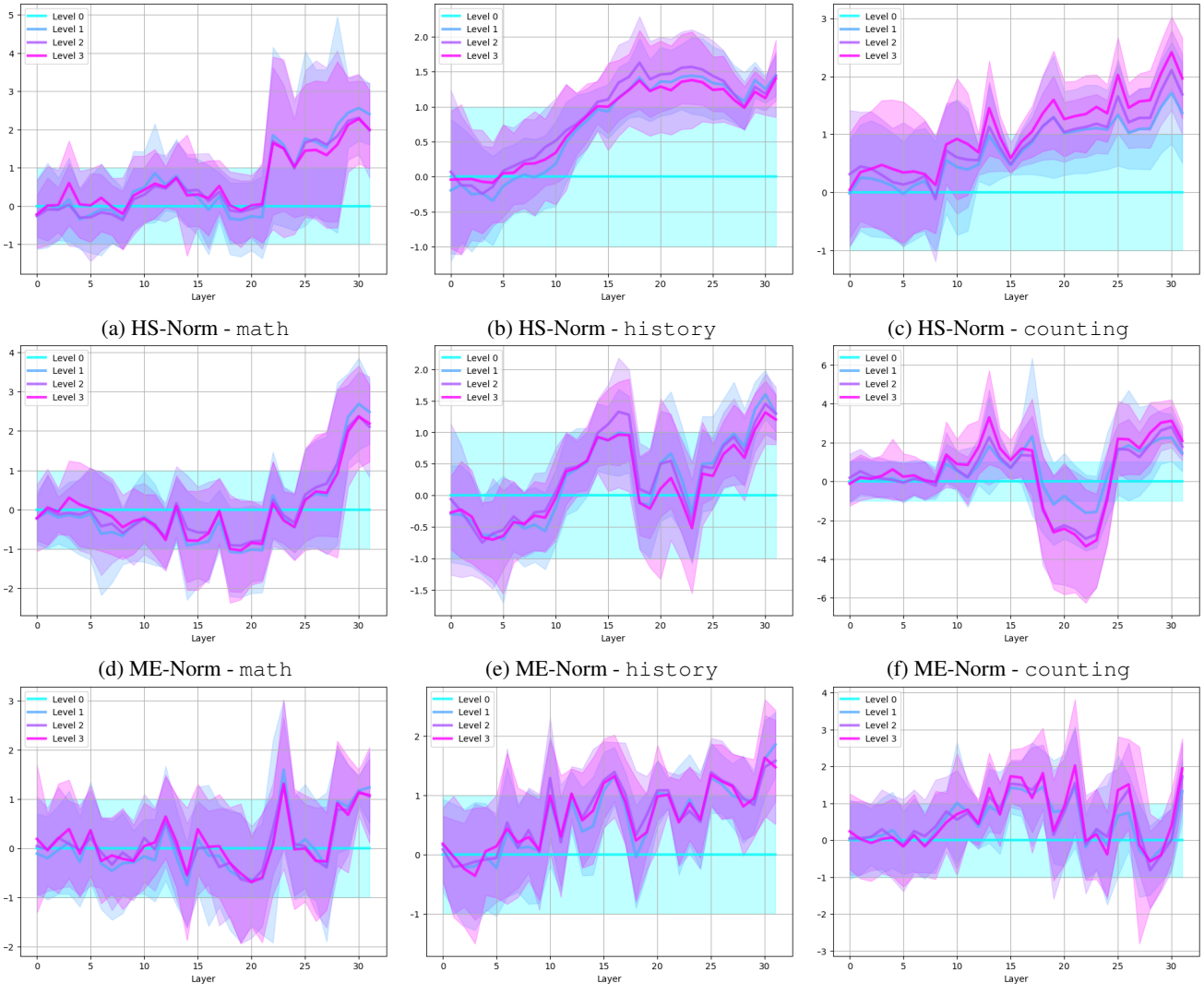
(f) ME - `math` (g) ME - `history` (h) ME - `counting` (i) ME - domain comp. (j) ME - `all`

(k) AS - `math` (l) AS - `history` (m) AS - `counting` (n) AS - domain comp. (o) AS - `all`

*Figure 4.* Geometric statistics (each row) are correlated with *(factual) incorrectness* on each domain (each column), motivating their use as hallucination detector scores. However, geometric statistics are impacted by domain shift, harming their cross-domain performance.


(a) HS-Norm - Domain Alignment (b) ME-Norm - Domain Alignment (c) AS-Norm - Domain Alignment

*Figure 5.* Domain alignment of the three geometric statistics using our proposed perturbation normalization technique.




(a) HS-Norm - `math` (b) HS-Norm - `history` (c) HS-Norm - `counting`

(d) ME-Norm - `math` (e) ME-Norm - `history` (f) ME-Norm - `counting`

(g) AS-Norm - `math` (h) AS-Norm - `history` (i) AS-Norm - `counting`

*Figure 6.* Geometric statistic distributions (augmented with perturbation normalization) for *incorrectness* on individual domains. Perturbation normalization reduces intra-domain variance as well, enhancing single-domain detection capability.