Learning Uncertainty from Sequential Internal Dispersion in Large Language Models

Ponhvoan Srey[1] **Xiaobao Wu**[2*] **Cong-Duy Nguyen**[3] **Anh Tuan Luu**[1,3*]
[1]Nanyang Technological University [2]Shanghai Jiao Tong University
[3]Centre for AI Research, VinUniversity
{ponhvoan002, anhtuan.luu}@ntu.edu.sg
duy.ntc@vinuni.edu.vn xiaobaowu@sjtu.edu.cn

Abstract

Uncertainty estimation is a promising approach to detect hallucinations in large language models (LLMs). Recent approaches commonly depend on model internal states to estimate uncertainty. However, they suffer from strict assumptions on how hidden states should evolve across layers, and from information loss by solely focusing on last or mean tokens. To address these issues, we present Sequential Internal Variance Representation (SIVR), a supervised hallucination detection framework that leverages token-wise, layer-wise features derived from hidden states. SIVR adopts a more basic assumption that uncertainty manifests in the degree of dispersion or variance of internal representations across layers, rather than relying on specific assumptions, which makes the method model and task agnostic. It additionally aggregates the full sequence of per-token variance features, learning temporal patterns indicative of factual errors and thereby preventing information loss. Experimental results demonstrate SIVR consistently outperforms strong baselines. Most importantly, SIVR enjoys stronger generalisation and avoids relying on large training sets, highlighting the potential for practical deployment.[†]

1 Introduction

Large Language Models (LLMs) have demonstrated impressive growth in performance across a wide array of tasks (Achiam et al., 2023; Grattafiori et al., 2024). Increasingly, they are being deployed in complex applications that require sophisticated reasoning, such as coding and mathematical reasoning (Guo et al., 2025). Despite widespread adoption, LLMs invariably suffer from unreliable generation or hallucination, frequently providing fictitious answers with complete confidence (Zhang et al., 2025). Due to their convincing response, it is challenging for users to determine factual correctness. This poses a major setback to LLM deployment, especially in high-risk domains.

Uncertainty estimation has emerged as a prominent solution to identify incorrect generations by aiming to accurately quantify the level of uncertainty in the response (Gal and Ghahramani, 2016; Hendrycks and Gimpel, 2016; Lakshminarayanan et al., 2017). Robust and accurate uncertainty estimation enables users to determine the level of trust to place in the LLM response, and intervene as necessary. However, recent approaches remain suboptimal, even compared to simple but effective methods, such as computing the entropy of the predictive output probability distribution (Vashurin et al., 2025). Second, a large majority relies on stochastic sampling to measure the consistency between a set of answers (Kuhn et al., 2023; Lin et al., 2024; Manakul et al., 2023). This introduces high computational overhead, rendering such methods impractical in real use cases.

Closely related is model probing, which involves training a lightweight classifier on top of last or mean hidden states to identify falsehoods by extracting world knowledge embedded in their internal states (Burns et al., 2022; Li et al., 2023; Azaria and Mitchell, 2023; Ji et al., 2024; Marks and Tegmark, 2023). These techniques are computationally efficient and enjoy state-of-the-art performance, but require training data and exhibit limited generalisability.

From current research, we identify two gaps in internal state approaches. First, reliance on task and model-specific heuristics for deriving scores reduces transferability across settings. Second, compression of evidence to last or mean-token summaries ignores important sequence-level patterns. In this work, we tackle hallucination de-

*Corresponding Authors.
†Our code repository is available online at https://github.com/ponhvoan/internal-variance.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 40088–40106

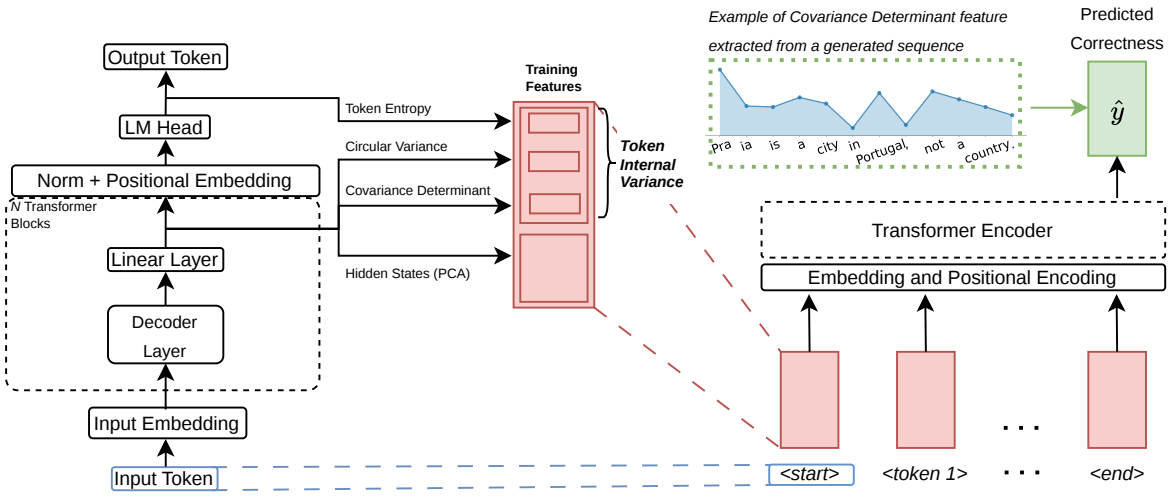


Figure 1: Illustration of our SIVR. At each generated token, we extract LLM hidden states, and compute their internal variance, consisting of token entropy, circular variance, and covariance determinant. We utilise these as informative features for sequence classification of response correctness with a simple transformer encoder architecture.

tection using internal LLM signals with an emphasis on cross-layer dynamics and ask: *Can uncertainty be inferred reliably from hidden states over the full token sequence with minimal assumptions to improve cross-task generalisability?* We introduce ***Sequential Internal Variance Representation (SIVR)***, a lightweight, supervised uncertainty estimation framework that (i) computes token-wise, layer-wise ***internal variance*** from hidden states, and (ii) dynamically learns to aggregate the full sequence to estimate uncertainty. SIVR operates on the simple, general premise that uncertainty is reflected in the degree of dispersion of internal representations across layers. To this end, we construct per-token internal variance that captures this dispersion, and learn patterns over the sequence to predict factual inaccuracy. Detailed formulation is presented in Section 2. SIVR is efficient to compute, and retains sequence dynamics that are critical for detecting factual errors. In summary, our contributions are as follows:

- We demonstrate the shortcomings of current works that use hidden states as a proxy signal for uncertainty. Internal variance, a novel and more robust feature that tracks the dispersion of hidden states across layers, is proposed.
- We introduce a pipeline that assesses all tokens, to fully take advantage of patterns suggestive of factual inaccuracy or hallucination.
- Our extensive experiments show that SIVR consistently outperforms strong baselines, and the incorporation of the proposed feature significantly enhances OOD generalisation.

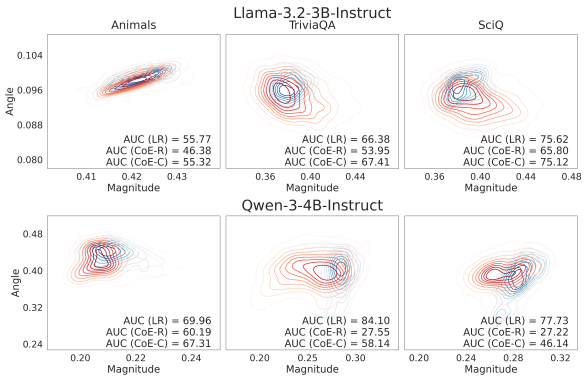


Figure 2: Visualisation of CoE features of correct and wrong answers. Significant overlap indicates CoE features provide poor discriminability.

2 Methodology

In this section, we present our proposed uncertainty estimation process in detail. First, we scrutinise a related method, Chain-of-Embeddings (CoE) (Wang et al., 2024), that produces an uncertainty score, and show its limitations. Then, we propose a more general feature and incorporate it into a unified framework that ascertains the uncertainty level in the model response.

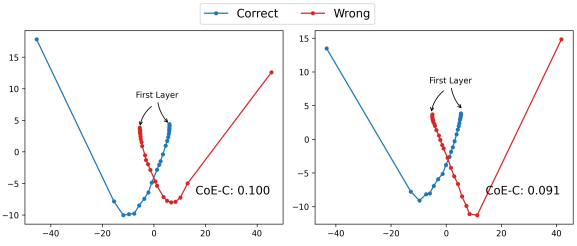


Figure 3: Visualisation (PCA-compressed) of two pairs of embeddings across layers. Each pair has identical CoE-C scores, but distinctive hidden states pattern, showing how CoE-C can degenerate even when the hidden states do not.

2.1 Current Limitations

Chain-of-Embeddings (CoE) (Wang et al., 2024) primarily inspects the magnitude and angle of the difference between successive layers of the hidden states. The authors found that answer correctness correlates positively with the magnitude, and negatively with angle, and combined the two features into an uncertainty score based on this observation. This places a strict constraint on the behavior of the hidden states that may not align across models and tasks. In Figure 2, the CoE features, Magnitude and Angle, of samples from other datasets are reproduced. Between the two models, the pattern of correct statements relative to the wrong ones varies, and there is little distinction between the two classes, resulting in poor discriminability.

Nonetheless, we argue that the hidden states across different layers remain a useful indicator of factual accuracy. For example, Figure 3 visualises two pairs of generated responses via PCA, where each pair has identical CoE-C scores, but the correct generations show distinctive hidden states features from the wrong ones. This showcases the situation where the score collapses as only the average "step-size" between layers is considered even though the latent paths are clearly different. Furthermore, in Figure 2, we feed the features to a logistic regression model, and by providing the correctness as labels, a consistently higher AUC is achievable. This confirms that the hidden states, particularly their behavior between layers, are useful features for discerning factual accuracy. Thus, we seek to define a more general and robust feature to encapsulate output uncertainty.

2.2 Internal Variance

To overcome the above limitations, we aim to formulate an uncertainty score based on a looser, more general assumption: *uncertainty is reflected in the spread of the hidden states across layers*.

As such, we aim to summarise *cross-layer* variability with compact statistics that capture complementary facets of dispersion and then *learn* patterns over the token sequence. To characterise latent space geometry, we distinguish between variation in magnitude and direction. Relying on a single metric obscures different degenerate modes. We therefore employ generalised variance to measure volumetric capacity and circular variance for directional diversity. Appendix A.1 shows that generalised variance is bounded by radial and directional components, demonstrating that jointly monitoring both is necessary to isolate distinct failure modes. We complement this geometric analysis by tracking output entropy, linking internal spatial characteristics to the predictive output distribution.

Generalised variance: First, we consider the determinant of the covariance matrix or generalised variance of hidden states for all layers. For the t-th output token s_t in a sequence $\boldsymbol{s}$ of total length T, denote the hidden states or embeddings at the l-th layer as $\boldsymbol{h}_t^l \in \mathbb{R}^d$, with sample covariance Σ where $l \in \{0, \ldots, L\}$ and d is the hidden states dimension. Since the hidden states are high-dimensional, for more stable covariance estimation, we compute the logarithm of the pseudo-determinant of the regularised covariance $\Sigma' = \Sigma + \alpha I_d$ for some small $\alpha > 0$. We approximate the generalised variance as

$$v_t = \log\det(\Sigma') = \sum_i \log \lambda_i, \qquad (1)$$

where λ_i are eigenvalues of Σ'. We use log-determinant $\log\det(\Sigma')$ as a compact summary of multidimensional spread. It aggregates the entire eigen-spectrum and is directly tied to differential entropy in the Gaussian case, an extension of Shannon entropy to continuous space (Zhouyin and Liu, 2025). The log-determinant is computed in a numerically stable and low-cost manner (see Appendix B).

Circular variance: Next, we adopt spherical or circular variance on normalised layer vectors $\hat{\boldsymbol{h}}_t^l$, which represents variation in different directions (Mardia and Jupp, 2009). This provides a complementary and scale-robust view of dispersion. We define circular variance c_t for token s_t as

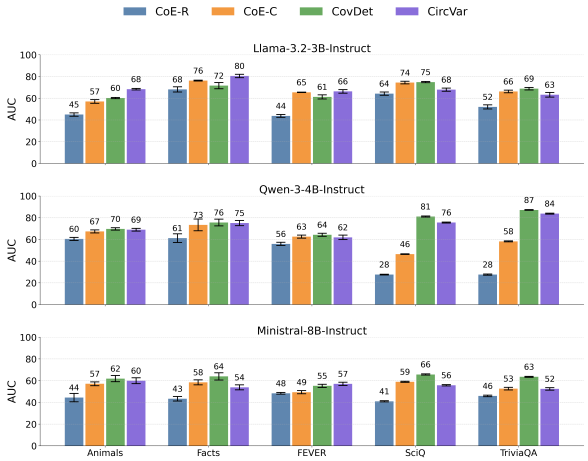


Figure 4: Performance of proposed features compared to CoE.

$$c_t = 1 - \left\| \frac{1}{L+1} \sum_{l=0}^{L} \hat{\boldsymbol{h}}_t^l \right\|. \tag{2}$$

Notably, circular variance also contains information on all pairwise relationships of hidden states at different layers (see Appendix A.2), instead of being restricted to relationships between successive layers as with CoE. To understand the effectiveness of these new metrics, we compare with CoE-R and CoE-C. In Figure 4, similar to CoE which gives a sequence score, circular variance (CirCVar) and covariance determinant (CovDet) is computed with the mean hidden states $\bar{\boldsymbol{h}}^l = \frac{1}{T}\sum_t^T \boldsymbol{h}_t^l$. CovDet and CircVar, on average, yields better AUC (Area Under the receiver operating Curve) than CoE-R and CoE-C. CoE-C generally is quite strong, but in some cases, such as with Qwen on SciQ, it severely underperforms. Meanwhile, the performance of CovDet and CircVar remains robust, indicating their broad applicability.

Token entropy: Finally, we include predictive entropy to quantify uncertainty in the probability distribution emitted by the decoder, which also encodes relevant information on output uncertainty. We define per-token entropy

$$e_t = H\left(p(s_t)\right) = -\sum_{s_t \in \mathcal{V}} p(s_t) \log p(s_t) \tag{3}$$

where $p(s_t)$ is the softmax probability of token $s_t \in \mathcal{V}$, and $\mathcal{V}$ is the model vocabulary set. Collectively, we formulate the ***internal variance*** at each token as $\boldsymbol{v}_t = [v_t, c_t, e_t]^\top$.

Together, these three signals — generalised variance for magnitude, circular variance for directionality, and predictive entropy for output-space uncertainty — offer a holistic characterisation of uncertainty that tracks how internal dispersion evolves across layers and how it manifests in the output distribution. Our ablation studies in Table 2 show that each component contributes and the combination yields the strongest performance.

2.3 Token Aggregation

We define a per-token dispersion feature $\boldsymbol{v}_t$ and seek a sequence-level uncertainty score. Relying on a single token can miss informative temporal structure. For example (Figure 1), in `Praia is a city in Portugal, not a country`, the generalised variance is initially high then stabilises, but a sharp spike at `Portugal` flags the error, a pattern obscured by last or mean-token summaries. Therefore, we propose to learn from the full token-wise sequence of dispersion features. We choose to focus on supervised uncertainty learning as this paradigm produces uncertainty scores that are aligned explicitly with a notion of correctness (Liu et al., 2024; Srey et al., 2026). Unlike UHead that mean-pool token encodings (Shelmanov et al., 2025) for a claim span, we retain order to capture patterns indicative of factual correctness.

Formally, we have responses of variable length $\mathcal{D} = \{\boldsymbol{s}_i, y_i\}_{i=1}^n$ where $\boldsymbol{s}_i \in \mathbb{R}^{T_i \times d}$ and $y_i \in \{0, 1\}$ with 1 for hallucinated instances. From each sequence $\boldsymbol{s}_i$, informative training features $\boldsymbol{x}_i$, with core components comprising the proposed internal variance $\boldsymbol{v}_t$, are extracted. The goal is to learn learn $f_\theta : \mathbb{R}^{T \times d_{\text{tr}}} \to [0, 1]$ to estimate $p_\theta(y = 1|\boldsymbol{x})$, which can be regarded as the uncertainty score for the sequence $\boldsymbol{s}$. d_{tr} is the dimension of the training features. The objective function is the binary cross entropy loss with l_2 regularisation

$$\frac{1}{n}\sum_{i=1}^{n} -y_i \log \hat{y}_i - (1 - y_i) \log (1 - \hat{y}_i) + \beta \|\theta\|_2^2 \tag{4}$$

We employ a lightweight sequence head with an embedding layer to project input to a hidden dimension of 128, followed by a transformer encoder block, and finally, a linear classifier layer to produce the sequence score. More details on architecture and hyperparameters are provided in Appendix C.

Method	TriviaQA			SciQ			MedMCQA			MGSM			MATH			MMLU			CommonsenseQA			Average			Mean Rank
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	
P(True)	66.63	93.75	80.36	58.66	93.13	79.04	56.37	92.59	51.15	50.79	96.67	42.17	68.07	81.67	71.30	**52.63**	93.53	74.28	50.06	91.41	30.58	58.69	81.54	50.54	11.06
Max Prob	76.82	75.19	83.24	71.54	78.01	84.05	60.77	86.88	56.76	60.49	95.34	51.94	60.68	93.57	66.27	47.67	96.27	71.42	58.24	87.23	40.31	66.04	81.92	59.12	9.54
Perplexity	78.40	73.43	85.06	72.42	77.66	84.63	61.57	87.26	57.77	61.42	94.90	52.78	61.66	94.21	67.05	47.76	96.52	71.57	58.38	89.50	40.38	66.31	82.05	59.00	9.51
Entropy	80.46	71.68	85.96	72.85	80.41	85.70	62.76	84.22	58.62	64.65	93.57	55.57	62.77	93.89	68.35	48.00	96.02	71.72	59.87	86.52	41.38	67.63	80.42	60.45	7.96
Temp-Scaled	74.27	77.44	81.62	70.71	78.35	82.71	60.26	89.35	56.46	57.33	94.46	48.79	60.18	93.57	65.69	47.66	96.27	71.38	57.30	90.45	39.54	64.53	82.72	57.34	11.19
Energy	76.50	74.94	82.01	72.73	79.73	86.21	62.20	82.32	57.79	67.14	67.85	55.47	64.33	88.10	69.14	47.33	96.02	72.44	60.82	86.75	41.75	68.19	76.62	61.20	7.40
SE	84.44	**40.26**	87.69	79.44	77.08	92.03	**66.88**	80.81	67.01	66.34	88.89	50.44	67.27	78.41	68.09	47.82	92.31	72.09	57.47	89.78	39.59	68.87	76.21	62.96	7.13
SAR	84.72	55.84	89.27	73.98	72.92	89.17	64.71	76.47	**72.08**	39.67	97.62	33.70	68.84	70.06	78.30	45.39	98.08	73.01	57.15	86.86	37.48	65.63	81.47	60.98	8.57
CoE-R	53.45	98.25	69.89	65.41	94.16	84.73	47.47	95.63	45.27	54.84	92.90	45.15	59.50	88.42	65.57	49.87	95.52	72.66	51.20	94.15	32.52	54.22	92.67	49.29	13.42
CoE-C	66.97	91.48	79.86	75.06	83.85	89.14	62.14	83.84	57.16	46.24	90.24	37.03	58.67	90.68	63.95	50.42	94.53	73.10	61.38	91.05	41.70	61.25	88.54	55.13	11.08
SATMD + MSP	85.79	65.79	92.14	78.75	80.70	91.07	63.51	90.99	54.82	72.27	77.89	55.40	77.14	71.43	71.22	47.03	94.25	68.08	64.04	87.65	40.16	71.16	77.35	62.42	7.22
Lookback Lens	76.95	76.92	84.06	80.54	77.59	92.09	58.25	83.18	49.86	71.05	83.33	62.66	75.19	77.36	78.25	47.14	96.10	69.58	53.37	89.71	36.83	72.20	71.78	65.42	6.49
SAPLMA	77.94	83.54	84.84	83.62	60.34	93.00	65.63	86.79	61.89	**76.48**	72.53	**70.33**	78.06	87.10	81.69	50.49	95.00	**75.42**	62.45	90.48	48.56	74.72	71.53	70.17	4.25
TAD	88.25	59.26	92.46	85.16	77.59	**94.07**	63.56	79.82	52.61	76.28	**41.51**	51.31	74.91	76.06	71.92	50.19	91.76	71.39	**68.00**	**82.80**	**53.38**	78.34	58.14	71.66	3.10
SIVR (Ours)	**89.31**	41.56	**93.67**	**85.70**	**52.54**	93.68	66.59	**72.82**	63.68	75.65	74.42	69.23	**83.48**	**42.11**	**86.00**	50.89	**84.42**	74.97	66.57	90.53	48.59	**79.53**	**55.46**	**76.53**	**2.08**

(a) Llama-3.2-3B-Instruct

Method	TriviaQA			SciQ			MedMCQA			MGSM			MATH			MMLU			CommonsenseQA			Average			Mean Rank
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	
P(True)	65.04	93.29	65.35	68.76	86.18	80.67	57.49	88.50	48.15	56.43	95.85	21.02	57.56	94.01	56.37	47.98	96.43	68.41	45.09	91.78	21.41	58.56	86.36	41.61	12.42
Max Prob	77.30	66.08	73.94	77.63	62.35	83.97	60.72	95.88	50.93	58.54	96.19	35.27	68.55	76.65	62.31	49.79	97.02	70.02	57.95	89.04	30.19	65.31	78.48	45.64	10.33
Perplexity	78.18	65.68	75.39	78.59	62.35	85.17	61.14	94.85	50.97	58.82	96.19	35.32	69.19	77.25	63.29	49.67	97.02	69.66	58.80	88.77	30.52	65.80	78.48	46.37	9.54
Entropy	78.58	67.06	76.25	77.49	66.47	85.13	62.39	93.81	52.44	61.45	96.19	38.87	71.20	71.86	65.12	49.10	96.43	69.16	60.63	88.49	32.36	67.04	77.79	47.17	8.74
Temp-Scaled	76.65	66.86	72.71	77.87	61.76	83.55	60.12	95.19	50.12	57.20	96.54	31.35	67.45	78.44	61.01	50.02	97.02	70.17	56.82	90.68	29.42	64.68	79.82	45.20	10.97
Energy	67.50	87.18	67.47	61.56	91.47	75.54	64.33	85.91	54.20	83.73	58.82	54.66	78.10	**55.69**	73.22	47.48	92.86	67.25	65.27	87.95	37.84	67.95	81.72	49.71	8.49
SE	81.65	54.76	83.37	79.01	74.63	87.85	66.95	88.46	**66.56**	66.00	76.67	56.88	73.58	70.73	**76.55**	50.56	96.83	71.36	62.46	79.45	34.27	69.17	75.25	53.09	6.83
SAR	83.66	45.24	84.35	72.78	82.09	82.69	68.95	82.69	63.57	63.71	66.67	52.09	70.30	78.43	66.78	49.54	92.06	69.51	53.58	91.78	37.03	65.83	75.21	49.59	8.65
CoE-R	37.75	100.00	44.87	32.31	100.00	57.84	50.90	94.85	41.52	66.09	95.85	32.20	53.39	91.62	50.66	48.34	93.45	69.83	51.30	93.97	25.31	47.24	95.72	36.04	13.86
CoE-C	65.14	93.89	68.58	69.56	90.88	82.59	56.47	94.16	46.91	52.18	94.46	30.61	70.85	84.43	70.53	50.53	93.45	70.47	54.01	92.60	27.23	57.17	91.31	42.11	12.54
SATMD + MSP	82.45	61.39	78.79	84.83	**51.67**	92.69	69.94	72.41	56.37	**90.79**	35.09	69.12	71.40	64.86	61.95	47.91	96.43	73.15	63.13	94.67	34.32	74.55	65.76	56.27	5.24
Lookback Lens	82.94	59.09	87.25	83.78	80.00	90.63	66.70	**63.03**	52.10	79.56	64.56	**71.44**	61.06	92.16	50.41	46.13	93.44	72.04	69.93	76.63	41.89	72.11	75.77	57.47	6.22
SAPLMA	83.70	62.38	85.63	87.41	57.35	92.91	63.53	81.03	51.66	81.03	41.38	48.44	**79.87**	75.76	69.40	54.64	97.06	74.80	61.37	78.38	31.75	76.72	64.05	62.01	4.46
TAD	87.02	50.53	**89.48**	83.60	65.67	91.69	65.09	92.44	57.04	77.37	56.03	55.88	76.96	75.61	69.68	50.38	95.89	75.34	**74.48**	73.63	**51.26**	75.75	67.19	61.58	4.33
SIVR (Ours)	**89.48**	**39.81**	89.27	**87.79**	54.55	**94.14**	**75.95**	67.24	59.57	89.04	**17.54**	58.13	79.05	67.57	72.30	**64.07**	**91.30**	**88.60**	68.32	**73.13**	50.28	**80.86**	**55.37**	**64.89**	**2.06**

(b) Llama-3.1-8B-Instruct

Method	TriviaQA			SciQ			MedMCQA			MGSM			MATH			MMLU			CommonsenseQA			Average			Mean Rank
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	
P(True)	51.12	94.51	48.24	49.92	94.52	56.88	53.21	95.23	52.92	52.28	96.16	31.90	57.22	98.55	56.84	48.85	95.53	74.96	48.64	96.09	29.20	50.57	94.99	37.51	13.61
Max Prob	80.95	51.70	75.91	72.56	74.20	74.93	52.87	94.40	53.95	68.87	82.27	47.84	64.56	92.55	65.10	52.56	94.41	76.37	54.30	90.53	32.44	65.11	82.66	50.74	9.35
Perplexity	82.00	51.70	78.34	73.78	72.83	76.20	53.84	94.61	54.51	68.97	80.35	47.65	65.22	91.51	65.77	51.90	94.97	76.33	54.89	91.12	32.67	66.50	80.38	51.90	7.96
Entropy	81.40	57.77	78.71	72.98	77.85	75.79	54.57	94.19	55.48	69.72	78.14	45.82	66.26	91.72	66.94	50.96	94.13	76.23	54.94	92.66	33.12	65.99	86.12	51.45	8.92
Temp-Scaled	80.61	56.25	75.08	72.10	73.06	74.59	52.74	94.61	53.50	67.67	83.46	47.03	64.02	91.51	64.46	52.86	94.69	76.48	54.40	89.47	32.29	65.41	80.00	50.53	9.11
Energy	74.48	75.38	72.22	71.65	81.51	75.28	52.95	92.32	54.18	69.46	80.65	43.50	70.85	83.85	68.19	49.25	93.85	75.35	53.81	91.95	32.87	66.31	80.56	50.45	8.86
SE	79.85	66.67	83.30	78.98	81.08	**87.83**	55.09	83.33	61.41	73.94	63.08	60.17	69.83	**65.31**	72.89	48.03	**87.50**	74.23	58.39	89.29	36.64	67.82	76.20	56.77	6.60
SAR	85.62	70.83	83.88	78.51	67.57	83.35	54.68	78.57	58.47	70.44	69.84	53.94	70.87	75.51	73.46	49.34	95.83	76.16	55.95	81.43	32.69	68.95	73.73	54.27	6.17
CoE-R	45.11	100.00	47.43	41.16	100.00	53.06	44.05	97.93	47.73	53.07	96.01	34.22	48.25	97.10	48.80	47.88	96.37	74.33	44.77	96.21	26.26	45.74	97.29	36.86	14.44
CoE-C	53.37	94.13	53.06	58.55	93.61	68.29	51.01	95.23	51.54	60.88	90.10	38.13	55.53	94.41	57.88	46.20	96.93	73.75	49.91	96.09	30.14	52.27	94.50	39.98	13.53
SATMD + MSP	84.81	56.48	82.27	77.46	**62.20**	83.34	62.64	83.33	61.25	**86.38**	**53.97**	**74.94**	77.71	82.50	78.41	49.46	97.22	76.97	64.57	**75.57**	34.52	72.84	72.81	61.41	4.68
Lookback Lens	76.17	86.96	80.96	72.93	88.16	83.28	**65.78**	91.43	**61.66**	78.06	66.67	43.78	75.54	84.48	71.81	52.66	92.54	77.79	63.53	89.53	38.20	73.09	77.67	58.33	5.64
SAPLMA	80.62	67.92	79.08	73.55	82.95	79.42	49.24	90.72	53.67	70.34	88.89	57.77	**81.64**	70.10	**82.49**	54.66	90.28	80.01	**68.01**	79.88	**47.81**	73.30	70.24	62.09	4.92
TAD	83.29	59.41	83.35	75.54	84.44	80.90	58.73	**77.05**	60.93	79.11	62.96	63.11	78.51	68.13	71.86	54.89	94.74	**81.29**	58.93	87.15	30.66	75.24	68.90	61.78	4.06
SIVR (Ours)	**87.29**	**40.38**	**87.07**	**80.81**	63.95	86.03	61.12	82.08	60.76	84.70	58.78	73.12	77.56	72.16	77.19	**55.54**	91.46	77.69	66.39	86.59	46.40	**78.11**	**60.81**	**66.18**	**2.17**

(c) Ministral-8B-Instruct

Table 1: Main results. Cooler color indicates better results, with the best bolded and second best underlined.

3 Experiments

3.1 Setup

Datasets. We evaluate extensively on twelve datasets for two general tasks, fact-checking and question answering (QA). In fact-checking, a statement is provided for an LLM to verify its factuality. The setting is identical to TruthEval (Khatun and Brown, 2024), and the anchor experiments by Jin et al. (2025). For this task, we select statements from a broad range of topics from the datasets Counterfact (Meng et al., 2022), Common Claims (Casper et al., 2023), Animals and Facts (Azaria and Mitchell, 2023), and FEVER (Thorne et al., 2018). For clearer presentation, we report performance for fact-checking tasks in Appendix D.2. In the ablation, we refer to the collection of the first four datasets as "True-False". For QA, we evaluate both free-form, and multiple-choice answers, across multiple important domains: TriviaQA (Joshi et al., 2017), SciQ (Welbl et al., 2017) for general knowledge; MedMCQA (Pal et al., 2022),

MMLU (Hendrycks et al., 2020) for more specialised medical and scientific knowledge; MGSM (Shi et al., 2022), MATH (Hendrycks et al., 2021), CommonsenseQA (Talmor et al., 2019) for mathematical and reasoning tasks. Queries from each dataset are provided as prompts for an LLM, and the responses are obtained for evaluation. Corresponding binary correctness labels are extracted via exact matching with reference answers *e.g.* MedMCQA and MMLU, or with ROUGE-L (Lin, 2004) *e.g.* TriviaQA and SciQ. More details are found in Appendix C.

Baselines. We compare with four classes of uncertainty estimation baselines: (1) **Logit-based**: Maximum Sequence Probability (MSP or Max Prob), Entropy (Huang et al., 2023), Perplexity (Si et al., 2022), Temperature-Scaled MSP (Shih et al., 2023), and Energy (Liu et al., 2020). (2) **Sampling-based**: Semantic Entropy (SE) (Kuhn et al., 2023), SAR (Duan et al., 2024). (3) **Confidence elicitation**: P(True) (Kadavath et al., 2022), which prompts the model to determine if the answer is correct, and the probability of the token "True" is extracted as a confidence score. (4) **Internal states:** CoE (Wang et al., 2024); and four supervised baselines: SAPLMA (Azaria and Mitchell, 2023): trained on last-token hidden states; SATMD + MSP (Vazhentsev et al., 2025c): trained on Mahalanobis distances of hidden states at each layer, averaged across tokens, and concatenated with MSP; Lookback Lens (Chuang et al., 2024): trained on lookback ratios, the proportion of attention on context to attention on both generated and context tokens; and TAD: a two-stage process that trains a probe on attention weights, token probabilities, and intermediate scores from the first stage (Vazhentsev et al., 2025a).

Language Models. For the main experiments, we evaluate with Llama-3.2-3B, Llama-3.1-8B (Grattafiori et al., 2024), and Ministral-8B (Mistral AI, 2024). In Appendix D.1, we provide supplementary results with Qwen-3-4B and Qwen-3-14B (Yang et al., 2025). We work with the instruction-tuned versions, except for Qwen-3-14B. For reproducibility, we generate all responses via greedy decoding.

Metrics. We employ AUC, FPR@95, and AUPR (Davis and Goadrich, 2006) as evaluation metrics to comprehensively assess the quality of the uncertainty scores for discerning correct from wrong generations. AUC (Area Under the receiver operating Curve) is a threshold-invariant metric that measures the classifier's ability to distinguish between positive and negative classes. A higher AUC indicates stronger discriminative power. FPR@95 quantifies the proportion of negatives incorrectly classified as positives when a true positive rate or recall of 95% is achieved, a particularly informative metric for safety-critical contexts with stringent recall requirements. A low FPR@95 is better. Finally, AUPR (Area Under the Precision-Recall curve) is a threshold-invariant metric that is concerned with the trade-off between precision and recall. A higher AUPR reflects the classifier's ability to maintain precision while retrieving positives effectively.

3.2 Results and Analysis

Main Results. Table 1 presents the main results. First, our approach consistently achieves better results compared to its competitors. On average, across all datasets and models, the strongest baseline, TAD, obtains 76.44%, 64.74%, and 65.01% on the three metrics. In comparison, SIVR achieves 79.50%, 57.21%, and 69.20%, amounting to an improvement of 3.06%, 7.53%, and 4.19% on AUC, FPR@95, and AUPR, respectively. Similar to previous findings (Fadeeva et al., 2023; Vashurin et al., 2025), simple methods such as Max Prob are relatively effective and stable. SAR and SE slightly outperform these methods, but are not consistent. This may be attributed to the difficulty sampling-based approaches face in detecting self-consistent errors (Tan et al., 2025). Likewise, most other methods lack stability. P(True) generally gives weak results, but on certain tasks, such as Common Claim, the performance is fair. This may be due to sensitivity of input prompts and the token selected for confidence elicitation. In some cases, CoE is competitive even with supervised baselines, such as on SciQ and MedMCQA (Llama-3.2-3B-Instruct), but performance varies significantly across tasks. This reinforces the observation that CoE could provide SOTA results, but typically underperforms because their assumption does not hold in general. In line with prior works, supervised methods are the most optimal and stable. They are comparable on most tasks, but on the whole, SAPLMA surpasses both SATMD and Lookback Lens, suggesting that hidden states themselves are more informative than Mahalanobis distances and lookback

Variant	True-False			Fever			TriviaQA			SciQ			MedMCQA			GSM			Math			MMLU			CommonsenseQA			Average		
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP
HS (PCA)	79.33	72.22	53.13	80.78	49.69	50.83	84.65	61.54	84.23	76.31	76.74	81.93	59.66	83.96	62.94	80.84	54.20	67.80	77.23	68.04	78.54	56.33	85.37	77.65	63.08	83.54	45.43	74.68	71.00	63.49
HS	82.98	47.48	52.86	78.54	55.35	47.43	81.77	62.50	79.73	75.29	68.60	77.98	55.89	90.57	53.13	82.49	50.38	68.71	80.01	62.89	79.21	55.00	90.24	74.40	64.58	95.12	47.22	75.46	63.80	61.60
CovDet, AngVar	77.38	60.36	53.43	75.04	66.67	38.91	69.85	73.08	70.00	76.58	61.63	79.54	63.41	97.17	60.67	70.28	77.10	54.48	72.85	75.26	71.69	57.88	89.02	79.78	60.32	85.98	45.09	71.31	72.28	59.49
Internal Variance	83.08	57.45	57.34	73.98	76.10	41.60	83.76	38.46	80.08	81.92	56.98	85.32	64.53	96.23	63.10	80.97	57.25	67.21	78.25	72.16	80.57	55.72	89.02	75.70	64.05	87.20	46.06	76.29	66.93	64.08
Full (RNN)	78.56	61.46	44.97	68.02	71.70	34.93	85.45	50.00	84.71	81.35	72.09	86.19	59.48	80.19	56.29	83.65	45.80	69.79	72.07	72.16	71.70	57.15	89.02	77.37	64.18	86.59	45.14	73.80	67.78	58.83
SIVR	85.37	45.06	57.54	82.44	54.09	55.71	87.29	40.38	87.07	80.81	63.95	86.03	61.12	82.08	60.76	84.70	58.78	73.12	77.56	72.16	77.19	55.54	91.46	77.69	66.39	86.59	46.40	78.11	60.81	66.18

Table 2: Ablation with Ministral-8B-Instruct

ratios. Meanwhile, our approach additionally inspects internal variance features sequentially, enabling us to more accurately determine the likelihood of incorrect or hallucinated generations. We further show that this inclusion of internal variance is crucial to maintain performance in out-of-distribution settings.

Ablation Study. To understand the effectiveness of the components in SIVR, we conduct an ablation. Two settings are considered: one only using the hidden states, and one only the internal variance. For the first setting, we test with hidden states, and with top-10 principal components (PC) to reduce overfitting. We conduct a more extensive ablation on the number of PCs in Appendix D.4. Then, we test only the proposed CovDet and CircVar, as well as the whole internal variance, which includes token entropy. Finally, we employ a recurrent neural network (RNN) as the architecture. The results are provided in Table 2. Using only the hidden states slightly outperforms SAPLMA. As using only the PC hidden states yield reasonable performance, these are adopted into the complete SIVR. On the other hand, with just CovDet and CircVar, we achieve an AUC of 71.31%, only slightly worse than SAPLMA. Further inclusion of token entropy and hidden states boost performance considerably. Notably, even though internal variance is only 3-dimensional compared to the much higher dimensional hidden states, using internal variance as features alone yields comparable results to using the full hidden states. This demonstrates that internal variance is informative as training features. The combination of both internal variance and hidden states in the full SIVR framework gives the optimal results. As the RNN is not as effective as the transformer encoder, transformer is used in the main method.

Out-of-distribution (OOD) Evaluation. While the supervised paradigm is SOTA, performance usually degrades drastically under the OOD setting where test data differ from the training set.

Test On	Trivia	SciQ	MCQA	MGSM	MATH	MMLU	CSQA	Avg	Test AUC
Lookback Ratios	-18.71	-14.89	-11.43	-18.43	-17.58	0.02	-13.4	-13.49	55.74
Hidden States	-43.32	-30.53	-18.05	-25.19	-24.12	-2.56	-8.26	-21.72	49.44
Internal Variance	-10.61	-15.32	-8.83	-21.82	-17.34	-3.68	-7.62	-12.18	60.56
SIVR	-17.44	-13.7	-2.7	-25.97	-23.36	-5.16	-14.36	-14.67	58.67

Table 3: Average change in AUC with out-of-distribution training data.

Specifically, to test and compare the robustness of SIVR, we utilise: (i) lookback ratios; (ii) only the hidden states; (iii) only the internal variance; and (iv) the full SIVR features. Lookback ratios were specifically selected for comparison as they boast strong transfer capability. Figure 5 visualises the change in AUC, and Table 3 summarises the result for each configuration. As expected, a considerable decline in AUC is observed in all cases. Nonetheless, internal variance is the most robust, leading to only a 12.18% decrease compared to using hidden states which results in a drop of 21.72%. OOD AUC with internal variance even exceeds that using lookback ratios by 4.8%, which were designed to transfer across tasks. We attribute this to the fact that hidden states are higher dimensional and can encode more task-specific information, making them easier to overfit, whereas internal variance captures more invariant cross-layer dispersion patterns associated with errors, hence better transfer. This behaviour is also aligned with prior observations that hidden-states-based probes can overfit and degrade substantially when transferred to OOD tasks (Chuang et al., 2024). AUPR and FPR@95, provided in Appendix D.5, further corroborate our findings. Overall, this underscores the contribution of internal variance, particularly to improve generalisability.

Training Data Size. Figure 6 plots the evaluation with the three metrics for varying number of training instances. More training data benefit performance. However, even with only 128 data points, a satisfactory result is achieved, reducing the dependence on training data size.

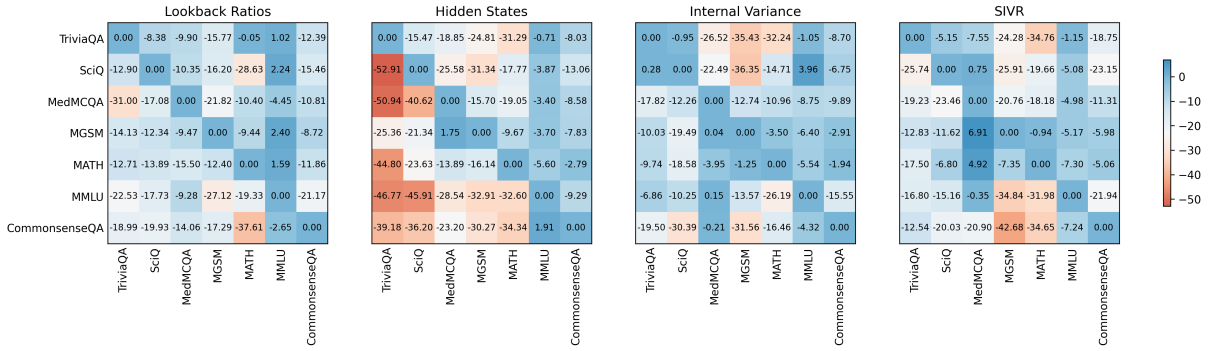


Figure 5: Change in AUC under OOD setting with Ministral-8B-Instruct. Training and test data are on vertical and horizontal axis, respectively.

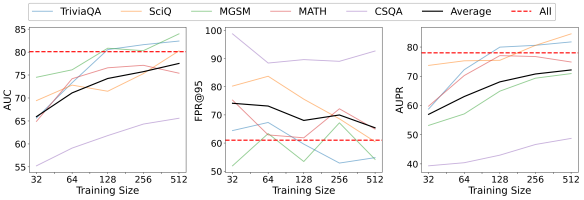


Figure 6: Effect of training data size.

4 Related Work

Interest in uncertainty estimation for language models has led to a surge in proposals. Recent works build on conventional information-based approaches, such as entropy, and propose adaptations specific to language generation by combining logit-level and language-level uncertainty estimation (Kuhn et al., 2023; Duan et al., 2024; Zhang et al., 2023). Kuhn et al. (2023) noticed that simple information-based methods inappropriately increased uncertainty for syntactically different but semantically identical generations, and introduced semantic uncertainty to address this discrepancy. Duan et al. (2024) proposed SAR to eliminate the effects of prevalent but semantically trivial tokens. Concerned with a similar issue, Stronger Focus (Zhang et al., 2023) estimates entropy, weighted by attention values. Vazhentsev et al. (2025b) discovered that, for hallucinated tokens, attention weights noticeably drop for "uncertainty-aware" heads, which are simply those with the highest average attention. Subsequently, confidence scores are computed for these heads and aggregated for the sequence.

To extract internal knowledge in LLMs, the next suite of techniques formulate an uncertainty score from model behaviour, mainly focusing on the hidden states. At the broadest level, we categorise these techniques into probe-free and probe-based methods. Probe-free methods utilise the hidden states to derive an uncertainty score. They can be further divided into (i) Sampling-based, such as EigenScore (Chen et al., 2024) which measures consistency in hidden states of a set of sampled responses; (ii) density-based approaches, which take inspiration from out-of-distribution detection (Yoo et al., 2022; Ren et al., 2022), and compute some measure of distance, notably the Mahalanobis distance (Lee et al., 2018) to the data centroid; and (iii) single-call approaches, such as CoE (Wang et al., 2024) which tracks the change in hidden states of successive layers from a single generation as a reflection of predictive uncertainty.

On the other hand, probe-based methods train a lightweight classifier or *probe* on features extracted from the LLM. Azaria and Mitchell (2023) took hidden states at the last layer of the last token directly. Subsequent works focus on extracting more informative features. For example, Vazhentsev et al. (2025c) computed their Mahalanobis distances of embeddings at each layer, averaged over tokens in the sequence, and concatenated them with the maximum sequence probability. On top of activation maps at each layer, He et al. (2024) also considered the top-k probabilities and indices of the last token. Other works seek to alleviate the burden of label annotation, for instance, through automatic data generation and labelling with high-capacity commercial models (Su et al., 2024; Vazhentsev et al., 2025b), or by utilising alternative signal sources from the model itself, like self-verbalised confidence (Srey et al., 2025), as soft pseudolabels. However, performance, especially for logits and language-based scores, is normally subpar compared to simpler methods (Vashurin et al., 2025), and often, they rely on sampling, resulting in high computational

costs. Methods that formulate uncertainty using hidden states, such as CoE (Wang et al., 2024), have strict assumptions. Meanwhile, their probe-based counterparts are empirically strong, but exhibit limited generalisation. To fill this gap, SIVR introduces a principled label-efficient, and generalisable sequence-aware signal.

Moreover, current methods mostly output an uncertainty score assessed solely on the last token (He et al., 2024), mean token (Wang et al., 2024), or as a mean of token scores (Vazhentsev et al., 2025b), disregarding relevant tokens and information that encode uncertainty. Shelmanov et al. (2025) passed attention features through a transformer, and averaged encoder outputs across claim tokens before classification. In contrast, SIVR operates without claim segmentation, uses a new family of cross-layer internal-variance features, and preserves token order to learn sequence-level dispersion patterns, enabling fine-grained token-level attribution even when claim spans are unavailable (see examples in Appendix D.7).

5 Conclusion

In this work, we introduced SIVR, a novel factual detection framework that tracks model internal dispersion as it generates a response. Given the hidden states at each token, we calculate their internal variance, a measure of their dispersion between layers. Internal variance, along with the hidden states, are extracted as training features for sequence classification. Following our framework, SIVR dynamically learns patterns indicative of factual correctness. SIVR demonstrates strong and stable performance, consistently surpassing competitive baselines. Crucially, we show that internal variance significantly enhances out-of-distribution generalisation. Moreover, a large training dataset is not needed, outlining the potential for real-world applications. In future work, there are several directions to consider. Other informative features, such as semantic importance, could be incorporated as extra signals for uncertainty. More broadly, to reduce reliance on annotated data, unsupervised approaches leveraging internal variance offer a promising direction towards practical solutions for factual detection. Further integration into the decoding process can enable inference-time hallucination detection and mitigation, and improve model reliability.

Limitations

We believe our work has the following limitations.

Supervised Learning. While SIVR remains relatively robust to OOD tasks and works well with few data points, it is still supervised and requires an annotated training data.

LLM Variety. We experimented with moderately sized LLMs, ranging from 3 to 14 billion parameters. Experiments with more capable models of larger size could further validate our claims.

Interpretability. Our work opens up new avenues for exploring the interpretability of the training features at each token. In Appendix D.7, we have demonstrated the potential to leverage attribution methods to study the contribution of each token to hallucination risk. Future work could consider token or claim-level evaluation with groundtruths.

Acknowledgements

We are grateful to all anonymous reviewers for providing constructive and helpful feedback to strengthen our work. This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-005).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, establish, exploit: Red teaming language models from scratch. *arXiv preprint arXiv:2306.09442*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye.

2024. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9367–9385, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao, and Kai Chen. 2024. LLM factoscope: Uncovering LLMs' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10218–10230, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104, Miami, Florida, US. Association for Computational Linguistics.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2025. Exploring concept depth: How large language models acquire knowledge and concept at different layers? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 558–573, Abu Dhabi, UAE. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli

Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Aisha Khatun and Daniel G Brown. 2024. Trutheval: A dataset to evaluate llm truthfulness and reliability. *arXiv preprint arXiv:2406.01855*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Preprint*, arXiv:2305.19187.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Kanti V Mardia and Peter E Jupp. 2009. *Directional statistics*. John Wiley & Sons.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Mistral AI. 2024. Ministral-8B-Instruct-2410. `https://huggingface.co/mistralai/Mistral-8B-Instruct-2410`.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*.

Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, Preslav Nakov, and Timothy Baldwin. 2025. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in LLM outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35712–35731, Suzhou, China. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Andy Shih, Dorsa Sadigh, and Stefano Ermon. 2023. Long horizon temperature scaling. In *International conference on machine learning*, pages 31422–31434. PMLR.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Ponhvoan Srey, Quang Minh Nguyen, Xiaobao Wu, and Anh Tuan Luu. 2026. Towards reliable truth-aligned uncertainty estimation in large language models. *arXiv preprint arXiv:2604.00445*.

Ponhvoan Srey, Xiaobao Wu, and Anh Tuan Luu. 2025. Unsupervised hallucination detection by inspecting reasoning processes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22117–22129, Suzhou, China. Association for Computational Linguistics.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. In

Findings of the Association for Computational Linguistics: ACL 2024, pages 14379–14391, Bangkok, Thailand. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. 2025. Too consistent to detect: A study of self-consistent errors in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4755–4765, Suzhou, China. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Gleb Kuzmin, Ivan Lazichny, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. 2025a. Unconditional truthfulness: Learning unconditional uncertainty of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35673–35694, Suzhou, China. Association for Computational Linguistics.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and 1 others. 2025b. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for llms. *arXiv preprint arXiv:2505.20045*.

Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2025c. Token-level density-based uncertainty quantification methods for eliciting truthfulness of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2246–2262, Albuquerque, New Mexico. Association for Computational Linguistics.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F Wong, and Rui Wang. 2024. Latent space chain-of-embedding enables output-free llm self-evaluation. *arXiv preprint arXiv:2410.13640*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

Zhanghao Zhouyin and Ding Liu. 2025. Understanding neural networks with logarithm determinant entropy estimator. *Neurocomputing*, 657:131520.

A Theoretical Analysis

A.1 Generalised Variance Bound

Proposition 1 demonstrates that generalised variance is constrained by both sources of magnitude (radial) and directional variation. Using generalised variance alone is insufficient as it loses information about the type of dispersion. For instance, when there is little variation in one dimension, generalised variance may vanish, even though the remaining directions exhibit high variability.

Proposition 1. *Let Σ denote the sample covariance for hidden states across layers. The generalised variance $\det\Sigma$ is bounded by the radial variation $\mathrm{Var}(\|\boldsymbol{h}\|)$ and (weighted) resultant length r_w as follows:*

$$\log\det\Sigma \leq d\log\left(\frac{\mathrm{Var}(\|\boldsymbol{h}\|)+\mu_{\|\boldsymbol{h}\|}^2(1-r_w^2)}{d}\right), \tag{5}$$

where $\mu_{\|\boldsymbol{h}\|} = \frac{1}{L+1}\sum_l\|\boldsymbol{h}^l\|$ and $r_w = \|\sum_l \boldsymbol{h}^l\|/\sum_l\|\boldsymbol{h}^l\| = \|\bar{\boldsymbol{h}}\|/\mu_{\|\boldsymbol{h}\|}$, with equality if and only if the distribution is isotropic, i.e. $\Sigma = \lambda I$.

Proof. Let $\lambda_1,\ldots,\lambda_d$ be the eigenvalues of Σ. Then, $\log\det\Sigma = \sum_{i=1}^d\log\lambda_i$. Since logarithm is concave, we apply Jensen's inequality:

$$\begin{aligned}\frac{1}{d}\sum_{i=1}^d\log\lambda_i &\leq \log\left(\frac{1}{d}\sum_{i=1}^d\lambda_i\right)\\ &= \log\left(\frac{\mathrm{tr}(\Sigma)}{d}\right)\end{aligned} \tag{6}$$

Next, we decompose the total variation or trace:

$$\begin{aligned}\mathrm{tr}(\Sigma) =& \frac{1}{L+1}\sum_{l=0}^L\|\boldsymbol{h}^l-\bar{\boldsymbol{h}}\|^2\\ =& \left(\frac{1}{L+1}\sum_{l=0}^L\|\boldsymbol{h}^l\|^2\right)-\|\bar{\boldsymbol{h}}\|^2\end{aligned} \tag{7}$$

Writing $\boldsymbol{h}^l$ as $\|\boldsymbol{h}^l\|\hat{\boldsymbol{h}}^l$, the first term is the second moment of radius, $\mu_{\|\boldsymbol{h}\|^2} = \frac{1}{L+1}\sum_{l=0}^L\|\boldsymbol{h}^l\|^2 = \mathrm{Var}(\|\boldsymbol{h}\|)+\mu_{\|\boldsymbol{h}\|}^2$. The norm of mean vector can be expressed as $\|\bar{\boldsymbol{h}}\| = \mu_{\|\boldsymbol{h}\|}r_w$. Thus,

$$\mathrm{tr}(\Sigma) = \mathrm{Var}(\|\boldsymbol{h}\|)+\mu_{\|\boldsymbol{h}\|}^2(1-r_w^2) \tag{8}$$

Substituting Equation (8) into Equation (6) yields the desired inequality. □

A.2 Circular Variance and Pairwise Similarity

Recall that circular variance at token t is defined as

$$c_t = 1-r_t = 1-\left\|\frac{1}{L+1}\sum_{l=0}^L\hat{\boldsymbol{h}}_t^l\right\| \tag{9}$$

Let the mean pairwise cosine similarity be $S_t = \frac{1}{L(L+1)}\sum_{l\neq m}\hat{\boldsymbol{h}}_t^l\cdot\hat{\boldsymbol{h}}_t^m$. The hidden states have been normalised such that $\|\hat{\boldsymbol{h}}_t^l\| = 1\ \forall l$.

$$\begin{aligned}r_t^2 &= \left\|\frac{1}{L+1}\sum_{l=0}^L\hat{\boldsymbol{h}}_t^l\right\|^2 = \frac{1}{(L+1)^2}\sum_{l,m}\hat{\boldsymbol{h}}_t^l\cdot\hat{\boldsymbol{h}}_t^m\\ &= \frac{1}{(L+1)^2}\left(\sum_{l=m}\hat{\boldsymbol{h}}_t^l\cdot\hat{\boldsymbol{h}}_t^m+\sum_{l\neq m}\hat{\boldsymbol{h}}_t^l\cdot\hat{\boldsymbol{h}}_t^m\right)\\ &= \frac{1+LS_t}{L+1}\end{aligned} \tag{10}$$

Expressing c_t in terms of S_t yields

$$c_t = 1-\sqrt{\frac{1+LS_t}{L+1}} \tag{11}$$

Circular variance accounts for the relationship between all pairs at different layers, instead of only subsequent pairs under CoE.

B Computational Cost

SIVR is far more efficient than existing sampling-based approaches as the additional computation to extract internal variance features is manageable. In particular, the determinant of the covariance is obtained without performing a large d^2 eigendecomposition. Instead, we compute the determinant from a small $(L+1)\times(L+1)$ Gram matrix using Sylvester's theorem or equivalently Singular Value Decomposition (SVD). The other terms, circular variance and predictive entropy, are efficient to compute.

In Table 4, we report the computational time with Llama-3.2-3B-Instruct on Nvidia RTX-6000 with 48 GB of GPU memory. We compare with two sampling-based methods, Semantic Entropy (SE) (Kuhn et al., 2023) and SAR (Duan et al., 2024) implemented via LM-Polygraph (Fadeeva et al., 2024), and SAPLMA. SIVR is

Method	Time (s) TriviaQA / SciQ	AUC TriviaQA / SciQ
SE	5.20 / 7.08	84.72 / 73.99
SAR	3.51 / 4.17	84.45 / 79.44
SAPLMA	0.25 / 0.35	77.94 / 83.62
SIVR	0.26 / 0.37	89.31 / 85.70

Table 4: Computational time of SIVR compared to sampling-based methods, SE and SAR, and SAPLMA.

lightweight in computation, adding modest overhead to SAPLMA, and achieves better performance.

C Experimental Details

C.1 Dataset

We evaluate on twelve datasets:

- Animals and Facts (Azaria and Mitchell, 2023): we use all 1008 and 613 statements.
- TriviaQA (Joshi et al., 2017), MedMCQA (Pal et al., 2022): we randomly sample 1000 questions from the validation set.
- SciQ (Welbl et al., 2017), MMLU (Hendrycks et al., 2020): the full validation and test set of 1531 and 1000 questions, respectively, are used.
- MGSM (Shi et al., 2022): we select queries from four languages, English, Bengali, Japanese, and Thai, each with 250 instances.
- Counterfact (Meng et al., 2022), Common Claims (Casper et al., 2023), FEVER (Thorne et al., 2018), MATH (Hendrycks et al., 2021), CommonsenseQA (Talmor et al., 2019): we randomly select 1000 instances. For FEVER, only those with labels "SUPPORTS" or "REFUTES" are chosen.

For probe-based methods, the datasets are split 80-20 for training and evaluation. Following Duan et al. (2024), for TriviaQA and SciQ, ROUGE-L (Lin, 2004) is adopted as the correctness metric, with a threshold of 0.7. Table 5 provides the prompts used.

C.2 Classifier Details

A simple transformer encoder model is used for the sequence classification task. The architecture is as shown in Figure 7. Adam optimiser is used with learning rate 10^{-4} and weight decay 10^{-5}.

Counterfact, Claims, Animals, Facts, FEVER
Determine whether the given statement is TRUE or FALSE. The Final Answer must only be either 'TRUE.' or 'FALSE.' only. `{statement}`
Let's think step-by-step.
TriviaQA, SciQ
Be concise, and output only the final answer. `{question}`
MedMCQA
Answer the following multiple choice medical question. The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of ABCD. Think step by step and output the reasoning process before answering. `{question}`
MGSM
Solve this math problem. Give the reasoning steps before giving the final answer on the last line by itself in the format of "Answer:". Do not add anything other than the integer answer after "Answer:" `{question}`
MATH
`{question}` Please reason step by step, and put your final answer within \\boxed \\
MMLU
Answer the following multiple choice question. The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of ABCD. Think step by step before answering. `{question}`
CommonsenseQA
Answer the following multiple choice common-sense reasoning question. The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of ABCDE. Think step by step and output the reasoning process before answering. `{question}`

Table 5: Prompts for each dataset.

```
TransformerClassifier(
  (input_proj): Linear(in_features=13, out_features=128, bias=True)
  (posenc): PositionalEncoding()
  (encoder): TransformerEncoder(
    (layers): ModuleList(
      (0-1): 2 x TransformerEncoderLayer(
        (self_attn): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=128, out_features=128, bias=True)
        )
        (linear1): Linear(in_features=128, out_features=256, bias=True)
        (dropout): Dropout(p=0.1, inplace=False)
        (linear2): Linear(in_features=256, out_features=128, bias=True)
        (norm1): LayerNorm((128,), eps=1e-05, elementwise_affine=True)
        (norm2): LayerNorm((128,), eps=1e-05, elementwise_affine=True)
        (dropout1): Dropout(p=0.1, inplace=False)
        (dropout2): Dropout(p=0.1, inplace=False)
      )
    )
  )
  (pool): AttnPool(
    (scorer): Sequential(
      (0): Linear(in_features=128, out_features=128, bias=True)
      (1): Tanh()
      (2): Linear(in_features=128, out_features=1, bias=True)
    )
  )
  (head): Linear(in_features=128, out_features=1, bias=True)
)
```

Figure 7: Model architecture of classifier.

D Additional Experimental Results

D.1 Qwen-3 Experiments

Table 6 shows the evaluation with Qwen-3 models of 4B and 14B parameter size. The results are in line with other model families and sizes, with SIVR consistently outperforming its competitors.

D.2 Fact-Checking Datasets

Table 10 provides the results for each dataset with explicit true-false groundtruths.

D.3 More Ablation

We report ablation for the Meta models in Table 9.

D.4 Number of Principal Components

We adopt principle component analysis to avoid overfitting to high-dimensional, model-specific hidden states and keep the supervised head lightweight. This practice is consistent with prior works, *e.g.* (Vazhentsev et al., 2025c). We analyse the dependence on the number of principal components (PC) in Table 12, and find that performance is stable with 10 PCs.

D.5 More OOD Evaluation

Figure 8, along with Appendix D.1, present the OOD change in FPR@95 and AUPR, respectively, with the Ministral-8B-Instruct. The results align with the observation with AUC as well, showing that internal variance improves robustness with respect to all three metrics.

D.6 Input Length

We provide additional analysis for SIVR with respect to the length of the generated sequence. In the table below, we bin the generated sequences into short, medium, and long. We observe that SIVR performs better on longer sequences. We believe that with longer sequences, patterns that indicate factual inaccuracy may be more pronounced.

D.7 Token Contribution to Hallucination

With the sequential classifier, we can get a fine-grained view of which tokens drive prediction by using attribution methods that assign contribution from the input. Integrated Gradients (Sundararajan et al., 2017) is a simple training-free method. Table 11 demonstrates how IG can highlight to the user which tokens contributed strongly to hallucination risk.

D.8 Failure Cases

Table 13 highlights demonstrative failure cases. From these examples, we observe the following:

- False negatives (missed hallucinations): when the error is concentrated in a single decisive token, *e.g.* an important entity, while the rest of the response is fluent and coherent, the dispersion pattern may be weak or not sharply localised, so sequence-level score does not cross hallucination threshold.
- False positives (flagging correct answers): when a correct answer contains rare entities or domain-specific terminology, dispersion can be elevated and calibration can be less reliable, leading SIVR to incorrectly flag the response as hallucination.

In summary, SIVR can miss localised factual errors, leading to false negatives, and flag rare or technical tokens as incorrect, resulting in false positives.

Method	TriviaQA			SciQ			MedMCQA			MGSM			MATH			MMLU			CommonsenseQA			Average		
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP
P(True)	79.18	84.80	89.79	49.61	99.74	66.59	55.78	91.97	39.61	59.14	93.66	17.07	55.52	92.66	18.26	49.62	96.04	70.65	43.12	94.43	17.28	59.47	84.98	37.64
Max Prob	81.29	56.42	89.30	71.29	77.30	75.51	66.37	83.63	50.67	74.82	76.08	27.47	54.43	87.88	17.24	50.40	95.34	72.25	64.79	85.14	33.08	67.35	80.78	44.26
Perplexity	82.23	56.08	89.79	71.88	77.30	76.11	66.35	84.11	50.39	74.84	75.72	27.56	54.59	87.71	17.22	50.57	95.10	72.23	64.59	84.31	32.89	67.51	77.68	48.01
Entropy	80.81	57.43	89.47	70.37	78.57	75.91	66.86	83.79	51.06	75.54	76.67	28.53	54.50	86.69	17.33	50.37	95.34	72.05	65.28	84.52	34.38	67.66	77.07	48.30
Temp-Scaled	82.30	55.41	89.13	72.65	76.53	75.95	65.96	82.99	50.14	74.14	73.33	27.49	54.52	89.08	17.19	50.69	94.87	72.41	64.09	83.90	31.76	67.12	78.10	47.60
Energy	39.78	96.96	70.11	41.48	100.00	59.92	63.48	84.59	47.89	67.51	75.00	26.70	56.32	82.25	17.93	51.21	95.10	72.54	62.17	86.17	28.86	58.95	84.01	43.15
SE	83.74	54.55	92.55	71.74	83.75	80.02	62.31	87.20	48.35	65.43	86.39	23.73	56.77	98.22	20.03	46.69	100.00	74.88	53.72	94.41	27.75	64.15	82.27	47.35
SAR	78.06	74.55	87.20	73.65	86.11	79.62	58.87	92.00	42.18	52.01	92.90	16.62	49.30	96.45	15.59	47.75	96.15	74.28	52.75	96.27	22.17	63.26	85.53	44.50
CoE-R	27.55	100.00	58.44	27.22	100.00	47.25	62.74	88.92	48.21	58.08	92.82	13.73	42.89	97.10	12.42	51.25	96.50	72.67	63.86	91.43	32.99	50.78	90.20	38.38
CoE-C	58.14	94.26	72.72	46.14	95.41	55.53	67.95	80.90	**54.26**	59.02	91.03	15.04	42.46	96.59	12.18	50.52	96.27	72.33	65.55	89.58	34.41	58.51	88.24	43.22
SATMD + MSP	89.08	64.91	95.85	84.40	73.13	91.09	68.75	85.50	50.80	**78.99**	74.39	40.56	50.91	94.52	16.69	54.41	**89.47**	72.64	64.92	89.64	39.53	70.38	77.03	48.28
Lookback Lens	87.37	57.81	94.40	81.26	86.90	88.05	**69.01**	**69.53**	47.78	56.17	**56.79**	4.01	59.09	**40.91**	11.96	46.33	95.38	72.63	**78.16**	**73.98**	45.76	71.16	67.33	48.23
SAPLMA	89.36	64.41	94.18	83.80	60.26	89.45	62.28	82.40	48.94	78.35	98.80	**56.34**	**67.97**	67.80	**39.93**	57.31	96.51	75.77	75.65	82.99	41.18	71.99	74.62	52.58
SIVR (Ours)	**93.60**	**20.00**	**96.06**	**88.99**	**52.56**	**93.04**	63.62	72.44	49.30	75.95	64.20	48.71	65.33	80.87	29.62	**58.94**	92.77	**78.48**	74.03	78.97	**46.34**	**75.40**	**62.88**	**57.25**

(a) Qwen-3-4B-Instruct

Method	TriviaQA			SciQ			MedMCQA			MGSM			MATH			MMLU			CommonsenseQA			Average		
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP
P(True)	65.14	96.49	66.20	49.70	98.40	60.60	57.23	86.80	37.11	76.23	81.49	26.33	56.78	95.58	31.64	53.64	96.10	75.54	43.95	98.03	15.01	58.01	88.20	34.27
Max Prob	81.95	59.59	82.09	60.58	86.04	63.17	66.38	82.64	47.27	68.41	93.13	24.77	63.81	85.64	38.24	49.62	92.86	73.43	66.53	88.18	32.25	71.80	79.72	50.04
Perplexity	82.12	59.59	82.45	61.12	86.04	63.83	66.46	80.75	46.57	68.86	92.24	24.88	63.98	82.87	38.63	49.48	93.51	73.36	66.62	87.93	32.09	71.96	77.58	50.15
Entropy	82.38	59.38	83.40	60.98	86.96	65.64	67.33	80.38	48.26	71.17	89.25	30.77	63.15	87.29	39.13	49.84	94.16	73.40	67.67	83.25	32.60	72.25	77.00	51.12
Temp-Scaled	81.55	62.27	80.83	61.28	86.04	62.94	65.75	83.02	45.58	66.90	94.93	22.14	64.32	81.77	38.11	49.39	94.81	73.38	65.45	91.87	30.86	69.63	91.22	47.82
Energy	66.43	96.91	72.28	50.06	100.00	61.74	69.84	71.70	50.27	78.17	71.94	31.61	61.88	81.22	35.43	49.50	94.16	73.78	67.55	84.24	33.68	51.08	90.66	35.18
SE	64.04	93.48	68.67	60.79	84.27	64.37	66.24	82.54	50.49	69.27	75.00	36.94	58.21	83.02	34.02	48.75	97.06	66.50	57.40	89.74	28.41	60.42	85.61	39.97
SAR	66.41	94.87	75.14	58.30	94.38	61.75	62.75	83.05	54.98	61.96	88.76	23.09	55.99	91.82	27.96	47.01	91.18	64.21	50.82	89.74	22.15	57.07	89.08	36.55
CoE-R	45.78	97.73	48.08	51.69	95.65	55.43	51.67	94.72	37.57	36.22	97.91	7.76	55.10	90.61	23.92	46.42	98.05	72.87	44.84	96.31	16.64	56.37	91.13	35.53
CoE-C	49.23	96.91	51.38	57.11	94.28	60.38	52.33	96.23	37.19	30.87	97.91	6.94	55.75	88.95	24.74	46.64	97.40	73.77	47.79	96.06	17.82	57.67	90.34	36.85
SATMD + MSP	82.50	69.89	85.68	77.52	78.38	87.20	63.54	79.59	47.32	**84.62**	63.08	40.48	69.26	51.35	37.55	50.37	94.29	72.55	72.62	83.54	**46.92**	73.17	71.93	52.81
Lookback Lens	79.96	71.03	81.41	**85.38**	53.19	**88.06**	**75.57**	**65.71**	**59.93**	65.28	94.44	22.84	64.40	71.01	38.83	38.49	91.43	69.78	**75.01**	75.25	44.15	75.48	68.56	59.50
SAPLMA	84.06	62.89	84.18	81.46	59.77	84.60	69.59	81.13	54.39	76.68	54.41	20.68	65.20	83.78	35.28	49.27	**90.32**	73.55	65.09	70.73	22.41	77.17	61.44	57.44
SIVR (Ours)	**90.90**	**35.87**	**91.54**	84.70	**49.44**	87.75	64.71	82.35	49.13	79.95	**53.62**	**43.22**	**77.35**	**46.15**	**41.61**	**54.07**	96.30	**81.65**	70.32	**61.18**	25.98	**80.62**	**51.59**	**65.13**

(b) Qwen-3-14B

Table 6: Qwen-3 performance.

Dataset	Short	Medium	Long	Overall
TriviaQA	84.74	83.19	99.43	89.31
SciQ	83.19	78.89	94.54	85.70

Table 7: AUC with Llama-3.2-3B-Instruct for varying input lengths.

Test On	Trivia	SciQ	Med	MGSM	MATH	MMLU	CSQA	Avg	Test FPR@95
Lookback Ratios	4.87	1.3	3.18	21.4	6.2	-0.55	6.16	6.08	91.76
Hidden States	36.22	17.45	8.34	34.6	20.79	4.87	10.16	18.92	92.26
Internal Variance	29.33	29.45	-6.29	27.48	14.09	3.66	2.33	14.29	85.33
SIVR	32.05	23.84	-3.62	32.82	18.39	4.27	8.33	16.58	87.35

(a) Average change in FPR@95.

Test On	Trivia	SciQ	Med	MGSM	MATH	MMLU	CSQA	Avg	Test AUPR
Lookback Ratios	-22.64	-22.15	-7.93	1.87	-19.27	-3.01	-8.93	-11.72	53.63
Hidden States	-40.2	-25.22	-16.12	-28.23	-25.94	-4.0	-7.67	-21.06	50.16
Internal Variance	-9.76	-12.38	-9.67	-22.57	-16.37	-1.57	-6.11	-11.21	59.94
SIVR	-20.19	-12.72	-4.5	-30.65	-25.22	-4.44	-11.07	-15.54	57.07

(b) Average change in AUPR.

Table 8: Summary of performance with OOD data.

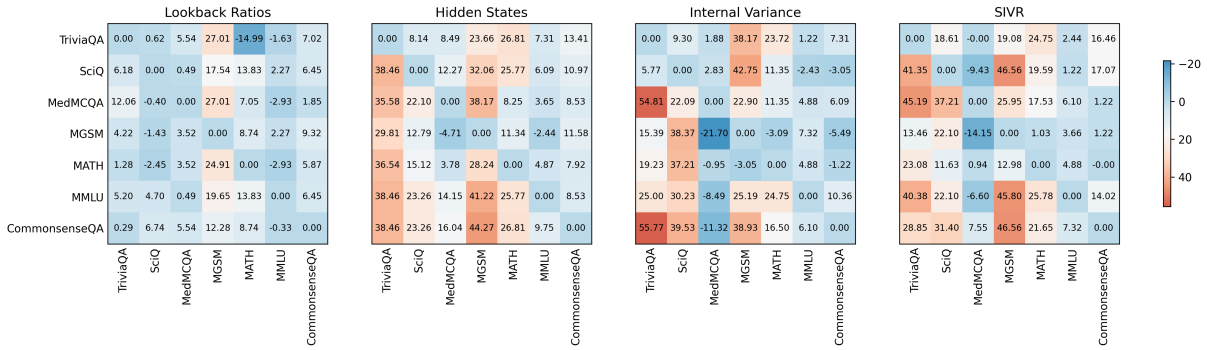
Variant	True-False			Fever			TriviaQA			SciQ			MedMCQA			GSM			Math			MMLU			CommonsenseQA			Average		
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP
HS (PCA)	86.24	52.01	72.53	85.62	37.69	70.17	80.00	60.27	88.50	77.08	83.05	90.38	56.73	93.33	53.31	65.26	87.64	63.29	75.94	67.86	84.70	51.51	93.67	72.93	62.84	81.71	50.13	74.99	67.77	71.96
HS	89.84	33.13	79.10	85.34	38.46	70.39	83.22	57.53	89.51	76.87	83.05	90.32	62.74	86.67	56.22	65.71	88.76	63.80	80.59	67.86	87.27	55.81	93.67	77.99	67.45	79.88	53.93	78.09	60.70	75.49
CovDet, AngVar	81.61	53.93	64.14	81.23	56.15	68.87	89.56	34.25	93.78	88.80	38.98	94.95	66.84	76.19	62.89	71.53	77.53	65.74	77.19	75.00	84.68	55.86	93.67	76.80	67.10	72.56	48.19	77.05	61.67	71.04
Internal Variance	81.80	58.19	65.79	84.32	50.77	70.34	89.40	39.73	93.75	88.77	40.68	94.99	67.78	76.19	64.37	72.89	75.28	67.44	78.10	78.57	85.44	57.01	96.20	78.64	66.61	81.10	47.71	77.67	64.27	72.15
Full (RNN)	83.88	52.55	66.52	73.78	60.00	50.27	86.52	49.32	90.56	84.89	57.63	93.15	60.28	88.57	56.29	66.46	78.65	56.83	76.81	62.50	82.97	46.95	96.20	73.19	64.82	76.83	46.50	74.67	64.99	67.99
SIVR	90.81	26.03	79.14	86.95	46.72	76.20	89.31	41.56	93.67	85.70	52.54	93.68	66.59	72.82	63.68	75.65	74.42	69.23	83.48	42.11	86.00	50.89	84.42	74.97	66.57	90.53	48.59	79.53	55.46	76.53

(a) Llama-3.2-3B-Instruct

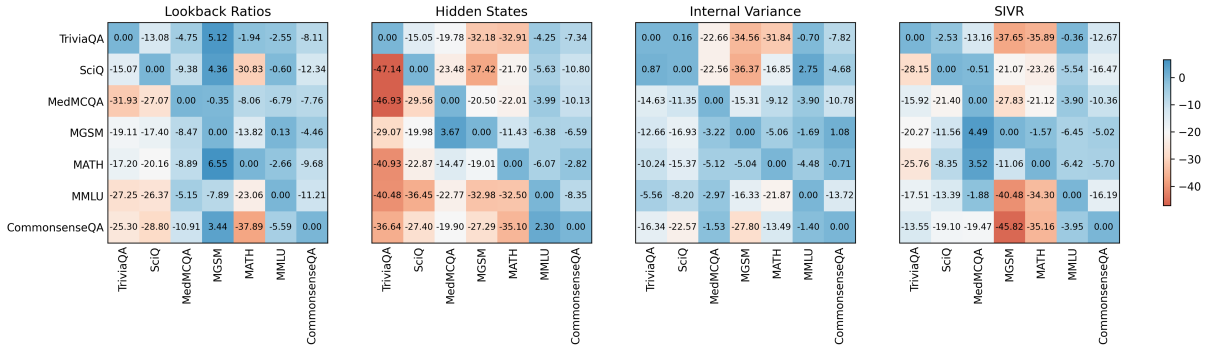
Variant	True-False			Fever			TriviaQA			SciQ			MedMCQA			GSM			Math			MMLU			CommonsenseQA			Average		
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP
HS (PCA)	76.98	54.49	45.04	81.71	47.47	45.79	83.74	69.16	84.91	85.50	50.00	92.76	66.70	71.31	52.82	70.08	71.97	44.47	74.79	73.30	70.58	57.10	87.32	80.85	72.00	61.33	46.40	74.96	62.49	58.23
HS	83.64	38.00	49.86	85.40	55.70	58.08	85.06	58.88	83.73	86.06	53.12	93.33	68.03	77.87	53.54	81.57	50.76	61.70	79.72	62.64	74.78	52.37	92.96	76.02	72.71	64.09	46.34	78.79	55.67	62.25
CovDet, AngVar	74.27	63.44	42.74	85.64	51.90	53.72	83.91	62.62	84.89	89.17	42.19	95.00	67.99	92.62	55.27	73.37	75.00	51.36	75.72	65.93	70.82	49.91	92.96	74.01	65.74	78.45	41.28	74.04	67.95	58.11
Internal Variance	77.17	52.83	42.57	86.08	36.71	51.60	83.46	64.49	84.33	90.80	37.50	95.81	65.84	84.43	54.45	74.43	75.00	52.65	74.34	83.52	71.96	54.21	91.55	77.45	65.72	77.90	40.86	75.30	63.54	58.28
Full (RNN)	76.86	61.50	45.77	78.57	52.53	32.53	86.51	57.01	86.54	88.82	43.75	94.48	69.76	92.31	66.78	71.26	90.91	54.01	69.76	92.31	66.78	54.40	92.96	76.83	58.46	96.69	29.74	73.75	72.04	57.56
SIVR	83.88	47.95	54.98	81.10	61.44	46.54	89.48	39.81	89.27	87.79	54.55	94.14	75.95	67.24	59.57	89.04	17.54	58.13	79.05	67.57	72.30	64.07	91.30	88.60	68.32	73.13	50.28	80.86	55.37	64.89

(b) Llama-3.1-8B-Instruct

Table 9: Ablation with Llama models.



(a) OOD FPR@95 change.



(b) OOD AUPR change.

Figure 8: More OOD performance with Ministral-8B-Instruct.

Method	Counterfact			Common Claim			Animals			Facts			FEVER		
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP
P(True)	49.85	78.36	33.85	71.00	47.56	61.45	54.09	72.55	24.31	60.07	67.19	19.12	66.09	70.04	38.86
Max Prob	67.16	85.03	52.46	67.92	69.46	64.32	69.36	92.83	41.36	85.03	54.37	57.45	66.80	68.88	39.81
Perplexity	66.66	85.19	51.72	67.47	69.05	63.07	69.03	92.83	40.66	84.48	54.56	54.16	66.40	69.45	39.21
Entropy	67.87	84.77	53.31	67.09	73.08	63.53	70.12	87.38	39.55	87.59	46.63	61.86	67.50	66.86	39.91
Temp-Scaled	65.40	85.11	50.55	67.45	67.54	62.87	66.95	92.70	40.29	81.40	57.54	49.41	65.48	69.88	38.81
Energy	72.50	75.04	58.67	71.23	71.07	67.72	67.47	77.69	37.14	87.28	49.60	64.57	68.76	70.32	41.54
SE	75.55	90.48	62.99	71.08	70.06	69.45	66.63	78.38	48.44	85.40	44.65	64.54	58.17	83.44	33.20
SAR	70.53	90.48	51.64	72.40	82.00	76.32	64.97	87.84	39.56	84.74	76.73	62.18	60.51	82.78	29.03
CoE-R	42.40	95.01	33.28	63.32	87.90	60.14	47.21	96.28	23.87	70.99	79.17	31.30	44.98	94.67	27.12
CoE-C	64.23	86.29	50.51	53.02	91.94	53.20	55.03	98.14	29.84	76.34	77.78	45.51	65.51	82.71	40.52
SATMD + MSP	70.89	79.66	58.10	66.50	65.59	65.27	75.82	94.08	57.68	82.82	41.75	48.34	69.30	78.42	46.75
Lookback Lens	80.39	62.50	64.53	79.13	52.38	71.28	78.32	63.76	52.87	92.84	**9.90**	80.26	73.17	88.67	42.79
SAPLMA	77.45	65.83	68.24	79.58	47.96	76.95	71.80	74.17	43.71	92.71	36.63	80.17	80.38	57.97	57.29
TAD	80.06	53.44	63.79	**86.16**	**35.48**	84.78	87.90	42.21	**70.93**	**93.61**	20.00	**88.63**	85.94	**37.76**	64.64
SIVR (Ours)	**84.12**	**46.78**	**71.75**	85.40	52.69	**86.70**	**88.13**	**30.32**	69.15	91.51	30.61	84.72	**86.95**	46.72	**76.20**

(a) Llama-3.2-3B-Instruct

Method	Counterfact			Common Claim			Animals			Facts			FEVER		
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP
P(True)	65.48	65.83	33.15	76.18	63.01	59.62	48.34	94.13	17.66	48.44	86.49	4.03	65.87	80.76	23.47
Max Prob	66.32	66.11	32.75	54.75	88.18	43.14	62.34	75.75	25.31	79.43	51.75	13.50	70.39	76.71	26.31
Perplexity	66.45	66.11	33.55	55.10	88.18	43.87	62.89	75.88	26.40	79.93	52.81	14.77	70.89	76.71	27.56
Entropy	69.80	67.09	35.56	55.59	86.82	44.25	66.72	71.00	27.01	80.05	51.23	13.64	71.50	77.09	26.26
Temp-Scaled	65.13	69.19	32.67	55.13	88.87	43.92	60.77	79.00	25.33	78.44	55.61	13.89	70.52	78.61	28.27
Energy	77.87	70.31	45.78	66.87	86.82	52.38	69.40	79.63	27.43	58.89	97.54	9.46	74.41	86.46	31.26
SE	71.04	66.67	44.62	56.15	90.48	42.09	71.18	82.56	27.76	80.79	38.02	18.11	70.66	83.72	27.62
SAR	70.91	63.13	37.44	52.67	77.42	42.08	64.17	78.79	27.26	73.44	54.69	11.33	66.26	89.53	20.98
CoE-R	56.88	93.98	33.74	49.23	90.92	38.55	40.91	99.63	14.20	29.07	100.00	3.06	50.77	94.43	20.66
CoE-C	42.57	98.32	20.60	45.83	92.64	35.82	52.53	92.63	22.67	72.81	76.67	10.07	53.51	91.65	19.21
SATMD + MSP	79.30	58.52	56.62	78.35	53.51	65.41	66.07	89.57	35.63	82.93	51.33	19.60	77.50	59.63	31.59
Lookback Lens	70.91	64.83	35.19	73.79	77.36	69.88	76.39	73.72	51.12	79.01	91.15	33.41	75.13	73.29	34.32
SAPLMA	**85.92**	56.64	**70.59**	80.62	60.68	**74.71**	77.64	55.00	41.91	**93.39**	**20.00**	**50.24**	71.46	82.91	52.06
TAD	80.79	63.31	60.38	78.30	65.04	64.77	76.44	63.13	37.87	73.68	47.37	25.08	**84.88**	**57.62**	**60.52**
SIVR (Ours)	84.17	**55.24**	63.02	**83.47**	**44.35**	73.06	**85.43**	**50.62**	**57.68**	82.43	41.59	26.14	81.10	61.44	46.54

(b) Llama-3.1-8B-Instruct

Method	Counterfact			Common Claim			Animals			Facts			FEVER		
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP
P(True)	46.65	93.38	23.88	47.99	92.15	33.54	50.14	96.83	15.91	51.88	90.18	7.16	48.98	96.77	18.67
Max Prob	63.35	90.93	39.54	71.21	83.84	55.87	67.19	71.24	24.05	68.61	76.53	31.93	64.28	89.33	31.01
Perplexity	67.72	87.86	42.35	72.03	77.21	55.90	68.04	71.95	26.27	71.56	67.78	33.03	68.05	82.63	33.80
Entropy	66.89	91.47	40.49	69.84	94.51	54.61	71.35	79.93	32.14	65.53	89.49	25.13	67.45	91.56	32.96
Temp-Scaled	65.11	86.02	39.63	71.06	71.27	54.66	65.04	77.46	22.35	72.83	64.10	34.92	66.43	78.16	31.35
Energy	69.06	74.35	40.76	66.15	83.31	50.62	71.69	75.70	35.12	78.82	55.52	26.43	67.54	78.29	30.82
SE	66.88	77.50	44.04	70.90	72.73	55.60	67.32	82.42	36.30	78.61	65.56	41.80	66.06	79.88	27.01
SAR	70.74	62.35	30.17	60.09	81.82	40.25	77.69	78.18	45.57	77.08	61.67	29.55	76.39	61.18	43.79
CoE-R	45.44	96.86	25.10	51.43	98.17	39.27	33.69	99.41	11.10	45.79	92.64	14.50	48.24	96.77	20.49
CoE-C	49.48	96.18	27.21	48.25	96.65	37.08	48.20	93.66	14.91	58.02	93.87	8.79	47.78	93.18	18.92
SATMD + MSP	72.34	67.55	43.12	77.86	49.62	58.84	79.97	99.41	60.92	74.42	80.19	47.76	66.47	65.62	34.56
Lookback Lens	70.01	89.47	43.46	79.61	65.67	66.49	74.37	79.12	27.42	90.49	29.63	**63.77**	77.93	68.32	41.33
SAPLMA	80.32	60.75	**61.48**	81.74	51.91	65.89	82.25	58.48	56.74	**92.88**	**20.18**	52.53	64.36	80.86	28.13
TAD	77.54	68.21	50.93	**85.62**	55.97	70.67	**84.69**	67.05	**65.95**	89.13	32.17	45.67	76.94	69.46	36.08
SIVR (Ours)	**82.73**	**54.30**	57.86	84.24	**47.88**	**72.58**	84.24	**51.72**	49.36	90.25	26.32	50.34	**82.44**	**54.09**	**55.71**

(c) Ministral-8B-Instruct

Method	Counterfact			Common Claim			Animals			Facts			FEVER		
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP
P(True)	60.03	82.08	24.33	75.72	55.21	54.36	60.95	85.09	23.66	67.24	65.19	7.21	57.76	78.87	22.85
Max Prob	65.31	87.76	30.94	60.63	90.22	45.21	73.40	68.72	38.64	78.85	69.28	18.93	66.57	91.64	31.91
Perplexity	65.73	87.13	31.76	60.57	89.59	45.45	73.26	69.34	38.69	78.97	63.48	18.22	66.52	91.25	31.62
Entropy	66.20	85.24	31.72	61.29	89.75	45.91	73.89	64.90	38.31	80.32	59.04	19.67	66.44	91.38	31.41
Temp-Scaled	64.73	87.13	30.52	60.05	88.96	44.73	72.61	72.04	38.64	77.18	71.67	17.24	66.48	90.24	31.72
Energy	62.68	82.15	28.70	58.84	89.59	43.05	65.36	79.68	30.53	75.33	72.87	14.29	63.27	94.55	30.80
SE	65.74	79.88	25.62	62.47	89.47	44.11	66.21	85.53	42.30	63.06	76.80	6.35	71.91	77.99	37.18
SAR	68.75	**57.14**	23.91	65.66	80.45	44.66	69.76	97.70	36.41	68.38	72.68	16.79	74.14	98.77	38.76
CoE-R	50.40	94.95	21.26	50.31	95.74	36.39	60.31	95.32	26.50	59.61	81.40	7.26	55.17	93.66	24.15
CoE-C	54.46	93.44	24.85	56.82	94.16	42.20	67.29	89.66	33.35	71.79	84.81	19.70	61.91	90.11	30.13
SATMD + MSP	71.81	69.30	34.82	57.67	80.99	47.80	75.28	85.45	36.65	**81.40**	**35.54**	18.89	66.89	81.53	33.98
Lookback Lens	71.80	85.71	32.51	72.17	64.91	60.07	82.11	56.97	50.07	77.10	48.74	31.16	73.28	70.32	40.35
SAPLMA	**73.68**	64.04	**39.16**	**82.95**	**34.65**	**67.41**	62.12	92.02	29.27	56.44	80.51	5.69	73.98	71.07	43.59
SIVR (Ours)	71.10	69.39	32.11	74.26	46.62	47.92	**85.66**	**40.74**	**59.51**	68.95	94.02	**56.07**	**84.41**	**41.92**	**49.81**

(d) Qwen-3-4B-Instruct

Method	Counterfact			Common Claim			Animals			Facts			FEVER		
	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP	AUC	F95	AUP
P(True)	51.36	75.41	19.26	66.04	69.06	41.02	43.06	96.21	12.94	68.30	86.05	5.14	64.70	78.82	20.45
Max Prob	78.04	71.88	42.97	68.96	85.92	47.88	79.01	80.47	40.94	96.69	46.55	62.30	81.66	83.79	45.21
Perplexity	78.04	71.50	43.04	69.03	85.92	48.02	79.09	80.47	41.16	96.91	38.66	62.36	81.82	71.48	45.37
Entropy	77.93	70.24	43.04	68.90	87.68	48.03	78.99	76.69	41.24	96.83	39.16	62.49	81.86	69.59	45.39
Temp-Scaled	72.87	100.00	38.64	65.42	100.00	43.90	70.88	100.00	34.68	93.45	100.00	60.41	78.35	100.00	42.39
Energy	27.51	98.87	13.78	26.44	98.68	21.92	43.99	92.07	13.63	37.82	98.82	2.72	33.81	99.29	11.30
SE	61.57	80.00	27.23	55.10	92.96	36.25	68.15	87.58	37.99	51.08	76.41	5.70	64.49	85.23	23.08
SAR	57.05	83.13	20.84	54.76	95.71	38.07	66.75	91.30	30.88	48.31	71.79	2.72	54.76	93.18	16.83
CoE-R	65.70	84.87	32.71	70.58	80.65	51.47	66.52	90.77	30.06	69.40	83.53	19.06	72.46	82.72	30.79
CoE-C	66.25	79.45	32.72	69.25	82.99	48.80	69.29	89.70	32.84	74.71	80.34	25.48	72.78	83.91	30.17
SATMD + MSP	72.12	72.15	42.01	71.32	63.83	45.18	73.36	83.93	31.71	84.58	60.17	56.95	76.28	63.01	40.16
Lookback Lens	83.02	58.13	61.62	84.74	**58.46**	75.53	87.23	39.64	59.44	83.48	82.05	62.03	83.26	62.42	50.42
SAPLMA	**87.62**	**37.11**	**65.80**	82.72	75.74	73.93	89.76	39.64	57.89	90.34	30.25	58.27	84.19	51.48	**58.29**
SIVR (Ours)	87.19	37.34	65.13	**85.00**	71.88	**77.54**	**90.50**	**27.56**	**75.67**	**98.53**	**5.88**	**84.09**	**86.41**	**25.57**	41.43

(e) Qwen-3-14B

Table 10: Results for individual true-false topics.

Question

Which score is used for wound infection? A: Glasgow coma scale, B: Southampton score, C: Apgar score, D: SIRS score.

Correct Answer: B

Response

The Glasgow Coma Scale is used to assess the level of consciousness in patients, not for wound infection. The Southampton score is used for assessing the severity of sepsis, not specifically for wound infection. The Apgar score is used to assess the health of newborns at birth, not for wound infection. The Systemic Inflammatory Response Syndrome (SIRS) score is used to identify patients with systemic inflammatory response syndrome, which can be associated with wound infection. Therefore, the correct answer is:

Answer: D

Question

Without supervision, who is allowed to give dentures? A: Denturist, B: Hygienist, C: School Dental nurse, D: Dental therapist.

Correct Answer: A

Response

Denturist : They are trained to make and fit dentures, but typically require supervision for complex cases.

Hygienist : They are trained in oral hygiene and preventive care, but not in making or fitting dentures.

School Dental Nurse: They provide basic dental care in schools, but not denture fitting.

Dental Therapist: They are trained to provide a range of dental services, including making and fitting dentures, and can do so without supervision. Therefore, the correct answer is:

Answer: D.

Table 11: Qualitative examples. Red implies contribution to hallucination risk.

# PC	2	5	10	20	30
Average AUC	78.60	80.02	79.29	79.68	79.24
Average FPR@95	60.01	57.35	58.46	61.68	57.89
Average AUPR	73.63	74.38	74.06	74.10	72.91

Table 12: Performance with varying number of principal components (#PC).

Failure Type	Query	Response	$\hat{p}(y=1)$	Comment
False Negative	Which Gilbert and Sullivan operetta is sub-titled 'The Witches Curse'?	The Mikado </s>	.251	Detector wrongly focused on a special token.
	All of the following are the complications in the new born of a diabetic mother except A. Hyperbilirubinemia, B. Hyperglycemia, C. Hypocalcemia, D. Hypomagnesemia	To determine the correct answer, let 's consider the complications that are typically associated with newborns of diabetic mothers: ... Hyperglycemia: This is a common complication in newborns of diabetic mothers due to the mother 's high blood sugar levels . ... Hypomagnesemia : This is a condition where the newborn has low magnesium levels , which can also be a complication in newborn s of diabetic mothers . Given this information, the complication that is not typically associated with newborns of diabetic mothers is: Answer : D </s>	.467	Detector did identify the final answer "D" as hallucinatory. However, sequence level score is dominated by contributions from other tokens, like semantically less important ones ("let's consider"), wrongly obscuring the predicted probability of hallucination.
False Positive	Under which British monarch was Warren Hastings appointed first Governor of the Presidency of Fort William (Bengal) and the head of the Supreme Council of Bengal, effectively the first Governor-General of India?	Warren Hastings was appointed by King George III .	.536	Detector focused on appropriate segment, but does not recognise it as hallucination.
	In female adrenal gland secretes which hormone? A. Progesterone, B. Testosterone, C. Estrogen, D. DHEA	The adrenal glands are responsible for producing several hormones, including cortisol, aldosterone, and androgens. Among the options given , DHEA (Dehydroepiandrosterone) is an androgen produced by the adrenal glands . It is a precursor to both testosterone and estrogen. Answer: D	.915	Detector flags technical terminology as hallucinatory, potentially due to limited training data.

Table 13: Failure cases.