INTRYGUE: Induction-Aware Entropy Gating for Reliable RAG Uncertainty Estimation

Anonymous ACL submission

Abstract

While retrieval-augmented generation (RAG) significantly improves the factual reliability of LLMs, it does not eliminate hallucinations, so robust uncertainty quantification (UQ) remains essential. In this paper, we reveal that standard entropy-based UQ methods often fail in RAG settings due to a mechanistic paradox. An internal "tug-of-war" inherent to context utilization appears: while induction heads promote grounded responses by copying the correct answer, they collaterally trigger the previously established "entropy neurons." This interaction inflates predictive entropy, causing the model to signal false uncertainty on accurate outputs. To address this, we propose INTRYGUE (Induction-Aware Entropy Gating for Uncertainty Estimation), a mechanistically grounded method that gates predictive entropy based on the activation patterns of induction heads. Evaluated across four RAG benchmarks and six open-source LLMs (4B to 13B parameters), INTRYGUE consistently matches or outperforms a wide range of UQ baselines. Our findings demonstrate that hallucination detection in RAG benefits from combining predictive uncertainty with interpretable, internal signals of context utilization[1].

1 Introduction

Large language models (LLMs) are widely deployed (Xie et al., 2024; Amorese et al., 2023), but their static memory is often insufficient. This is typically addressed with retrieval-augmented generation (RAG), which improves reliability by grounding output in external documents (Fan et al., 2024; Lewis et al., 2020). Nonetheless, RAG does not fully eliminate LLM hallucinations (Zhang et al., 2025). As ungrounded errors limit the use of LLMs in high-stakes environments, reliable methods for detecting these failures remain essential.

[1]Our code is available at the following link https://anonymous.4open.science/r/tda4hallucinations-1B39/README.md

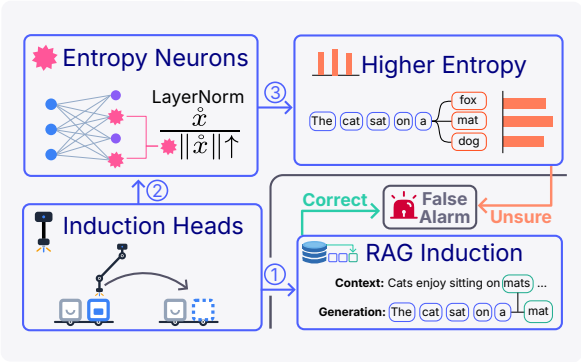


Figure 1: The mechanistic tug-of-war in RAG uncertainty quantification. ① During generation, Induction Heads exhibit intense attention activations, locking onto specific patterns in the retrieved context to promote relevant tokens into the output. This mechanism sharpens the logit distribution toward the correct answer, increasing entropy-based model confidence. However, ② it simultaneously activates Entropy Neurons. These neurons inflate the LayerNorm denominator, ③ artificially increasing the predictive entropy. This dual activation results in a False Alarm, where correct, context-grounded outputs are incorrectly flagged as highly uncertain.

Although many hallucination detection methods have been proposed (Alansari and Luqman, 2025), they frequently face practical limitations, such as relying on external databases (Li et al., 2024) or introducing substantial computational overhead via complex prompting and multi-step verification pipelines (Es et al., 2024; Friel and Sanyal, 2023). To avoid these bottlenecks, we approach hallucination detection through uncertainty quantification (UQ) lens, directly exploiting the LLM's internal confidence scores to gauge the reliability of its outputs.

While numerous UQ methods have been developed for closed-book settings (Kuhn et al., 2023; Qiu and Miikkulainen, 2024; Vazhentsev et al., 2025), there is only a very limited amount of work on UQ for the RAG scenario. Recent studies note that standard UQ techniques are often poorly suited

for RAG, as their design does not incorporate specific RAG-setup properties, e.g. prompt-response semantic alignment (Soudani et al., 2025). To address this gap, the existing approaches inject the document–response relationship into the uncertainty score; however, because these solutions rely on external models or multiple model generations, their applicability remains limited (Soudani et al., 2025; Perez-Beltrachini and Lapata, 2025). Moreover, the internal model mechanisms causing poor performance of UQ in RAG remain underexplored.

In this work, we explore the mechanistic limitation that entropy-based uncertainty estimation (Fomicheva et al., 2020) — a widely used baseline for hallucination detection (Nikitin et al., 2024; Fadeeva et al., 2025; Qiao et al., 2026) — faces in RAG scenarios. We demonstrate that correct RAG-based responses are associated with the activation of induction heads, which are attention mechanisms that copy and propagate retrieved information (Elhage et al., 2021). While these heads promote context-aligned responses and increase model confidence, they simultaneously trigger an opposing mechanism: entropy neurons (Stolfo et al., 2024). These specialized units modulate predictive entropy and conversely act to reduce the LLM's confidence. As a result, this internal tug-of-war makes entropy-based scores less reliable estimators of model uncertainty (Figure 1).

To address this issue, we introduce INTRYGUE (Induction-Aware Entropy Gating), a mechanistically grounded method that adjusts entropy scores based on the activation strength of induction heads. INTRYGUE accounts for induction-driven entropy inflation and improves entropy's effectiveness as a hallucination signal in the RAG settings.

Our contributions:

- We demonstrate that the correctness of LLM responses in RAG settings is strongly associated with the activation of the model's induction heads, which significantly *increase* both the probability of the correct answer and overall model confidence;
- We validate a connection between induction heads and entropy neurons, previously identified only in controlled induction settings involving exact token repetition and short naturally occurring n-gram repetitions. For real RAG-based decoding, induction heads activate entropy neurons, thereby not deliberately *decreasing* LLM confidence;
- Consequently, the process of generating accurate RAG responses is shown to inherently involve two interconnected but competing mechanisms, which pull LLM confidence scores in opposite directions, resulting in unreliable output entropy-based uncertainty estimation;
- To address this conflict, we propose **INTRYGUE**, a novel method that gates entropy using induction head attention sinks, effectively correcting induction-driven entropy inflation in RAG settings;
- Our **INTRYGUE** is demonstrated to consistently outperform or perform on par with a wide range of uncertainty estimation baselines across four RAG benchmarks and six models, ranging from 4B to 13B parameters in size.

2 Background

2.1 Induction Heads

Induction heads, which fundamentally drive in-context learning in Transformers, were formalized in Olsson et al. (2022). They implement a strict sequence-copying heuristic: given a token pattern $(t_A, t_B, \ldots, t_A)$, the head attends strongly to t_B to predict it as the subsequent token. To empirically identify them, one evaluates the model on a randomly generated, repeated sequence $S = (s_1, \ldots, s_L, s_1, \ldots, s_L)$ and computes an *induction score*. This score is defined as the average attention mass allocated to the correct historical offsets of S:

$$\text{InductionScore}(A^{(l,h)}) = \frac{1}{L} \sum_{j=1}^{L} \alpha^{(l,h)}_{L+j,1+j}, \quad (1)$$

where $A^{(l,h)}$ is the attention matrix of head h at layer l, and $\alpha^{(l,h)}_{t,k}$ denotes the attention weight from the current position t to the past position k.

2.2 Entropy Neurons

In this paper, the term *neurons* refers to individual entries of the pre-activation hidden states of the MLP sublayer in the final transformer block:

$$\text{MLP}(\mathbf{x}) = \mathbf{W}_{\text{out}}\sigma(\mathbf{W}_{\text{in}}\mathbf{x} + \boldsymbol{\beta}_{\text{in}}) + \boldsymbol{\beta}_{\text{out}}, \quad (2)$$

where $\mathbf{x} \in \mathbb{R}^{d_{\text{model}}}$ is a normalized residual stream hidden state, $\mathbf{W}^{\top}_{\text{out}}, \mathbf{W}_{\text{in}} \in \mathbb{R}^{d_{\text{mlp}} \times d_{\text{model}}}$ are learned weight matrices, $\boldsymbol{\beta}_{\text{in}}, \boldsymbol{\beta}_{\text{out}} \in \mathbb{R}^{d_{\text{mlp}}}$ are learned bias vectors, and σ is an activation function. The output of the final transformer block is

then passed through layer normalization:

$$\mathrm{LN}(\mathbf{x}) = \frac{\mathbf{x} - m(\mathbf{x})}{\sqrt{\mathrm{Var}(\mathbf{x}) + \epsilon}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}, \qquad (3)$$

where $m(\mathbf{x})$ and $\mathrm{Var}(\mathbf{x})$ denote the mean and variance of the entries of $\mathbf{x}$, and $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{d_{\text{model}}}$ are learned biases. The resulting vector is then projected onto the vocabulary logits by the unembedding matrix $\mathbf{W}_{\mathrm{U}} \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$, with $\mathcal{V}$ being a dictionary size, followed by a softmax to obtain a probability distribution over tokens.

Entropy neurons (Gurnee et al., 2024) are characterized by a large norm of their corresponding column $\mathbf{w}_{\text{out}}$ of $\mathbf{W}_{\text{out}}$ and a near-uniform shift to the logits, quantified by the LogitVar score:

$$\mathrm{LogitVar}(\mathbf{w}_{\text{out}}) = \mathrm{Var}\left(\frac{\mathbf{W}_{\mathrm{U}}\mathbf{w}_{\text{out}}}{\|\mathbf{W}_{\mathrm{U}}\|_{\text{dim}=1}\|\mathbf{w}_{\text{out}}\|}\right). \qquad (4)$$

The large norm, in turn, affects the LayerNorm scaling, squeezing the logit distribution and reducing model confidence, similarly to temperature scaling.

The authors of Stolfo et al. (2024) have hinted that induction heads activate entropy neurons, but validation of this fact was limited to controlled induction settings with synthetic exact token repetition and short naturally occurring n-gram repetitions of fixed length leaving open whether this mechanism generalizes to broader contexts where the model must reconcile retrieved external knowledge with its own parametric memory.

3 Empirical Study: Entropy-based Uncertainty Estimation Underperforms in RAG

In this section, we investigate why entropy-based uncertainty estimates tend to underperform in RAG settings. Our evidence suggests that this failure may be linked to the dual role played by induction heads, which we analyze in detail below.

3.1 Entropy-based Uncertainty Estimation

Although entropy-based methods are a widely adopted baseline in hallucination detection (Nikitin et al., 2024; Fadeeva et al., 2025; Qiao et al., 2026), their effectiveness in RAG settings remains limited. To illustrate this, we evaluate the performance of the Maximum Entropy (Fomicheva et al., 2020) on the MS MARCO subset of the RAGTruth dataset (Niu et al., 2024) and XSum (Narayan et al., 2018). The results are shown in Figure 2. Although offering a weak signal regarding generation errors,

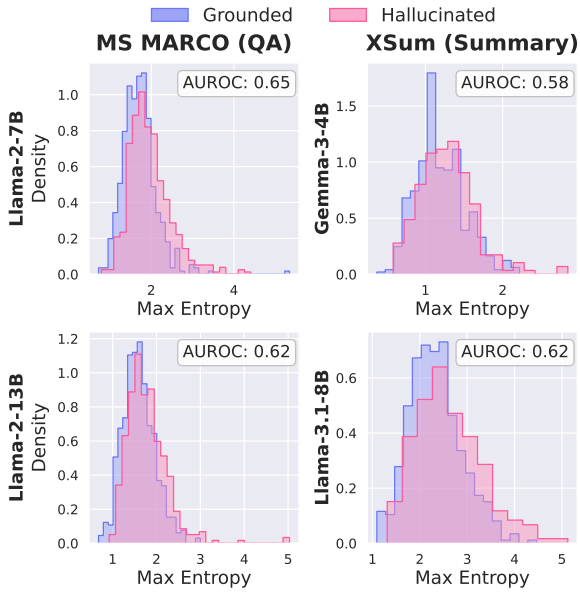


Figure 2: Maximum Entropy score distributions for hallucinated (pink) and grounded (blue) samples heavily overlap, demonstrating that Maximum Entropy fails to reliably separate hallucinated from grounded LLM responses.

the distributions of scores for hallucinated and non-hallucinated responses overlap heavily, demonstrating that entropy-based metrics provide unreliable uncertainty estimates in RAG settings.

3.2 Dual Role of Induction Heads

Induction heads play a dual role in RAG response generation: they both boost and lower the model's confidence in the correct response. On one hand, induction heads are responsible for copying relevant tokens from the context, directly contributing to the generation of faithful responses, increasing confidence. On the other hand, they activate entropy neurons, which inflate the entropy of the output distribution, lowering the model's confidence. This dual role may help explain why entropy-based uncertainty estimates tend to be unreliable in RAG settings.

3.2.1 Induction Heads Pull Confidence Up...

Correlation. To investigate whether induction head activation is predictive of hallucination in RAG-based responses, we compute the average SinkRate for the top-5 induction heads[2] and evaluate its ability to discriminate between hallucinated and grounded responses using the AUROC score. For an attention matrix $A \in \mathbb{R}^{N \times N}$, where N is the

[2]We compute the InductionScore (1) for all heads and select the 5 heads with the greatest scores.

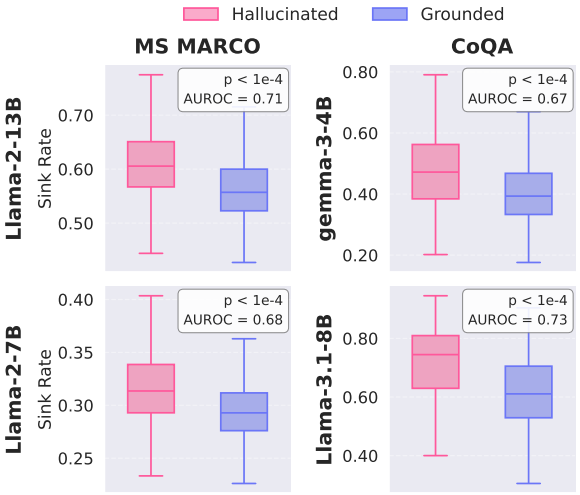


Figure 3: SinkRate scores distributions: hallucinated (pink) vs grounded (blue). AUROC scores and p-values of the Mann-Whitney U-tests are provided in the legend.

total sequence length, the SinkRate[3] is defined as:

$$\text{SinkRate}(A, n) = \max_j \frac{\sum_{i=N-n}^{N} \alpha_{i,j}}{w_j}, \quad (5)$$

where n is the response length and $w_j = \min(n, N - j + 1)$ is a normalization factor that accounts for the number of response tokens able to attend to position j. A greater SinkRate indicates deactivation of the induction head, suggesting that induction has not occurred (Elhage et al., 2021).

As shown in Figure 3, hallucinated samples consistently exhibit higher SinkRate across all four settings, with statistically significant differences ($p < 10^{-4}$, Mann-Whitney U-test) and AUROC scores ranging from 0.63 to 0.74, implying that the deactivation of induction heads may serve as a potential signal of hallucination in RAG settings.

Causality. To establish a causal link between induction head activation and correct answer generation, we ran a mean ablation experiment on the CoQA dataset (Reddy et al., 2019). Mean ablation replaces a head's output with its mean activation computed over a reference dataset, effectively removing the head's contribution to the forward pass. For correctly answered examples with the greatest induction head activation (SinkRate below the 0.33 quantile), we ablated the top-10 induction heads and compared the effect to ablating 10 randomly chosen heads.

[3]This definition is motivated by our empirical observation that attention frequently sinks to utility tokens (e.g., punctuation marks and brackets) that are not necessarily located at the beginning of the prompt (see Appendix B).

Figure 4a)-c) shows that ablating induction heads sharply increases loss and entropy, drastically increasing uncertainty compared to randomly chosen heads. This asymmetry indicates that induction heads play a causal role in context retrieval, rather than just correlating with it, and contribute substantially to accurate RAG generation.

3.2.2 ...and Push It Back Down

Correlation. To validate the connection between entropy neurons and induction heads in RAG settings, we examine the correlations between the average SinkRate (over the top-10 induction heads) and the activation scores of the top-15 entropy neurons[4], where the activation score refers to the maximum activation value of a neuron taken across the tokens of the generated response.

Figure 4d)-e) shows that for both Qwen3-8B and Llama-3.1-8B, most entropy neurons have a negative Spearman correlation with the SinkRate. This supports the idea that induction heads drive entropy neuron excitation, confirming that the findings of Stolfo et al. (2024) translate from controlled setups to real scenarios.

Causality. To establish the causal influence of induction heads on entropy neurons, we performed mean ablation of the top-k ($k = 4$ for Gemma; $k = 1$ for Llama) induction heads and measured the resulting change in entropy neuron activations via the l_2-norm — a natural choice given that entropy neurons act by scaling the LayerNorm denominator (3). As a control, we repeated the same procedure for k randomly selected heads.

Figure 5a) shows that ablating induction head consistently decreases the l_2-norm of entropy neuron activations for Llama-3.1-8B and Mistral-7B, while this observation doesn't stand for random heads. This confirms a causal link between induction head activity and entropy neuron excitation.

Impact on entropy scores. We further investigated whether identified entropy neurons causally influence model entropy-based uncertainty by performing targeted ablation experiments on the CoQA dataset. We compared the effect of mean-ablating the top-30 entropy neurons with the ablation of 50 randomly selected neurons within the same MLP layer. Figure 5b) illustrates that ablating entropy neurons yields a substantial increase

[4]We compute the LogitVar (4) for all neurons in the MLP layer of the last transformer block and select the 15 neurons with the lowest scores.

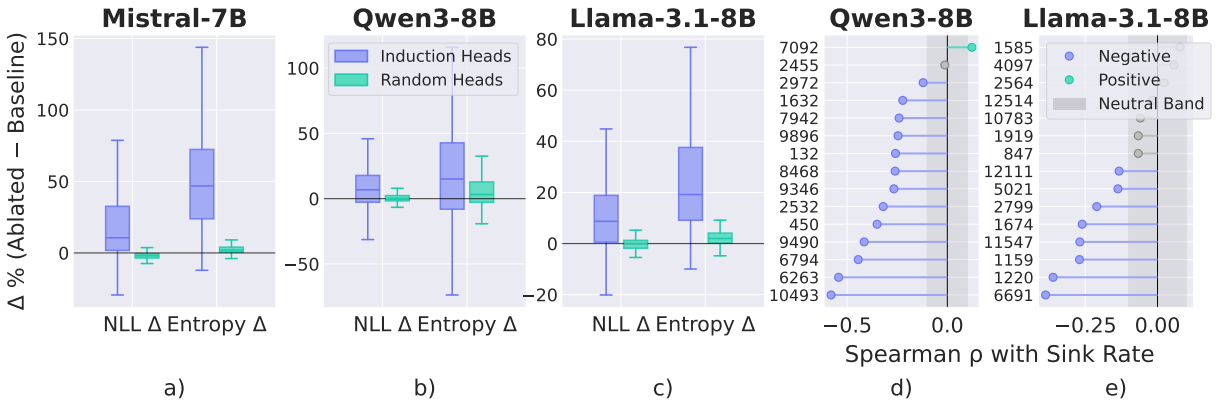


Figure 4: (a)–(c) Percentage change in Negative Log-Likelihood (NLL) and Entropy following mean ablation of induction heads (blue) versus random heads (green). Ablating induction heads results in significantly higher uncertainty and loss across all models. (d)–(e) Spearman rank correlation (ρ) between the Sink Rate of induction heads and the activations of the top-15 neurons with the lowest LogitVar scores; this indicates that induction head behavior is associated with the activity of these neurons.

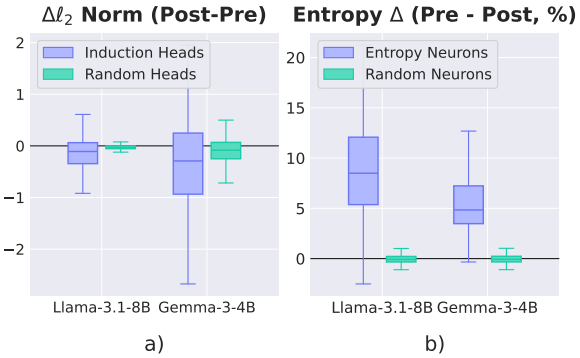


Figure 5: (a) Drop in entropy neuron activations ($\Delta\ell_2$ Post − Pre) when ablating induction heads (blue), confirming causal dependence. (b) Change in entropy (Pre − Post) after ablating entropy neurons; the positive values indicate that activations of these neurons inflate model confidence.

in generation entropy, while random ablation produces no significant change. These findings confirm that these specific neurons actively modulate the model's predictive confidence.

4 Mechanistic Uncertainty Quantification

Our findings indicate that standard entropy-based estimators overlook the controversial impact of induction heads activation, a mechanism critical for faithful RAG generation. To address this, we propose **INTRYGUE** (Induction-Aware Entropy Gating for Uncertainty Estimation). Our method uses aggregated SinkRate (5), which quantifies activity of induction heads, as a gating term; it modulates the raw entropy score, amplifying the estimated uncertainty when the model fails to mechanistically attend to the retrieved context.

Let P denote the input prompt and $R = (r_1, \ldots, r_n)$ be the generated response sequence of tokens. At each generation step t, the model outputs a probability distribution over the vocabulary $\mathcal{V}$. We derive the sequence $\mathcal{E} = \{e_t\}_{t=1}^{n}$ by calculating the entropy of these distributions to quantify the uncertainty of each predicted token:

$$e_t = -\sum_{v \in \mathcal{V}} P(v \mid P, R_{<t}) \log P(v \mid P, R_{<t}). \tag{6}$$

Simultaneously, we extract a set of SinkRates $\mathcal{S} = \{s_i\}_{i=1}^{k}$ defined in Eq. 5 corresponding to top-k induction heads with greatest InductionScore (1) values. These rates serve as a mechanistic proxy for quantifying the extent of induction during the generation process. The hyperparameter k is selected on a validation set.

Let f and g be aggregation functions (e.g., mean or max) that map these sequences to scalars. We employ the aggregated SinkRate as a coefficient that gates the predictive entropy: when induction fails (high SinkRate), the uncertainty signal is amplified; when induction is successful (low SinkRate), the signal is dampened. We thus formulate our score, INTRYGUE, as:

$$\text{INTRYGUE}(P, R) = \underbrace{f\left(\{s_i\}_{i=1}^{k}\right)}_{\substack{\text{Induction}\\ \text{Activity}}} \cdot \underbrace{g\left(\{e_t\}_{t=1}^{n}\right)}_{\substack{\text{Predictive}\\ \text{Uncertainty}}}. \tag{7}$$

In practice, the choice of aggregation functions f and g allows the metric to be tuned to capture either the average uncertainty or the most severe failure points across the generated response.

	MS MARCO			CNN/DM		
	Mistral-7B	Llama-2-7B	Llama-2-13B	Mistral-7B	Llama-2-7B	Llama-2-13B
INTRYGUE$_{\text{min-max}}$	**0.77** $\pm$ **0.03**	**0.72** $\pm$ **0.04**	**0.73** $\pm$ **0.04**	**0.69** $\pm$ **0.03**	0.60 $\pm$ 0.03	**0.60** $\pm$ **0.08**
INTRYGUE$_{\text{mean}}$	0.71 $\pm$ 0.06	0.67 $\pm$ 0.02	0.67 $\pm$ 0.04	0.65 $\pm$ 0.03	**0.62** $\pm$ **0.08**	0.58 $\pm$ 0.03
Attention Score	0.49 $\pm$ 0.05	0.47 $\pm$ 0.07	0.51 $\pm$ 0.05	0.57 $\pm$ 0.07	0.52 $\pm$ 0.03	0.51 $\pm$ 0.06
Focus	0.67 $\pm$ 0.06	0.63 $\pm$ 0.03	0.61 $\pm$ 0.03	0.62 $\pm$ 0.04	0.51 $\pm$ 0.09	0.50 $\pm$ 0.06
LN-Entropy	0.65 $\pm$ 0.06	0.63 $\pm$ 0.03	0.62 $\pm$ 0.04	0.64 $\pm$ 0.04	**0.62** $\pm$ **0.09**	0.56 $\pm$ 0.05
LSC	0.68 $\pm$ 0.04	0.57 $\pm$ 0.03	0.57 $\pm$ 0.05	0.58 $\pm$ 0.05	0.49 $\pm$ 0.07	0.52 $\pm$ 0.07
MaxEntropy	0.67 $\pm$ 0.03	0.65 $\pm$ 0.06	0.62 $\pm$ 0.05	0.61 $\pm$ 0.04	0.60 $\pm$ 0.04	0.54 $\pm$ 0.09
MaxProb	0.75 $\pm$ 0.04	0.69 $\pm$ 0.03	0.67 $\pm$ 0.04	0.61 $\pm$ 0.07	0.51 $\pm$ 0.07	0.53 $\pm$ 0.08
Perplexity	0.64 $\pm$ 0.05	0.48 $\pm$ 0.03	0.46 $\pm$ 0.03	0.49 $\pm$ 0.08	0.52 $\pm$ 0.04	0.48 $\pm$ 0.04
RAUQ	0.56 $\pm$ 0.06	0.53 $\pm$ 0.06	0.54 $\pm$ 0.04	0.49 $\pm$ 0.08	0.52 $\pm$ 0.07	0.52 $\pm$ 0.05
ReDeEP	0.72 $\pm$ 0.03	0.63 $\pm$ 0.05	0.66 $\pm$ 0.03	0.56 $\pm$ 0.06	0.57 $\pm$ 0.06	0.52 $\pm$ 0.03
EigenScore	0.60 $\pm$ 0.04	0.58 $\pm$ 0.03	0.58 $\pm$ 0.04	0.63 $\pm$ 0.08	0.59 $\pm$ 0.03	0.57 $\pm$ 0.08
SAR	0.67 $\pm$ 0.04	0.55 $\pm$ 0.02	0.48 $\pm$ 0.03	0.57 $\pm$ 0.04	0.48 $\pm$ 0.10	0.52 $\pm$ 0.05
Semantic Density	0.55 $\pm$ 0.02	0.55 $\pm$ 0.05	0.53 $\pm$ 0.02	0.56 $\pm$ 0.08	0.52 $\pm$ 0.04	0.57 $\pm$ 0.07
Semantic Entropy	0.60 $\pm$ 0.05	0.53 $\pm$ 0.04	0.62 $\pm$ 0.04	0.54 $\pm$ 0.04	0.50 $\pm$ 0.06	0.54 $\pm$ 0.03

Table 1: AUROC scores for hallucination detection on subsamples of RAGTruth for QA (MS MARCO) and Summarization (CNN/DM). Bold and underline indicate the best and second-best results, respectively. All results are averaged over 5 runs.

5 Results

5.1 Experiment setting

Datasets and models. We evaluate on four question answering (QA) and summarization tasks: long-form QA (MS MARCO) and summarization (CNN/DM) from RAGTruth (Niu et al., 2024), conversational QA (CoQA (Reddy et al., 2019)), and extreme summarization (XSum (Narayan et al., 2018)). Additional details are in Appendix C.

We evaluate six popular open-source LLMs that provide access to their internal states: gemma-3-4b-it (Team, 2025a), LLaMA-2-7B-chat, LLaMA-2-13B-chat (Touvron et al., 2023), Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), LLaMA-3.1-8B-Instruct (Weerawardhena et al., 2025), and Qwen3-8B (Team, 2025b). As annotated responses for gemma-3-4b-it, LLaMA-3.1-8B, and Qwen3-8B are not available in the RAGTruth dataset, we limit the evaluation of these models to CoQA and XSum.

Baselines. We compare our method against a comprehensive set of baselines, including information-based: AttentionScore (Sriramanan et al., 2024), Focus (Zhang et al., 2023), Length-Normalized Entropy (Malinin and Gales), Lowest Span Confidence (LSC) (Qiao et al., 2026), MaxEntropy (Fomicheva et al., 2020), MaxProb (Fomicheva et al., 2020), Perplexity (Fomicheva et al., 2020), and RAUQ (Vazhentsev et al., 2025); sampling diversity based: EigenScore (Chen et al.), SAR (Duan et al., 2024), Semantic Density (Qiu and Miikkulainen, 2024), and Semantic Entropy (Kuhn et al., 2023); mechanistic intepretability-based: ReDeEP (Sun et al., 2025). Implementation details are available in Appendix D.

5.2 Main results

The main experimental results are presented in Tables 1-2. We report two implementations of our proposed method: INTRYGUE$_{\text{min-max}}$ (where f and g are the sequence minimum and maximum, respectively) and INTRYGUE$_{\text{mean}}$ (where both are the average).

As shown in Tables 1 and 2, INTRYGUE consistently matches or outperforms both attention-based (e.g., RAUQ, Attention Score) and logit-based baselines (e.g., MaxProb, LN-Entropy) across all evaluated benchmarks. This suggests that integrating distinct uncertainty signals, internal induction dynamics and predictive entropy, enables more robust uncertainty quantification. Importantly, the optimal aggregation variant depends heavily on response length. For the long, detailed generations of RAGTruth (MS MARCO, CNN/DM) and XSum, averaging over the sequence blurs localized hallucination signals. Consequently, INTRYGUE$_{\text{min-max}}$ dominates these settings, achieving the top score in 9 out of 12 corresponding configurations. Conversely, INTRYGUE$_{\text{mean}}$ proves most effective for CoQA with responses consisting of only a few tokens and averaging is more appropriate.

Illustrative study. Figures 6b)-c) compare MaxEntropy and mean SinkRate predictions on CoQA (thresholds for both methods are selected on the same validation set to maximize accuracy). Quadrants show sample counts in bold and subset accuracies percentages. Notably, the bottom-left quadrant

	CoQA					
	Mistral-7B	Llama-2-7B	Llama-2-13B	Llama-3.1-8B	Qwen3-8B	gemma-3-4b
INTRYGUE$_{\text{min-max}}$	0.79 ± 0.04	0.76 ± 0.04	0.72 ± 0.05	0.78 ± 0.08	0.77 ± 0.07	0.73 ± 0.05
INTRYGUE$_{\text{mean}}$	0.82 ± 0.04	**0.79 ± 0.03**	**0.81 ± 0.03**	0.81 ± 0.07	**0.80 ± 0.06**	**0.76 ± 0.04**
Attention Score	0.55 ± 0.03	0.54 ± 0.06	0.60 ± 0.06	0.56 ± 0.04	0.58 ± 0.09	0.51 ± 0.10
Focus	0.53 ± 0.05	0.33 ± 0.02	0.31 ± 0.03	0.53 ± 0.09	0.42 ± 0.04	0.48 ± 0.03
LN-Entropy	**0.83 ± 0.03**	0.74 ± 0.03	0.80 ± 0.04	0.81 ± 0.08	0.79 ± 0.05	0.75 ± 0.04
LSC	0.72 ± 0.03	0.70 ± 0.03	0.55 ± 0.05	0.77 ± 0.07	0.75 ± 0.06	0.58 ± 0.09
MaxEntropy	0.78 ± 0.03	0.60 ± 0.03	0.65 ± 0.06	0.75 ± 0.06	0.73 ± 0.05	0.70 ± 0.07
MaxProb	0.72 ± 0.05	0.54 ± 0.03	0.60 ± 0.08	0.68 ± 0.09	0.50 ± 0.09	0.58 ± 0.06
Perplexity	0.59 ± 0.04	0.52 ± 0.05	0.58 ± 0.10	0.50 ± 0.08	0.43 ± 0.07	0.51 ± 0.04
RAUQ	0.40 ± 0.02	0.41 ± 0.05	0.35 ± 0.02	0.45 ± 0.06	0.36 ± 0.04	0.41 ± 0.05
ReDeEP	0.62 ± 0.04	0.67 ± 0.06	0.77 ± 0.03	0.67 ± 0.05	0.59 ± 0.01	0.59 ± 0.05
EigenScore	0.69 ± 0.05	0.55 ± 0.05	0.60 ± 0.03	0.73 ± 0.05	0.70 ± 0.05	0.71 ± 0.05
SAR	0.72 ± 0.02	0.68 ± 0.04	0.76 ± 0.06	0.64 ± 0.06	0.63 ± 0.07	0.58 ± 0.06
Semantic Density	0.61 ± 0.03	0.59 ± 0.05	0.63 ± 0.04	0.59 ± 0.08	0.63 ± 0.06	0.58 ± 0.08
Semantic Entropy	0.79 ± 0.03	0.64 ± 0.05	0.72 ± 0.05	**0.82 ± 0.06**	0.72 ± 0.05	0.70 ± 0.05
	XSum					
INTRYGUE$_{\text{min-max}}$	**0.73 ± 0.07**	0.66 ± 0.07	**0.63 ± 0.05**	**0.66 ± 0.02**	0.68 ± 0.06	**0.62 ± 0.06**
INTRYGUE$_{\text{mean}}$	0.70 ± 0.03	**0.70 ± 0.03**	0.62 ± 0.03	0.64 ± 0.03	0.66 ± 0.04	0.58 ± 0.06
Attention Score	0.59 ± 0.03	0.50 ± 0.08	0.56 ± 0.04	0.51 ± 0.05	0.46 ± 0.06	0.50 ± 0.02
Focus	0.63 ± 0.08	0.65 ± 0.06	0.58 ± 0.04	0.62 ± 0.05	0.61 ± 0.03	0.61 ± 0.04
LN-Entropy	0.69 ± 0.03	0.61 ± 0.02	0.62 ± 0.04	0.64 ± 0.03	0.65 ± 0.04	0.57 ± 0.06
LSC	0.64 ± 0.05	0.57 ± 0.05	0.58 ± 0.04	0.59 ± 0.08	0.62 ± 0.04	0.53 ± 0.07
MaxEntropy	0.67 ± 0.02	0.60 ± 0.06	0.57 ± 0.07	0.62 ± 0.02	0.67 ± 0.06	0.58 ± 0.06
MaxProb	0.60 ± 0.07	0.58 ± 0.05	0.58 ± 0.04	0.62 ± 0.04	0.63 ± 0.04	0.53 ± 0.05
Perplexity	0.58 ± 0.05	0.56 ± 0.05	0.53 ± 0.04	0.61 ± 0.05	0.63 ± 0.03	0.47 ± 0.07
RAUQ	0.46 ± 0.06	0.43 ± 0.04	0.47 ± 0.08	0.46 ± 0.07	0.43 ± 0.04	0.50 ± 0.01
ReDeEP	0.62 ± 0.04	0.61 ± 0.04	0.60 ± 0.04	0.63 ± 0.02	**0.69 ± 0.05**	**0.62 ± 0.06**
EigenScore	0.60 ± 0.08	0.60 ± 0.05	0.52 ± 0.03	**0.66 ± 0.04**	0.62 ± 0.06	0.61 ± 0.08
SAR	0.66 ± 0.03	0.60 ± 0.03	0.58 ± 0.05	0.63 ± 0.04	0.64 ± 0.03	0.50 ± 0.05
Semantic Density	0.63 ± 0.03	0.63 ± 0.06	0.59 ± 0.05	0.55 ± 0.06	0.60 ± 0.07	0.58 ± 0.06
Semantic Entropy	0.64 ± 0.05	0.63 ± 0.03	0.62 ± 0.02	0.61 ± 0.05	0.67 ± 0.04	**0.62 ± 0.06**

Table 2: Hallucination detection results (AUROC score) on CoQA and XSum datasets. Bold and underline indicate the best and second-best results, respectively. All results are averaged over 5 runs.

reveals that when Entropy predicts "hallucination" but SinkRate predicts "grounded", Entropy significantly underperforms (e.g., 8.1% accuracy for Qwen). This false uncertainty is likely caused by induction head dynamics (Section 3). INTRYGUE effectively mitigates this by lowering entropy scores wherever the SinkRate is low. While the reverse scenario (top-right quadrant) also occurs, it affects a smaller fraction of the dataset (8–10% for both models).

5.3 Ablation studies

Gating functions We tested whether applying non-linear transformations (e.g., Tanh, Softsign) to the aggregated SinkRates could improve the INTRYGUE score (7). The results in Figure 7 (and Appendix B) demonstrate that these adjustments offered no significant benefit over using the raw values directly. This confirms that our method works effectively without the need for added complexity.

INTRYGUE$_{\text{min-max}}$ vs. isolated signals Table 4 (Appendix B) compares our method to isolated mechanistic (SinkRate) and uncertainty (MaxEntropy) signals. As individual metric performance varies by dataset, single-signal detection proves unreliable. By fusing both features, INTRYGUE consistently outperforms the strongest baselines, demonstrating that these signals are insufficient alone but highly effective combined.

Number of induction heads Figure 6a demonstrates the stability of INTRYGUE$_{\text{mean}}$ across different values of the number of induction heads k. Despite high run-to-run variance, the average AUROC remains stable. Tables 1-2 confirm this variance being inherent evaluation noise shared by all baselines. Thus, INTRYGUE can be employed without hyperparameter tuning.

Efficiency To highlight the efficiency of our method, we compare its computational complexity with that of the Entropy and Semantic Entropy baselines in Figure 8. INTRYGUE achieves a runtime comparable to LN-Entropy, while remaining substantially faster than the sampling-based Semantic Entropy. Similar runtimes are observed for other information-based baselines, including

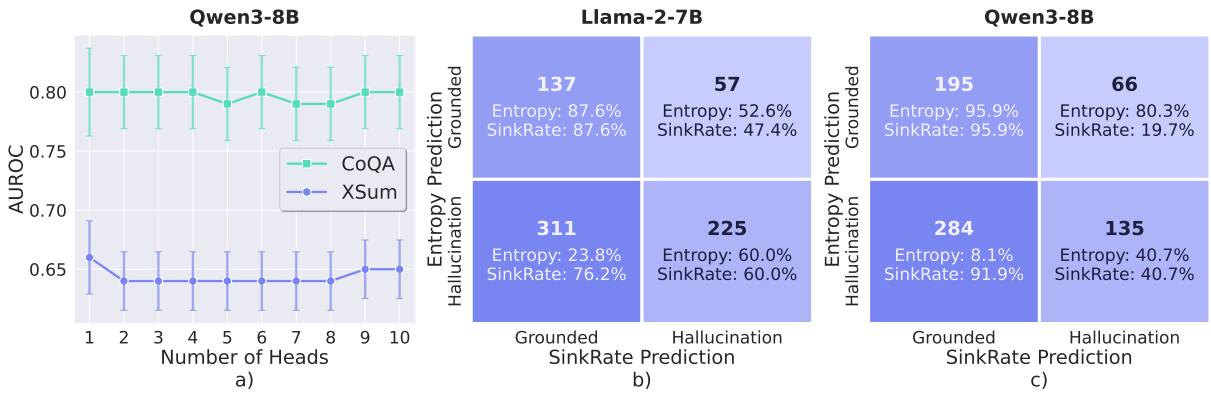


Figure 6: (a) Performance stability of INTRYGUE$_{\text{mean}}$ across different values of k (number of induction heads). (b-c) LN-Entropy vs. SinkRate prediction accuracy on CoQA. The bottom-left quadrant demonstrates SinkRate's ability to correct LN-Entropy errors, a mechanism directly exploited by INTRYGUE.

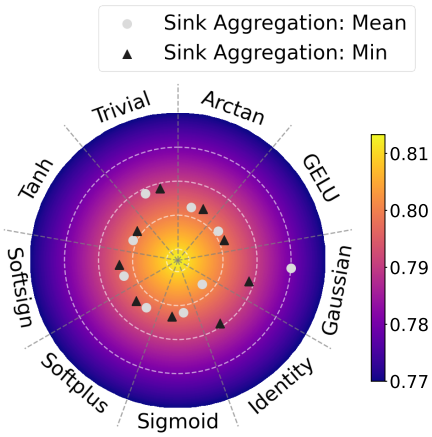


Figure 7: Impact of various non-linear transformations (f) on INTRYGUE performance. Higher AUROC scores are mapped closer to the center. Applying non-linear transformations does not allow for significant improvement in performance.

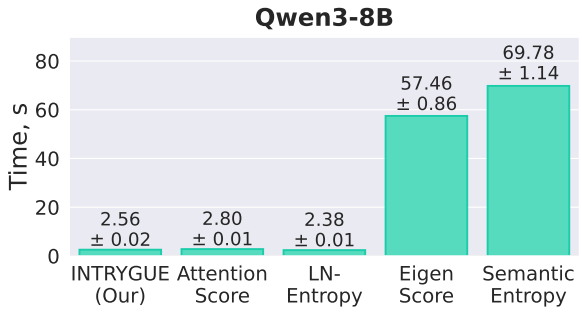


Figure 8: Computational complexity. Runtime in seconds of our method compared to information- and sampling-based baselines. Results are obtained using 20 samples from the XSum dataset with the average over ten runs presented.

RAUQ (2.74±0.03s) and Perplexity (2.31±0.01s), with differences within a few tenths of a second.

6 Conclusion

In this work, we investigated the mechanistic limitations of entropy-based uncertainty estimators in RAG scenarios. Previous evidence suggested that for highly controlled settings induction heads can activate entropy neurons; our analysis verifies that this exact mechanism occurs in a realistic RAG. While induction heads are crucial for copying relevant information from context and supporting grounded generation, their concurrent triggering of entropy neurons inflates predictive entropy, suppressing model confidence. Thus, a standard entropy-based score is unreliable with common cases of inappropriately high uncertainty to context-grounded answers.

To resolve this issue, our INTRYGUE, a training-free uncertainty estimation method, combines predictive entropy with a mechanistic score of induction head activity. By gating entropy with an estimate of induction strength, we compensate for induction-driven entropy inflation to yield a more reliable hallucination score. Experiments across four benchmarks spanning question answering and summarization, evaluated on six open-source LLMs, demonstrate that our method consistently performs on par with or better than a broad set of baselines.

Ultimately, we demonstrate that RAG uncertainty quantification must account for output distributions and the internal mechanisms of evidence utilization. The findings motivate a promising future direction: integration of mechanistic interpretability with uncertainty estimation for building more accurate and explainable methods.

Limitations

While INTRYGUE provides a mechanistically grounded approach to uncertainty estimation in RAG, we identify the following limitations:

- **Requirement of White-Box Access:** INTRYGUE relies on extracting internal attention matrices during the forward pass to compute the SinkRate and monitor induction heads. Consequently, it cannot be applied to proprietary, black-box large language models restricted behind APIs like GPT-4 or Claude.
- **Transformer Architecture:** Because our findings are fundamentally rooted in standard Transformer attention mechanics, the method's applicability to alternative architectures (such as state-space models) remains an open question.
- **Task-Dependent Aggregation:** As demonstrated in our results from Section 5.2, the optimal sequence aggregation strategy depends on the generation length. While $\text{INTRYGUE}_{\text{min-max}}$ is highly effective for long-form generation (MS MARCO, CNN/DM), $\text{INTRYGUE}_{\text{mean}}$ is required for short responses (CoQA). This prevents the score from being entirely plug-and-play across varying output lengths, requiring practitioners to select the appropriate aggregation heuristic for their specific RAG use case.
- **Faithfulness vs. Objective Factuality:** Our mechanistic proxy evaluates how successfully the model attends to and copies from the retrieved context. Therefore, INTRYGUE strictly measures faithfulness to the provided documents, not objective truth. If the upstream retrieval pipeline fetches false or conflicting information and the LLM successfully grounds its answer in that text, INTRYGUE will report low uncertainty, which users must not confuse with factual correctness.

References

Aisha Alansari and Hamzah Luqman. 2025. Large language models hallucination: A comprehensive survey. *arXiv preprint arXiv:2510.06265*.

T Amorese, C Greco, M Cuciniello, R Milo, O Sheveleva, N Glackin, and 1 others. 2023. Automatic speech recognition (asr) with whisper: Testing performances in different languages. In *CEUR WORKSHOP PROCEEDINGS*, volume 3574, pages 1–8.

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.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

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.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. Transformer Circuits Thread.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Dzianis Piatrashyn, Roman Vashurin, Shehzaad Dhuliawala, Artem Shelmanov, Timothy Baldwin, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. 2025. Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval augmented generation. *arXiv preprint arXiv:2505.21072*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-Jussà. 2024. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *arXiv preprint arXiv:2310.18344*.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research*.

Haichuan Hu, Congqing He, Xiaochen Xie, and Quanjun Zhang. 2024. Lrp4rag: Detecting hallucinations in retrieval-augmented generation via layer-wise relevance propagation. *arXiv preprint arXiv:2408.15533*.

Yihang Jiang, Xiaoyang Li, Guangxu Zhu, Hang Li, Jing Deng, Kaifeng Han, Chao Shen, Qingjiang Shi, and Rui Zhang. 2023. 6g non-terrestrial networks enabled low-altitude economy: Opportunities and challenges. *arXiv preprint arXiv:2311.09047*.

Mandar Joshi, Eunsol Choi, Daniel S 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.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zixuan Li, Jing Xiong, Fanghua Ye, Chuanyang Zheng, Xun Wu, Jianqiao Lu, Zhongwei Wan, Xiaodan Liang, Chengming Li, Zhenan Sun, and 1 others. 2024. Uncertaintyrag: Span-level uncertainty enhanced long-context modeling for retrieval-augmented generation. *arXiv preprint arXiv:2410.02719*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1797–1807.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Laura Perez-Beltrachini and Mirella Lapata. 2025. Uncertainty quantification in retrieval augmented question answering. *Transactions on Machine Learning Research*.

Yitong Qiao, Licheng Pan, Yu Mi, Lei Liu, Yue Shen, Fei Sun, and Zhixuan Chu. 2026. Lowest span confidence: A zero-shot metric for efficient and black-box hallucination detection in llms. *arXiv preprint arXiv:2601.19918*.

Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in neural information processing systems*, 37:134507–134533.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

C.H.-Wang Sky, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4401–4420.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2025. Why uncertainty estimation methods fall short in rag: An axiomatic analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16596–16616.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

Alessandro Stolfo, Ben Peng Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. 2024. Confidence regulation neurons in language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*.

Likun Tan, Kuan-Wei Huang, Joy Shi, and Kevin Wu. Interpdetect: Interpretable signals for detecting hallucinations in retrieval-augmented generation. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Gemma Team. 2025a. Gemma 3.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Sajana Weerawardhena, Paul Kassianik, Blaine Nelson, Baturay Saglam, Anu Vellore, Aman Priyanshu, Supriti Vijay, Massimo Aufiero, Arthur Goldblatt, Fraser Burch, and 1 others. 2025. Llama-3.1-foundationai-securityllm-8b-instruct technical report. *arXiv preprint arXiv:2508.01059*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, and 1 others. 2024. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3.

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.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

A Related works

UQ for Hallucination Detection in RAG scenario. While many UQ-based hallucination detection methods have been developed (Duan et al., 2024; Vazhentsev et al., 2025; Qiao et al., 2026), they have mainly been evaluated within closed-book settings (Joshi et al., 2017; Lin et al., 2022; Kwiatkowski et al., 2019; Wei et al., 2024). Existing work on uncertainty quantification in RAG is very limited, with recent works highlighting that standard UQ-based scores are unfit for RAG scenarios because their design does not account for the specific properties of a RAG setup (Soudani et al., 2025). While some methods have been proposed to address this issue, their computational overhead and external dependencies make them highly impractical (Soudani et al., 2025; Perez-Beltrachini and Lapata, 2025).

Hallucination Detection Through Mechanistic Interpretability. Mechanistic interpretability (MI) (Ferrando et al., 2024) seeks to explain how individual internal components of LLMs contribute to the final prediction. MI has previously been applied to hallucination detection in a few works: (Hu et al., 2024) proposes comparing representations of external and parametric knowledge using layer-wise relevance propagation (Bach et al., 2015), while (Sun et al., 2025; Tan et al.) suggest that hallucinations occur when feedforward layers overemphasize parametric knowledge in the residual stream, while Copying Heads (Olsson et al., 2022) fail to effectively employ external knowledge from retrieved content. However, the existing

Model	CoQA		XSum	
	Total	Hal. Rate (%)	Total	Hal. Rate (%)
gemma-3-4b-it	703	22.1	650	31.3
Qwen3-8B	698	14.6	750	24.3
Llama-3.1-8B-Instruct	691	22.1	650	28.9
Llama-2-7b-chat-hf	741	34.7	746	34.9
Llama-2-13b-chat-hf	661	23.9	730	32.3
Mistral-7B-Instruct-v0.1	1525	30.8	749	39.4

Table 3: Hallucination statistics for CoQA and XSum tasks (generated at $t = 1.0$ and annotated using GPT-4.1).

approaches rather design novel mechanistic features for hallucination detection than address the downsides of classic methods by considering their mechanistic limitations.

B Additional results

B.1 Ablation: gating functions

The additional results on applying non-linear mappings to the SinkRate are demonstrated in Figures 11 and 12. They confirm that additional transformations are uncecessary for our method.

B.2 Sink tokens

Figures 9 and 10 illustrate the sink rate patterns observed in Mistral-7B. The title of each plot lists the top three tokens with the highest sink rates, formatted as (token, position, sink rate). Notably, attention sinks not only to the initial token but also to other utility tokens (e.g., `<0x0A>` in these examples).

C Datasets

To evaluate the effectiveness of our approach across diverse RAG scenarios, we conduct experiments on four established benchmarks spanning both question answering (QA) and summarization. These datasets present varying levels of difficulty for context grounding and hallucination detection:

- **RAGTruth (MS MARCO & CNN/DM):** We utilize the long-form QA (MS MARCO) and standard summarization (CNN/DM) tasks from the RAGTruth benchmark (Niu et al., 2024). These subsets provide high-quality annotated data specifically designed for evaluating hallucinations in retrieval-augmented generation.
- **CoQA:** To test uncertainty quantification in interactive, context-dependent settings, we

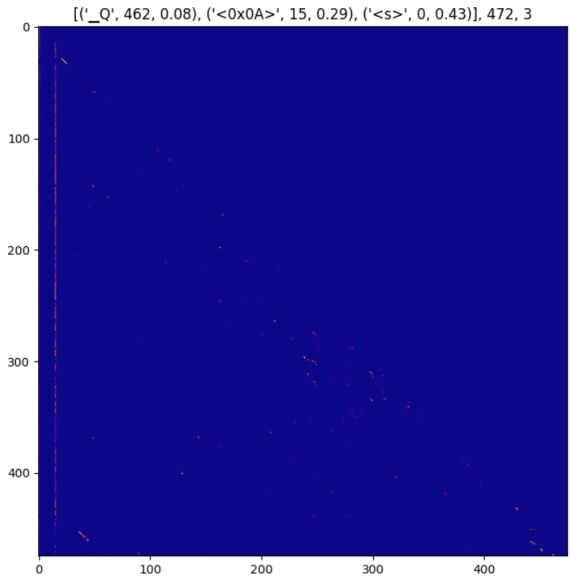


Figure 9: Attention matrix example with the sink patterns. Model: Mistral-7B, dataset: CoQA, head: L18H2.

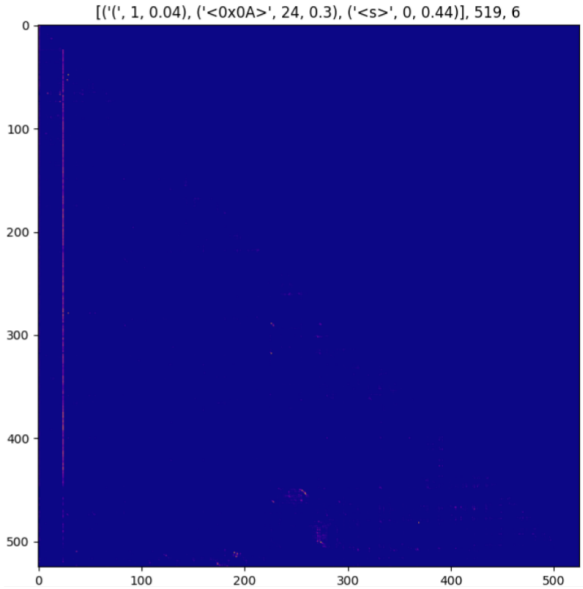


Figure 10: Attention matrix example with the sink patterns. Model: Mistral-7B, dataset: CoQA, head: L18H3.

CoQA			
Metric	Llama-2-13B	Qwen3-8B	Gemma-3-4B
SinkRate (min)	0.64 ± 0.03	0.57 ± 0.04	0.63 ± 0.04
MaxEntropy	0.65 ± 0.06	0.73 ± 0.05	0.70 ± 0.07
INTRYGUE$_{\text{min-max}}$	$\mathbf{0.72 \pm 0.05}$	$\mathbf{0.77 \pm 0.07}$	$\mathbf{0.73 \pm 0.05}$
MS MARCO			
Metric	Mistral-7B	Llama-2-7B	Llama-2-13B
SinkRate (min)	0.73 ± 0.03	0.70 ± 0.02	0.72 ± 0.04
MaxEntropy	0.67 ± 0.03	0.65 ± 0.06	0.62 ± 0.05
INTRYGUE$_{\text{min-max}}$	$\mathbf{0.77 \pm 0.03}$	$\mathbf{0.72 \pm 0.04}$	$\mathbf{0.73 \pm 0.04}$

Table 4: Performance comparison of minimum Sink Rate, maximum Entropy, and INTRYGUE across selected models on the CoQA and MS MARCO datasets. INTRYGUE outperforms both individual baseline metrics, demonstrating that combining uncertainty estimates with mechanistic signals yields superior detection compared to relying on either signal alone.

evaluate on the Conversational Question Answering (CoQA) dataset (Reddy et al., 2019), which requires models to ground responses over multi-turn dialogues. To construct our evaluation instances for this task, we generated model responses at a temperature of $t = 1.0$ and subsequently annotated them for hallucinations using GPT-4.1.

- **XSum:** Finally, we include the Extreme Summarization dataset (XSum) (Narayan et al., 2018). Because XSum requires highly abstractive, single-sentence summaries, models are notoriously prone to hallucination here, providing a rigorous stress-test for our uncertainty estimator. As with CoQA, the summaries used for evaluation were generated at $t = 1.0$ and automatically annotated using GPT-4.1.

The statistics of the CoQA and XSum datasets are provided in Table 3.

D Implementation details

In this section, we describe the key implementation choices:

- For EigenScore, we used the representation of the last token to embed sentences, as suggested by (Chen et al.). We extracted outputs from the 16th layer, as middle layers have been shown to contain the most factual information (Sky et al., 2024; Azaria and Mitchell, 2023).
- All methods based on multiple generations employed 5 additional model generations, with the exception of Semantic Density, which employed 10, as a lower number of samples failed to achieve better-than-random prediction quality.
- The hyperparameter α for the RaUQ method was set to 0.2 for the QA task and 0.0 for the summarization task, following the original paper settings.
- For a fair comparison with other baselines and our INTRYGUE score, the hyperparameters for the ReDeEP method were selected using response-wise labels (instead of token-wise labels, as proposed in the original paper).

All experiments were carried out across 5 random splits, corresponding to seeds $[42, \ldots, 46]$. The train/val/test split fractions were set to 0.4/0.4/0.2, respectively. We used NVidia L40 and NVidia H100 GPUs.

E Use of scientific artifacts

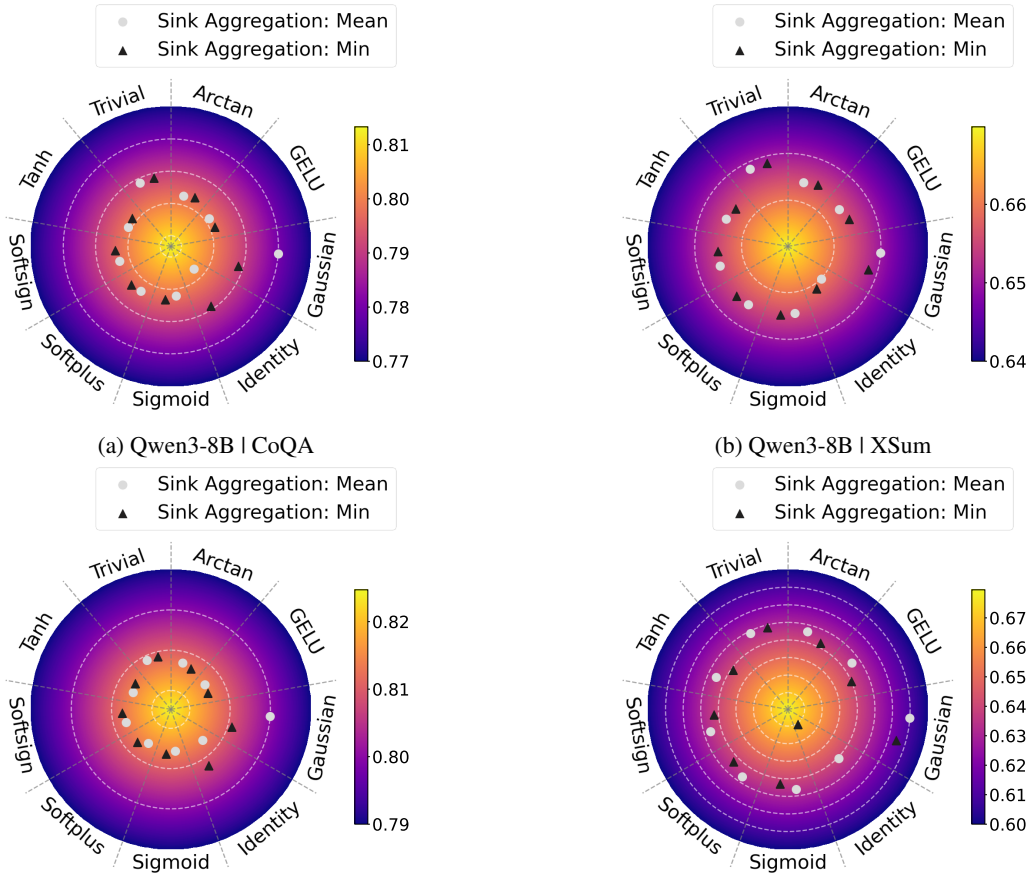
The CoQA dataset contains passages from seven domains operating under various licenses: literature and Wikipedia (CC BY-SA 4.0), MCTest children's stories (MSR-LA), RACE exam passages (custom license), and DeepMind CNN news passages (Apache). All dataset artifacts were utilized strictly following their intended licensing terms. The considered datasets contain no personally identifiable information or offensive content. Finally,

the text of this paper was proofread and refined for clarity using Gemini-3.1-Pro.

F Potential risks

Faithfulness vs. Factual Truth. INTRYGUE measures how well an output is grounded in the retrieved text, not its objective truth. If a RAG pipeline retrieves false, biased, or malicious context, the model may faithfully reproduce it. INTRYGUE will accurately report low uncertainty in these cases, which end-users could dangerously misinterpret as a guarantee of factual correctness.

Adversarial Vulnerabilities. Tying uncertainty scores to a specific mechanistic proxy introduces a theoretical attack vector. Adversaries could craft prompt injections specifically designed to artificially spike induction head activity, potentially tricking the gating mechanism into heavily suppressing the uncertainty score for manipulated outputs.



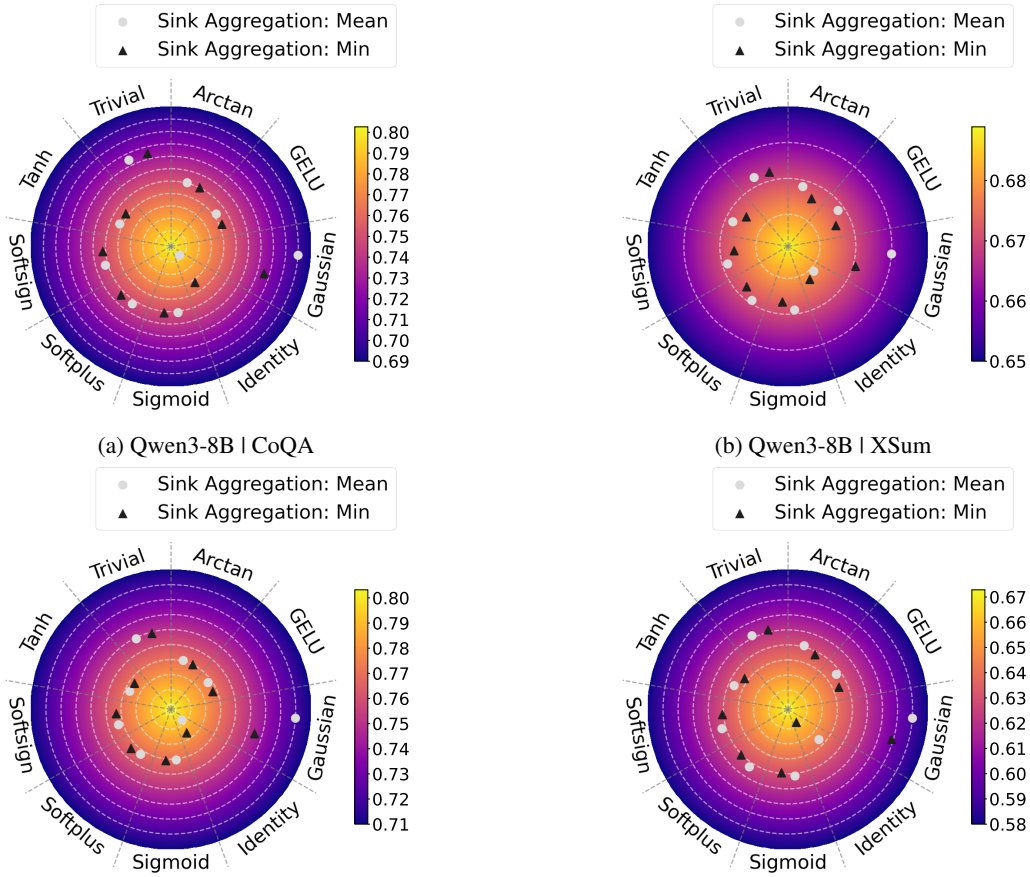
(a) Qwen3-8B | CoQA

(b) Qwen3-8B | XSum

(c) Llama-3.1-8B-Instruct | CoQA

(d) Llama-3.1-8B-Instruct | XSum

Figure 11: **Mean entropy aggregation.** Ablation of non-linear mappings that project the SinkRate to the $[0, 1]$ interval before multiplying it with entropy (Eq. 7). Trivial denotes that the SinkRate is set to 1, while Identity indicates that no mapping is applied. The heatmap shows AUC scores, where the center corresponds to AUC=1. Mappings are applied after Sink aggregation, and results are reported for both Mean and Min aggregations.



(a) Qwen3-8B | CoQA

(b) Qwen3-8B | XSum

(c) Llama-3.1-8B-Instruct | CoQA

(d) Llama-3.1-8B-Instruct | XSum

Figure 12: **Max entropy aggregation.** Ablation of non-linear mappings that project the SinkRate to the $[0, 1]$ interval before multiplying it with entropy (Eq. 7). Trivial denotes that the SinkRate is set to 1, while Identity indicates that no mapping is applied. The heatmap shows AUC scores, where the center corresponds to AUC=1. Mappings are applied after Sink aggregation, and results are reported for both Mean and Min aggregations.