Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detecting RAG hallucinations via cascade influence from attention to feedforward networks

Qiyang Gu
Academy of Military Sciences
Min Tang
Academy of Military Sciences
Fei Wang
Academy of Military Sciences
Hui Chen
Academy of Military Sciences
Fei Zhang
feizhang100@sina.com

Academy of Military Sciences
Pengyu Guo
Academy of Military Sciences

Research Article

Keywords: Retrieval-augmented generation (RAG), Hallucination detection, Attention–FFN cascade, Logit Lens, Mechanistic interpretability

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11527734/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Detecting RAG hallucinations via cascade influence from attention to feedforward networks

Qiyang Gu[1], Min Tang[1], Fei Wang[1], Hui Chen[1], Fei Zhang[1*] and Pengyu Guo[1*]

[1]National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing, 100071, China.

*Corresponding author(s). E-mail(s): feizhang100@sina.com; pengyu.guo@alumni.nudt.edu.cn;
Contributing authors: guqyams@outlook.com; mtangcn@163.com; feiwang@bupt.cn; chenhui@bupt.cn;

Abstract

Retrieval-augmented generation (RAG) mitigates the problem of insufficient knowledge when large language models (LLMs) are applied to specific domains, but models may still produce responses that are either unsupported or contradict the retrieved information, a phenomenon we term RAG hallucination. Recent mechanistic-interpretability detection methods attribute it to insufficient external-context utilization and excessive parametric-knowledge utilization. They quantify the two utilization signals and obtain a hallucination risk score by weighting or regression, but overlook the cascade influence between them. We propose CASDET, a training-free method considering Attention–FFN cascade influence. For each Transformer block, we form a local control state by removing only the current layer's external-context contribution at the FFN input, separate the influence signal as the difference between the two FFN outputs, and quantify it with the Logit Lens as the change in the current token's predicted probability. Layer-wise results are summarized into influence breadth, depth, and intensity, which are combined into a response-level hallucination risk score without thresholds or attention-head/layer selection. Under graded local intervention, a Page's L test confirms the monotonic strengthening of the influence signal. Across RAGTruth and HalluRAG with four open-source LLMs, CASDET is competitive with the stronger baselines, with a detection time comparable to that of single-forward uncertainty methods and 2–134× shorter than that of the remaining baselines. Our results indicate that the influence of external-context

introduction on the expression and use of parametric knowledge differs systematically between non-hallucinated and hallucinated responses, enabling training-free detection.

Keywords: Retrieval-augmented generation (RAG), Hallucination detection, Attention–FFN cascade, Logit Lens, Mechanistic interpretability

1 Introduction

Large language models (LLMs) have demonstrated strong language understanding and generation abilities in tasks such as question answering, summarization, and dialogue. However, limited by their finite parametric knowledge, LLMs frequently produce content that contradicts real-world facts or the provided input—a phenomenon known as LLM hallucination [1]. Retrieval-augmented generation (RAG) [2, 3] mitigates this problem by retrieving contextual information from external knowledge bases to improve answer accuracy. Yet even when the retrieved external context is accurate and relevant, models still generate responses that cannot be supported by or that contradict the external context. This phenomenon is defined as RAG hallucination [4, 5]. Reliably detecting such hallucinations is essential for the trustworthy deployment of RAG systems.

Recent studies have linked RAG hallucination to the relationship between the external context and the model's parametric knowledge. Research on knowledge conflict [6] shows that when the external context conflicts with the model's parametric knowledge, LLMs cannot always choose the more reliable knowledge source. They may instead continue to rely on their parametric knowledge. This indicates that the reliability of RAG generation depends not only on whether the retrieved external context is accurate and relevant, but also on whether this context can persistently and effectively influence subsequent predictions within the model. Building on this understanding, existing work [7, 8] formalizes RAG generation in terms of two types of signals, external-context utilization and parametric-knowledge utilization. These works attribute RAG hallucination to insufficient external-context utilization and excessive parametric-knowledge utilization.

These methods quantify the two utilization signals as relatively independent and fuse them by weighting or regression for RAG hallucination risk detection. This modeling, however, does not account for the cascade computation between the two in the Transformer. In a standard Transformer block [9], multi-head attention (hereafter attention) and the feedforward network (FFN) do not operate independently during generation. Attention first updates the residual stream, and the FFN then directly takes the updated residual state as its input. Consequently, the information that attention derives from the external context does not remain confined to the attention stage. Once written into the residual stream, it necessarily enters the input of the subsequent FFN.

Mechanistic interpretability research explains the mechanism behind this cascade. The FFN can be interpreted as a key-value memory structure [10], whose input state

determines which value vectors are activated. Its output can be viewed as an additive update to vocabulary predictions [11], in which different value vectors promote vocabulary concepts with specific semantics. External-context introduction therefore changes the expression and use of parametric knowledge by changing the FFN input. This paper addresses a more direct and specific question about RAG hallucination. How does external-context introduction affect the subsequent FFN's expression and use of parametric knowledge, and is this influence related to the occurrence of RAG hallucination?

To answer this question, we conduct the study presented in this work. The main contributions of this paper are as follows:

- We study the internal mechanism of RAG hallucination through the Attention–FFN cascade, proposing a testable conjecture on how external-context introduction affects the expression and use of parametric knowledge. This advances RAG hallucination analysis from the separate utilization strength of the two knowledge sources to their interaction inside the model.
- We propose CASDET (Cascade Detection), a training-free method for RAG hallucination detection that converts the influence signal δf, obtained through a local control state and the Logit Lens, into influence breadth, depth, and intensity, whose weighted sum yields a response-level hallucination risk score $\mathcal{H}$. It requires no thresholds and no selection of specific attention heads or Transformer layers.
- Under graded local intervention, a Page's L test confirms that the influence signal strengthens monotonically as the external-context contribution is restored, for every evaluated response. This establishes the validity of the underlying signal, not only of the detection method built on it.
- We evaluate CASDET on two benchmarks, RAGTruth and HalluRAG, across four open-source LLMs. CASDET achieves competitive detection performance on both benchmarks. Its detection time is comparable to single-forward uncertainty methods and substantially lower than the remaining baselines, without any parameter-selection set.

The remainder of the paper is organized as follows. Section 2 reviews the main categories of hallucination detection methods and states the conjecture that motivates this work. Section 3 presents the Attention–FFN cascade, the separation and quantification of the influence signal, its statistical validation, and the construction of the hallucination risk score in CASDET. Section 4 reports the experimental settings, the statistical validation results, the main results, the ablation study, and the sensitivity analysis of the weights. Section 5 discusses the support for the conjecture, the relation of the influence to existing methods, and open questions. Section 6 concludes the paper.

2 Related Work

We group existing hallucination detection methods into four categories according to the primary source of discriminative information.

Uncertainty methods. These methods assess the reliability of a response from the uncertainty of the prediction distribution during generation. Perplexity [12] is computed from the average per-token negative log-likelihood of the response, so a confident generation yields a small value. Length-Normalized Entropy (LN-Entropy) [13] averages the token-level predictive entropy over the response, which reduces the length bias of the uncertainty estimate. These signals reuse a single forward pass and thus add little cost, but they are computed from the prediction distribution alone. A confident generation may still lack support from the external context.

Generation-consistency methods. Hallucinations are identified by comparing the stability of responses across multiple generations. SelfCheckGPT [14] samples additional responses to the same query and scores each sentence of the target response by its factual consistency with the samples, since hallucinated content tends to vary across generations. EigenScore [15] moves this comparison into the internal representation space. It computes the covariance of the hidden states of the sampled generations and takes the average log-determinant as a measure of semantic dispersion, so that consistently generated responses receive small scores. These methods, however, require several generations per query, which multiplies the inference cost, and consistency indicates factual correctness only indirectly.

Context-verification methods. Methods in this category directly evaluate whether a response can be supported by the external context. RAGAs [16] estimates faithfulness with an LLM judge that decomposes the response into individual statements and computes the fraction of them that can be inferred from the external context. RefChecker [17] adopts a finer granularity. It extracts claim triplets from the response and classifies each triplet as entailed, contradictory, or neutral with respect to a reference, which also localizes the unsupported parts of the response. Such methods evaluate grounding directly but depend on an additional verification model, so their accuracy is bounded by the quality of that model's judgments.

Mechanistic-interpretability methods. Such methods detect hallucinations by examining the model's internal prediction mechanism, hidden states, and knowledge-utilization signals. ReDeEP [7] measures external-context utilization with specific attention heads and parametric-knowledge utilization with the change in the vocabulary distribution before and after the FFN, fusing the two into a hallucination score by regression. LUMINA [8] avoids relying on specific attention heads or Transformer layers. It measures external-context utilization by comparing the prediction distributions under correct-document and random-document conditions, and estimates parametric-knowledge utilization from the evolution of predictions across layers, fusing the two signals with manually set weights.

CASDET also exploits internal model signals. However, it focuses not on the separate utilization strength of the two knowledge sources, but on the influence of external-context introduction on the subsequent FFN's expression and use of parametric knowledge within the same Transformer block. Thus, CASDET differs not merely in introducing a new type of intermediate observation signal, but in using the Attention–FFN cascade directly as the basis for RAG hallucination detection. Accordingly, we propose the following conjecture.

Conjecture 1 (Influence of external-context introduction on the expression and use of parametric knowledge). When the external context is accurate and relevant to the query, its influence on the subsequent FFN's expression and use of parametric knowledge differs systematically between non-hallucinated and hallucinated responses.

3 Methodology

Overview. To investigate this conjecture, this section proposes CASDET, which detects RAG hallucinations by analyzing the Attention–FFN cascade computation inside each Transformer block. The signal it focuses on is the influence that the external context, after entering the current layer through attention, exerts on how the subsequent FFN expresses and uses parametric knowledge. Because attention outputs can be decomposed by source position, only the current layer's external-context contribution needs to be removed. The method separates this influence by constructing a local control state at the FFN input, takes the difference between the two FFN outputs as the influence signal, and quantifies it with the Logit Lens as a change in the predicted probability of the current generated token. The layer-wise results are finally combined into a response-level hallucination risk score. Figure 1 summarizes the pipeline.

Post-hoc detection setting. CASDET is designed for post-hoc hallucination detection rather than decoding-time prediction. Given a query, its retrieved context, and an already generated response, we perform a forward pass over the complete input sequence and examine the internal representations at response-token positions. At position t, the representation includes the observed token a_t, as permitted by the detection setting [7]. The Logit Lens projects this representation into the vocabulary space and yields a probability distribution. We refer to the probability of a_t under this distribution as the predicted probability of the current token. This probability reflects how strongly the representation supports a_t, not the probability with which the model would have generated it.

The remainder of this section proceeds in five steps. Subsection 3.1 analyzes the Attention–FFN cascade influence in the Transformer block. Subsection 3.2 separates the influence signal at the FFN input. Subsection 3.3 quantifies it as the change in the predicted probability of the current generated token with the Logit Lens. Subsection 3.4 validates it through graded local intervention. Subsection 3.5 defines influence breadth, influence depth, and influence intensity on the layer-wise results and combines them into a response-level hallucination risk score.

3.1 Analysis of the Attention–FFN cascade

CASDET builds on the computation order of the Transformer block. Mechanistic interpretability studies [18, 19] explain how model predictions are formed by analyzing the internal components of Transformers and how information flows between them. We consider the decoder-only Transformer adopted by current large language models. For notational simplicity, LayerNorm or RMSNorm is omitted below, whereas actual implementations retain the normalization structure of each model.

Cascade influence. For the l-th Transformer block, let $x_t^{\mathrm{mid},l}$ be the intermediate hidden state at position t after the attention update of this layer and before the FFN

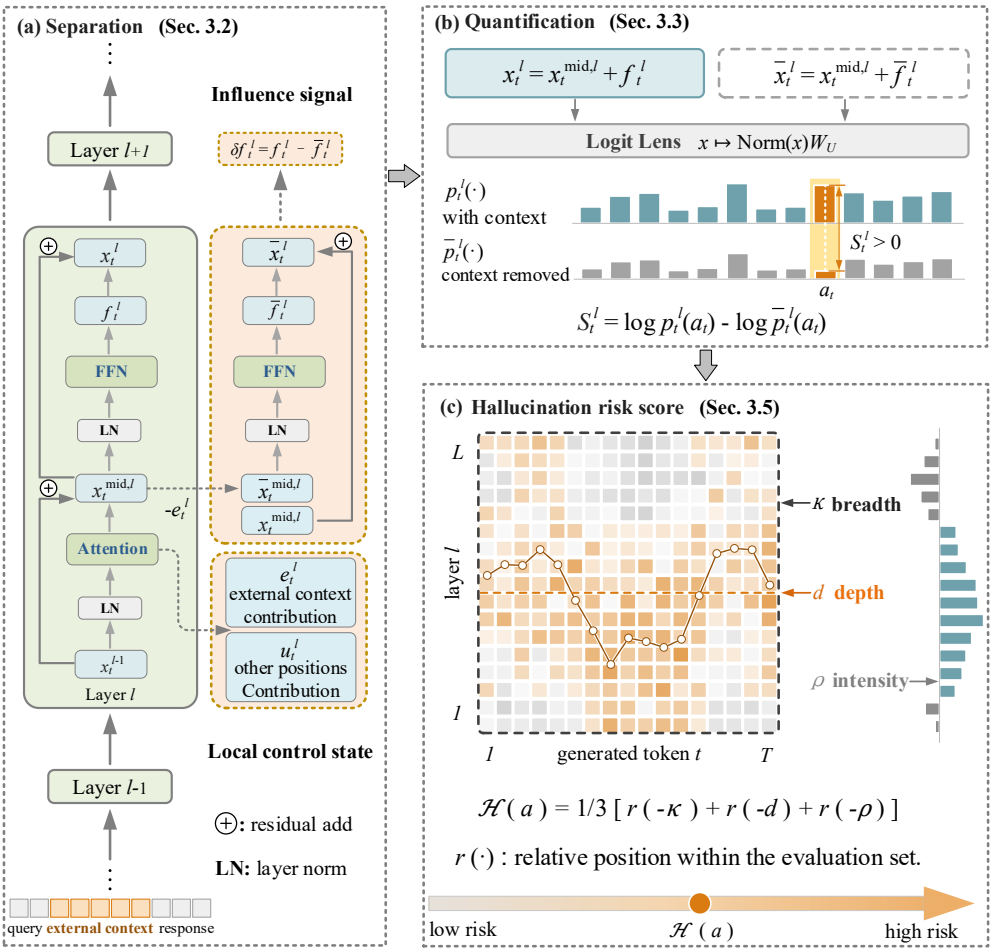


Fig. 1 Overview of CASDET. (a) Inside one Transformer block, the contribution the external context writes into the residual stream through attention is separated and removed at the FFN input to form a local control state. The difference between the two FFN outputs is the influence signal. (b) With the attention contribution kept identical, only the FFN output is replaced. Projecting the two residual states into the vocabulary space with the Logit Lens gives the difference in the log-probability of the current generated token. (c) The layer-wise results are summarized into influence breadth, influence depth, and influence intensity, whose relative positions are combined into a response-level hallucination risk score

computation. The attention update of this layer can then be written as:

$$x_t^{\mathrm{mid},l} = x_t^{l-1} + \sum_{h=1}^{H} \mathrm{Attn}^{l,h}\left(X_{\le t}^{l-1}\right) \tag{1}$$

where x_t^{l-1} is the hidden state at position t before entering layer l, H is the number of attention heads, $\mathrm{Attn}^{l,h}$ is the h-th attention head of layer l, and $X_{\le t}^{l-1}$ is the part of the hidden-state sequence entering this layer at positions up to t.

The FFN then updates the residual stream:

$$x_t^l = x_t^{\mathrm{mid},l} + \mathrm{FFN}^l\left(x_t^{\mathrm{mid},l}\right) \tag{2}$$

where x_t^l is the output of this layer and FFN^l is the feedforward network of layer l.

The cascade relation in Eqs. (1)–(2) indicates that the intermediate hidden state is both the result of the attention update and the input of the FFN. Once the part of the attention output produced by the external context is separated, its effect on the subsequent FFN computation can be analyzed. This requires that the attention output be decomposable by source position.

Attention decomposition. For the h-th attention head of layer l, its output at position t is a weighted sum of the contributions of all positions up to t, where the contribution of each position j is weighted by the attention weight $a_{t,j}^{l,h}$:

$$\mathrm{Attn}^{l,h}\left(X_{\le t}^{l-1}\right) = \sum_{j\le t} a_{t,j}^{l,h}\, x_j^{l-1} W_V^{l,h} W_O^{l,h} = \sum_{j\le t} a_{t,j}^{l,h}\, x_j^{l-1} W_{OV}^{l,h} \tag{3}$$

where $a_{t,j}^{l,h}$ is the attention weight of the h-th attention head at position t on position j, computed from the query–key inner product through softmax [18]. x_j^{l-1} is the hidden state at position j before entering layer l. $W_V^{l,h}$ and $W_O^{l,h}$ are the value projection matrix and the output projection matrix of this head, respectively, whose product $W_{OV}^{l,h}$ determines the information written into the residual stream [20]. The range $j \le t$ is the causal attention range.

The external context occupies a subset of these positions. Summing the contributions of these positions over all heads gives the external-context contribution e_t^l:

$$e_t^l = \sum_{h=1}^{H} \sum_{j\in \mathrm{ctx}} a_{t,j}^{l,h}\, x_j^{l-1} W_{OV}^{l,h} \tag{4}$$

where ctx is the set of token positions occupied by the external context in the input sequence.

The decomposition in Eq. (4) is the basis for CASDET. Since the query, the external context, and the model response in a RAG input sequence each occupy a well-defined interval of token positions, the summation over $j \le t$ can be split by source position, separating the vector contribution e_t^l of the external context to the current attention output from those of the remaining positions.

FFN computation. Prior mechanistic interpretability research shows that the FFN can be interpreted as a key-value memory structure [10]. Its input state determines the activation coefficient of each value vector, and its output can be viewed as an additive update to vocabulary predictions [11]. For layer l, the FFN computation

on input x can be written as:

$$\mathrm{FFN}^l(x) = \sum_{k=1}^{d_m} m_k^l(x)\, v_k^l \tag{5}$$

where d_m is the intermediate dimension of the FFN; v_k^l is the k-th value vector, corresponding to a piece of parametric knowledge that can be written into the residual stream; $m_k^l(x)$ is the activation coefficient of this value vector for input x, obtained from the inner product of x with the corresponding key vector through an activation function. Changing the input of the FFN is therefore equivalent to changing the expression and use of parametric knowledge, which motivates the local control construction in Subsection 3.2.

3.2 Separation of the influence signal

Subsection 3.1 established that the external context, once written into the residual stream through attention, changes the input of the subsequent FFN. This subsection addresses how to separate this influence from the model's normal computation. Figure 1(a) illustrates this separation, from the local control state to the resulting influence signal.

After the attention update of the l-th Transformer block, the residual state at a generation position t consists of three parts: the hidden state x_t^{l-1} entering the layer, the attention contributions of the remaining source positions in the current layer, and the current-layer external-context contribution e_t^l. Merging the contributions of the remaining positions into u_t^l, we construct a local control state at the FFN input that removes only the external-context contribution:

$$\bar{x}_t^{\mathrm{mid},l} = x_t^{\mathrm{mid},l} - e_t^l = x_t^{l-1} + u_t^l \tag{6}$$

where $\bar{x}_t^{\mathrm{mid},l}$ is the FFN input in the local control state, with the overline marking the local-control counterpart of a normal-state quantity, and u_t^l is the sum of the attention contributions of all source positions other than the external context in Eq. (3).

Because only the newly written external-context contribution is removed at the FFN input, the representation of the external context already written into the residual stream at earlier layers is not removed. The contributions of the query, the model response, and the remaining source positions of the current layer also stay unchanged. The difference between the two FFN inputs therefore comes solely from the current-layer external-context contribution. Replacing a single source contribution while holding the remaining activations fixed follows the intervention approach [21–23] of mechanistic interpretability studies that examine the causal role of a component. Compared with directly deleting the external context from the model input [24], this local control leaves the computation of the earlier layers intact.

The difference between the FFN outputs under the two inputs is exactly the influence of the external-context introduction on the expression and use of parametric knowledge. Let f_t^l and $\bar{f}_t^l$ be the FFN outputs of layer l in the normal state and the

local control state, respectively, with their difference defining the influence signal:

$$\delta f_t^l = f_t^l - \bar{f}_t^l = \mathrm{FFN}^l\left(x_t^{\mathrm{mid},l}\right) - \mathrm{FFN}^l\left(\bar{x}_t^{\mathrm{mid},l}\right) \tag{7}$$

where δf_t^l is the influence signal. The two FFN inputs differ only in the input-side term e_t^l, so the difference between the two outputs separates the effect of the external-context contribution.

Substituting the FFN representation in Eq. (5) into Eq. (7), we obtain the sum of the changes in the activation coefficients of the value vectors:

$$\delta f_t^l = \sum_{k=1}^{d_m} \left[m_k^l\left(x_t^{\mathrm{mid},l}\right) - m_k^l\left(\bar{x}_t^{\mathrm{mid},l}\right) \right] v_k^l \tag{8}$$

where $m_k^l(\cdot)$ and v_k^l are as defined in Subsection 3.1. As Eq. (8) shows, the external-context contribution changes the activation coefficients of the value vectors once it enters the FFN input. The coefficient changes in turn alter what the FFN writes into the residual stream. This equation only explains how the influence signal relates to the internal FFN computation. CASDET operates directly on the model's original FFN and requires no specific FFN parameterization, so gated FFN structures such as SwiGLU and GeGLU are handled equally well.

3.3 Quantification of the influence signal

The influence signal is a high-dimensional vector and cannot be used directly for hallucination detection. This subsection quantifies it as a change in predicted probability. Figure 1(b) depicts this quantification, from the two residual states to the difference in predicted probability.

With the attention contribution kept identical, we replace only the FFN output, thereby obtaining two residual states. Let x_t^l and $\bar{x}_t^l$ be the outputs of layer l in the normal state and the local control state, respectively. The two states share the same intermediate hidden state, so the external-context contribution that the current-layer attention writes into the residual stream is fully preserved in both states. The only difference between them is the influence signal defined in Eq. (7).

To translate this difference into a change in the prediction, we project the two states into the vocabulary space with the Logit Lens [25, 26], and define the local relative prediction change S_t^l as the difference in the log-probability of the current generated token a_t under the two states:

$$\begin{aligned} S_t^l &= \log p_t^l(a_t) - \log \bar{p}_t^l(a_t) \\ &= \log \left[\mathrm{Softmax}\big(\mathrm{LogitLens}(x_t^l)\big)\right]_{a_t} - \log \left[\mathrm{Softmax}\big(\mathrm{LogitLens}(\bar{x}_t^l)\big)\right]_{a_t} \end{aligned} \tag{9}$$

where a_t is the current generated token at position t, p_t^l and $\bar{p}_t^l$ are the vocabulary distributions obtained at layer l under the two states, and $p_t^l(a_t)$ denotes the probability that the distribution p_t^l assigns to the token a_t. A positive local relative prediction change indicates that the external-context introduction at this layer increases the predicted probability of the current generated token.

Since this influence may be distributed across different layers, CASDET does not pre-select specific Transformer layers. It averages over all L layers to obtain the token-level relative prediction effect of the current generated token:

$$S_t(a_t) = \frac{1}{L}\sum_{l=1}^{L} S_t^l = \log\left[\prod_{l=1}^{L}\frac{p_t^l(a_t)}{\bar{p}_t^l(a_t)}\right]^{1/L} \tag{10}$$

where L is the total number of Transformer blocks. A value above zero indicates that the influence overall increases the predicted probability of the current generated token. A larger value corresponds to a lower risk of hallucination generation, and a value below zero indicates the opposite. Rather than measuring the influence signal by vector magnitude, CASDET quantifies it by the probability change it induces in the prediction. The high-dimensional effect inside the Attention–FFN computation is thus converted into a scalar signal directly related to RAG hallucination generation.

3.4 Statistical validation of the influence signal

The previous subsections separate and quantify the influence signal, but observing a correlation between $S_t(a_t)$ and RAG hallucination generation alone is not sufficient to establish the validity of this signal. Following the statistical-validation approach of LUMINA [8], this subsection controllably adjusts the external-context contribution entering the FFN input and examines whether the relative prediction effect changes as expected. This determines whether the signal genuinely reflects the influence of the external-context introduction on the subsequent FFN's expression and use of parametric knowledge.

Specifically, we introduce a graded local intervention between the normal state and the local control state. Given an intervention coefficient $\gamma \in [0, 1]$, the external-context contribution entering the FFN of layer l is replaced by γ times its original value. Equivalently, the intervention input of the FFN at this layer is the normal FFN input minus $(1-\gamma)$ times the external-context contribution. The remaining computation of the layer then proceeds with this input, and the token-level relative prediction effect under this graded local intervention is denoted as $S_t^{(\gamma)}(a_t)$:

$$S_t^{(\gamma)}(a_t) = \frac{1}{L}\sum_{l=1}^{L}\log\frac{p_{t,\gamma}^l(a_t)}{p_{t,0}^l(a_t)} \tag{11}$$

where γ is the intervention coefficient, $p_{t,\gamma}^l$ is the vocabulary distribution obtained at layer l under the coefficient γ, and $p_{t,0}^l$ corresponds to the $\gamma = 0$ local control state. $\gamma = 0$ corresponds to the local control state in Subsection 3.2, which fully removes the external-context contribution entering the FFN input of this layer, whereas $\gamma = 1$ corresponds to the normal state. As γ increases from 0 to 1, this contribution is progressively restored toward its value in the normal state.

The mean absolute token-level relative prediction effect of the i-th response is defined as $M_i(\gamma)$:

$$M_i(\gamma) = \frac{1}{T_i} \sum_{t=1}^{T_i} \left| S_t^{(\gamma)}(a_t) \right| \tag{12}$$

where T_i is the number of generated tokens of this response, and $|\cdot|$ denotes the absolute value. Because the sign of $S_t^{(\gamma)}(a_t)$ indicates only the direction of the effect on the predicted probability, whereas this subsection concerns whether the strength of the relative prediction effect increases as the contribution is restored, we take the absolute value and average it within the response.

Accordingly, we propose the following validation hypothesis.

H1. If the influence signal obtained by CASDET is valid, then as γ increases, the token-level relative prediction effect induced by the external-context contribution exhibits an overall increasing trend. In other words, its effect on the prediction of the current generated token strengthens as more of the external-context contribution enters the FFN input.

Because each response is measured repeatedly at five values of γ and H1 posits an ordered trend, we adopt Page's L test for an ordered trend [27]. Its null hypothesis is that the effect is independent of the intervention coefficient. The test first assigns ranks to the effects of each response at the five values of γ by magnitude (ties receive average ranks). It then sums the ranks of all responses at each value of γ, weighted by its ordinal index, to obtain the Page statistic L. The statistic is standardized into z using its expectation and standard deviation under the null hypothesis. The value of z is the number of standard deviations by which L exceeds the null expectation, and a larger value indicates a stronger increasing trend of the effect with the intervention coefficient.

Since H1 specifies the direction of the effect, the test is one-sided. The null hypothesis is rejected when the one-sided p-value is below the significance level $\alpha = 0.05$. We perform this test on the evaluated responses in Subsection 4.2. A positive result establishes that the separated and quantified influence signal reflects the influence of the external-context introduction on the subsequent FFN's expression and use of parametric knowledge.

3.5 Computation of the hallucination risk score

Statistical validation concerns the influence signal itself. Response-level hallucination detection additionally requires information about the signal's distribution across layers. Prior studies [15, 28] have shown that the internal states and prediction evolution of different layers contain discriminative information independent of the final output, and that the predictions of different layers contribute unequally to the final output [29]. For a response a with T generated tokens, Eq. (9) yields a set of layer-wise results S_t^l ($t = 1, \ldots, T$; $l = 1, \ldots, L$). The aggregation in Eq. (10) summarizes this set over layers into $S_t(a_t)$, giving the overall direction of the influence signal. By the cascade relation of Subsection 3.1, the output of the l-th FFN is written into the residual stream and continues to participate in the computation of all subsequent layers.

The layers that carry the influence, its reach toward layers close to the output, and its strength therefore jointly determine how much it can change the final prediction. Figure 1(c) gives an overview of the scoring stage, from the layer-wise results to the response-level hallucination risk score.

We define three statistics on this set of results, measuring the breadth, depth, and intensity of the influence that the signal provides to the current generated token across layers. Influence breadth $\kappa(a)$ counts the proportion of layer–token pairs with $S_t^l > 0$:

$$\kappa(a) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{l=1}^{L}\mathbf{1}\big[S_t^l > 0\big] \tag{13}$$

where T is the number of generated tokens in the response, L is the total number of Transformer blocks, and $\mathbf{1}[\cdot]$ is the indicator function.

Influence depth $d(a)$ takes the relative prediction changes above zero at each layer as weights and computes a weighted average of the layer indices, averaged over all generated tokens:

$$d(a) = \frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{l=1}^{L} l \cdot \max\big(S_t^l, 0\big)}{\sum_{l=1}^{L}\max\big(S_t^l, 0\big)} \tag{14}$$

where $\max(\cdot, 0)$ takes the positive part.

Influence intensity $\rho(a)$ first averages the relative prediction changes at each layer, then normalizes them layer by layer and takes the root mean square, so that layers of different magnitudes receive equal weight:

$$\rho(a) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\frac{\bar{S}_l(a)}{\sigma_l}\right)^2} \tag{15}$$

where $\bar{S}_l(a)$ is the mean relative prediction change of all generated tokens in the response at layer l, and σ_l is the standard deviation of S_t^l at layer l within the same evaluation set.

The three statistics have different value ranges and reflect different dimensions of this influence. They describe how this influence is distributed across layers. Smaller values of κ and d indicate that positive influence involves fewer layers and is positioned earlier. Smaller ρ indicates weaker layer-wise mean influence or stronger cancellation between positive and negative token effects. The generation of the current token may then be insufficiently or inconsistently supported by the external context, and the risk of RAG hallucination is higher. To place the three statistics on a common scale, we rank each statistic over the entire evaluation set along the risk direction and replace the raw value with the relative position of each response in that ranking. A weighted sum of the three relative positions then gives the hallucination risk score of the response a:

$$\mathcal{H}(a) = w_\kappa\, r(-\kappa(a)) + w_d\, r(-d(a)) + w_\rho\, r(-\rho(a)) \tag{16}$$

Algorithm 1 Computation of the CASDET hallucination risk score

Require: Set of evaluated responses $\mathcal{A}$ with their external contexts
Require: Number of Transformer layers L
Ensure: Hallucination risk score $\mathcal{H}(a)$ of each response

for each response $a \in \mathcal{A}$ **do**
  Run the forward pass and record the attention outputs, FFN inputs $x_t^{\mathrm{mid},l}$, and FFN outputs f_t^l of the generated tokens at each layer
  for $l = 1$ to L and each generation position t **do**
    Decompose the attention outputs by source position and separate the external-context contribution e_t^l
    $\bar{x}_t^{\mathrm{mid},l} \Leftarrow x_t^{\mathrm{mid},l} - e_t^l$ ▷ local control state
    $\bar{f}_t^l \Leftarrow \mathrm{FFN}^l(\bar{x}_t^{\mathrm{mid},l})$
    $\delta f_t^l \Leftarrow f_t^l - \bar{f}_t^l$ ▷ influence signal
    Replace f_t^l with $\bar{f}_t^l$ to obtain the local-control output $\bar{x}_t^l$
    Project x_t^l and $\bar{x}_t^l$ with the Logit Lens to obtain p_t^l and $\bar{p}_t^l$
    $S_t^l \Leftarrow \log p_t^l(a_t) - \log \bar{p}_t^l(a_t)$ ▷ local relative prediction change
  end for
end for
Compute the standard deviations σ_l of S_t^l within $\mathcal{A}$
Compute $\kappa(a)$, $d(a)$, and $\rho(a)$ ▷ influence breadth, depth, and intensity
Rank $\kappa(a)$, $d(a)$, and $\rho(a)$ within $\mathcal{A}$ along the risk direction and replace each raw value with its relative position $r(\cdot) \in [0, 1]$
Compute the hallucination risk score $\mathcal{H}(a)$ of each response
return the hallucination risk scores $\{\mathcal{H}(a) \mid a \in \mathcal{A}\}$

where $r(\cdot)$ denotes the relative position after ranking along the risk direction, between 0 and 1, with $w_\kappa + w_d + w_\rho = 1$. The relative position is independent of the original scales of the three statistics, so differences in scale do not affect the weighted sum. The three statistics share a common relative-position scale, so we use equal weights, $w_\kappa = w_d = w_\rho = 1/3$, as the default setting, which favors none of them. Subsection 4.5 tests whether tuning them yields a reliable gain across benchmarks. In addition, the evaluation uses threshold-free metrics, so no decision threshold needs to be set for either dataset.

The complete procedure is summarized in Algorithm 1. The first stage is performed independently for each response, and the second stage operates within the set of evaluated responses. The numerical calibration of $\mathcal{H}(a)$ therefore depends on that set, while the discriminative information still comes from the response itself. All three statistics are aggregated within the response. If a single response must be scored in isolation, σ_l and $r(\cdot)$ can be fixed on an unlabeled reference set, then $\mathcal{H}(a)$ depends only on the response a. Subsection 4.3 discusses the corresponding experiment.

4 Experiments

This section evaluates CASDET on two public RAG hallucination detection benchmarks. We first validate the influence signal under the graded local intervention of

Subsection 3.4. This establishes whether the signal that CASDET relies on behaves as conjectured. We then report the detection performance and detection time of CASDET against representative baselines on four open-source LLMs. We next examine, through ablation, the roles of the Attention–FFN cascade relation, the local control state, and the three statistics. The final subsection examines the sensitivity of the results to the weights of the three statistics.

4.1 Experimental settings

This subsection describes the experimental setup, including the LLMs, the datasets, the baselines, the evaluation metrics, the experimental environment, and the implementation details of CASDET.

LLMs. We conduct experiments on four open-source LLMs: Llama2-7B, Llama2-13B [30], Llama3-8B [31], and Mistral-7B [32]. These models cover different scales and FFN implementation forms, which allows us to examine the applicability of CASDET to different Transformer structures. For the model responses already provided by the benchmark datasets, CASDET uses the corresponding LLM to re-run the forward pass. It records the attention outputs, FFN inputs, and FFN outputs of the generated tokens at each Transformer layer. The response-level hallucination risk score is then computed from these signals.

Datasets. Experiments use two representative benchmark datasets for RAG hallucination detection: RAGTruth [4] and HalluRAG [33]. RAGTruth is a human-annotated RAG hallucination dataset covering three tasks—question answering, data-to-text writing, and news summarization. Each sample consists of a query, retrieved documents, a model response, and the corresponding hallucination annotation. The original RAGTruth data do not contain the generated responses of Llama3-8B. We adopt the Llama3-8B extension test set constructed and annotated by Sun et al. [7]. HalluRAG is a dataset for detecting closed-domain hallucination in RAG systems, which constructs answerable and unanswerable prompts for each question. We keep only the answerable samples, i.e., those whose queries can be answered from the given context. Because HalluRAG provides responses only for Llama2-7B, Llama2-13B, and Mistral-7B, we do not report Llama3-8B. After this filter, the official test split contains only 81 answerable samples (10–20 hallucinated per model), a size at which AUROC/PCC estimates have large sampling variability. We therefore evaluate all 540/540/537 answerable samples with five-fold cross-validation.

Baselines. According to the categorization of hallucination detection methods above, we use all eight methods discussed in Section 2 as baselines, two from each category: the uncertainty methods Perplexity and LN-Entropy, the generation-consistency methods SelfCheckGPT and EigenScore, the context-verification methods RAGAs and RefChecker, as well as the mechanistic-interpretability methods ReDeEP and LUMINA. ReDeEP and LUMINA belong to the same mechanistic-interpretability category as CASDET and are the most direct points of comparison. To compare all methods on the same evaluation set and input convention, we run every baseline ourselves. ReDeEP and LUMINA differ from their original implementations on the HalluRAG benchmark. ReDeEP selects its heads and layers on the other four folds rather than on a separate parameter-selection set, and holds out the remaining fold for

evaluation in each round. LUMINA, besides the different construction of this benchmark's evaluation set, uses the unified input format, omitting the space before the response in the original implementation.

Evaluation metrics. CASDET is a RAG hallucination detection method based on a continuous score and does not rely on a fixed hallucination threshold. We evaluate the detection performance of CASDET with three metrics: the Area Under the Receiver Operating Characteristic Curve (AUROC), the Area Under the Precision–Recall Curve (AUPRC), and the Pearson correlation coefficient (PCC). AUROC measures the overall ability of a method to distinguish hallucinated from non-hallucinated responses across all decision thresholds, reflecting the ranking validity of a continuous score. AUPRC reflects the trade-off between precision and recall and, under a small proportion of hallucinated samples, better reflects practical detection performance than AUROC [34]. PCC evaluates the degree of linear correlation between the continuous hallucination risk score and the binary hallucination labels, complementarily verifying the numerical calibration quality of the score beyond the ranking relation.

Experimental environment. All experiments are conducted on a server with two AMD EPYC 9334 processors, 256 GB of memory, and three NVIDIA GeForce RTX 4090 GPUs (24 GB each). The software environment consists of Windows 10 (build 19045), Python 3.12.0, PyTorch 2.3.1 with CUDA 12.1, and the Hugging Face transformers library 4.57.6.

CASDET preserves the original LayerNorm or RMSNorm, attention, and FFN structures of each LLM. It does not pre-select specific attention heads or Transformer layers and instead computes over all layers. Its only hyperparameters are the three weights in Eq. (16). The main experiments use equal weights as the default setting, and Subsection 4.5 analyzes their sensitivity. Unlike mechanistic-interpretability methods that select attention heads or Transformer layers on a parameter-selection set, CASDET uses no additional parameter-selection samples. Its test-set labels are not used to set these weights or for any other parameter selection.

4.2 Statistical validation of the influence signal

To validate the influence signal, we test the following implication. If the signal is valid, its strength should increase as the external-context contribution entering the FFN is restored. Following the graded local intervention of Subsection 3.4, we set $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$, where $\gamma = 0$ is the local control state and $\gamma = 1$ is the normal state. For each response on RAGTruth, we compute the mean absolute token-level relative prediction effect $M_i(\gamma)$.

Figure 2 plots $M_i(\gamma)$ against γ for the four LLMs. Every curve rises monotonically from the local control state at $\gamma = 0$ to the normal state at $\gamma = 1$, and the rise is close to linear across the five values of γ. The prediction thus responds to the restored external-context contribution in a graded rather than threshold-like way, strengthening overall. On these averaged curves alone, however, monotonicity could still be an artifact of averaging. To this end, Table 1 tests the ordering at the level of individual responses.

Page's L test confirms the ordering for every response. In all four models, the statistic reaches the theoretical upper limit of this setting, $L = 55N$ (the maximum

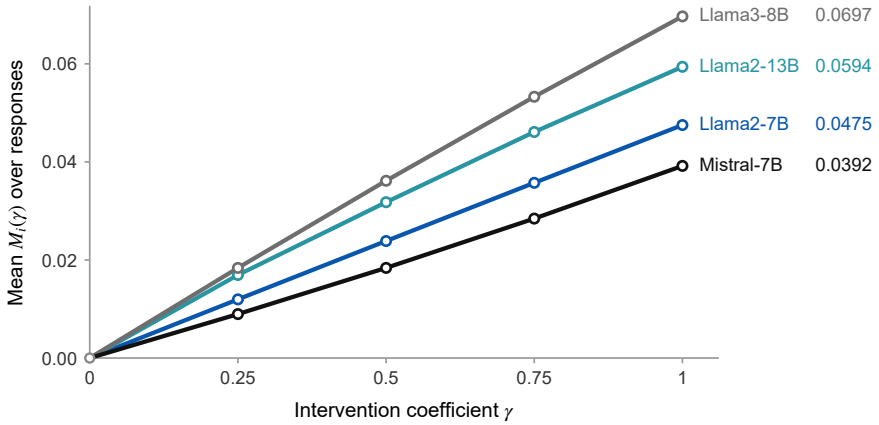


Fig. 2 Mean absolute token-level relative prediction effect $M_i(\gamma)$ under graded local intervention. Each curve gives the average of $M_i(\gamma)$ over the responses of one LLM, and the value at $\gamma = 1$ is listed on the right

Table 1 Page's L test for an ordered trend in the mean absolute token-level relative prediction effect $M_i(\gamma)$ across the five values of γ on RAGTruth

LLM	N	L	$E[L]$	z
Llama2-7B	449	24695	20205	42.37***
Llama2-13B	450	24750	20250	42.42***
Llama3-8B	360	19800	16200	37.94***
Mistral-7B	436	23980	19620	41.76***

N: number of responses; L: Page statistic; $E[L]$: expectation of L under the null hypothesis of no trend; z: standardized statistic; asterisks denote the one-sided significance level: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

of the range $[35N, 55N]$). The standardized statistic z lies far beyond the one-sided critical value ($p < 0.001$). The null hypothesis of no trend is thus rejected at $\alpha = 0.05$ for every model. Reaching the upper limit is stronger than plain significance. It holds only if, for every response, the five $M_i(\gamma)$ values are strictly increasing in γ, not merely increasing on average. The separated signal thus satisfies H1 strictly, for every response rather than only on average.

Overall, these results validate the influence signal and confirm that it reflects the influence of external-context introduction on the subsequent FFN's expression and use of parametric knowledge.

4.3 Main results

Having established the validity of the influence signal, we next evaluate its use for hallucination detection. Table 2 reports the response-level RAG hallucination detection results of CASDET and the baselines on RAGTruth and HalluRAG. For each LLM, AUROC, AUPRC, and PCC are listed separately for each benchmark, followed by the detection time for completing the whole evaluation set. Detection time covers the

Table 2 Performance comparison of CASDET and baselines on RAGTruth and HalluRAG. Detection time is the end-to-end time in seconds for the whole evaluation set

LLM	Baseline	RAGTruth				HalluRAG			
		AUROC ↑	AUPRC ↑	PCC ↑	Time ↓	AUROC ↑	AUPRC ↑	PCC ↑	Time ↓
Llama2-7B	Perplexity	0.5181	0.4945	-0.0054	**82.1**	0.3229	0.1954	-0.0265	**71.2**
	LN-Entropy	0.6927	0.6632	0.3279	95.4	**0.8004**	0.5941	0.4564	79.4
	SelfCheckGPT	0.5615	0.5499	0.1165	7419.2	0.7423	0.5460	0.4030	8566.6
	EigenScore	0.6243	0.5876	0.2078	1164.8	0.7515	0.6119	0.4607	1043.3
	RAGAs	0.7208	0.6546	0.2803	6447.7	0.7463	0.5878	0.5010	2020.5
	RefChecker	0.7189	0.6471	0.2396	2842.8	0.7320	0.5665	0.4914	975.6
	ReDeEP	0.7354	0.6951	0.3942	8510.2	0.7597	0.6003	0.5037	3053.5
	LUMINA	0.7253	0.7408	0.4295	442.9	0.7815	0.6128	**0.5195**	503.1
	CASDET	**0.7485**	**0.7449**	**0.4374**	111.2	0.7829	**0.6133**	0.4443	94.9
Llama2-13B	Perplexity	0.4365	0.3942	-0.1230	**161.9**	0.2262	0.1281	-0.2336	**144.4**
	LN-Entropy	0.7659	0.6776	0.4426	189.4	0.7631	0.4638	0.3917	157.6
	SelfCheckGPT	0.7482	0.6488	0.4203	19105.5	0.7701	0.4840	0.4021	11449.2
	EigenScore	0.7541	0.6254	0.4399	2267.2	0.7817	0.4922	0.3745	1964.8
	RAGAs	0.7470	0.6187	0.2793	7049.3	0.7502	0.5113	0.3425	2248.8
	RefChecker	0.7612	0.6508	0.4987	3157.1	0.6632	0.4740	0.2968	981.4
	ReDeEP	0.7674	0.7568	0.5225	6418.5	0.7836	0.5071	**0.4180**	672.9
	LUMINA	0.7883	0.7523	**0.5520**	713.2	0.7605	0.5370	0.3845	533.2
	CASDET	**0.7980**	**0.7670**	0.5261	312.9	**0.7894**	**0.5574**	0.4044	266.5
Llama3-8B	Perplexity	0.7107	0.7132	0.3540	**104.3**	—	—	—	—
	LN-Entropy	0.7019	0.7092	0.3440	129.0	—	—	—	—
	SelfCheckGPT	0.6303	0.6112	0.2680	4913.1	—	—	—	—
	EigenScore	0.6829	0.6587	0.2812	1395.5	—	—	—	—
	RAGAs	0.5841	0.5967	0.1281	9746.4	—	—	—	—
	RefChecker	0.6754	0.6393	0.3371	4276.0	—	—	—	—
	ReDeEP	0.7048	0.7013	0.3418	436.2	—	—	—	—
	LUMINA	0.7210	0.7394	0.3644	501.9	—	—	—	—
	CASDET	**0.7228**	**0.7503**	**0.3737**	114.8	—	—	—	—
Mistral-7B	Perplexity	0.6588	0.6645	0.2003	**80.2**	0.6347	0.2240	0.1803	**70.3**
	LN-Entropy	0.7464	0.7464	0.4096	91.4	**0.8946**	0.5086	0.5675	76.9
	SelfCheckGPT	0.5725	0.6158	0.1456	6238.1	0.8075	0.4945	0.5158	3473.6
	EigenScore	0.5652	0.6277	0.1992	870.3	0.8135	0.4778	0.5167	605.0
	RAGAs	0.7136	0.6800	0.2533	13395.1	0.7687	0.4274	0.5230	1953.8
	RefChecker	0.7306	0.7885	0.3791	6589.2	0.6466	0.2530	0.1002	715.5
	ReDeEP	0.7664	0.8014	**0.5396**	2546.2	0.8195	0.5040	0.3942	294.4
	LUMINA	0.7703	0.8060	0.4995	361.0	0.8207	**0.6080**	**0.6259**	248.9
	CASDET	**0.7748**	**0.8085**	0.4713	100.0	0.8290	0.5138	0.3950	90.2

The best result in each column is shown in bold and the second best is underlined. An em dash indicates an unavailable result.

end-to-end scoring process excluding model loading, and all methods are computed on the same evaluation set. For ReDeEP, the parameter-selection pass that its design requires outside the evaluation set is counted toward its time.

Table 2 shows that, on RAGTruth, CASDET achieves the best AUROC and AUPRC for all four LLMs, and its PCC is in the top three for all four. On HalluRAG, CASDET ranks in the top two on both AUPRC and AUROC for all three LLMs, and its PCC is within 0.10 of LUMINA, the best baseline on this metric, in aggregate. Across the two benchmarks, CASDET ranks in the top three in 19 of the 21 model–metric combinations, including 13 best results, placing its detection performance at the level of the stronger methods.

Table 3 Fixed-reference scoring versus full-batch scoring. AUROC is shown as a representative metric, with values for both scoring schemes (fixed reference: mean over five random five-fold splits); AUROC, AUPRC, and PCC report the maximum absolute changes relative to full-batch scoring over the same splits; Spearman is the minimum rank correlation between the two score orderings

Dataset	LLM	N	AUROC		Max absolute change			Spearman
			Full batch	Fixed reference	AUROC	AUPRC	PCC	
RAGTruth	Llama2-7B	449	0.7485	0.7492	0.0025	0.0029	0.0021	0.9991
	Llama2-13B	450	0.7980	0.7977	0.0010	0.0035	0.0020	0.9983
	Llama3-8B	360	0.7228	0.7228	0.0005	0.0018	0.0016	0.9990
	Mistral-7B	436	0.7748	0.7753	0.0009	0.0018	0.0008	0.9991
HalluRAG	Llama2-7B	540	0.7829	0.7821	0.0016	0.0013	0.0023	0.9989
	Llama2-13B	540	0.7894	0.7887	0.0012	0.0053	0.0024	0.9990
	Mistral-7B	537	0.8290	0.8281	0.0019	0.0048	0.0022	0.9988

The reference set consists of the other four folds and uses no hallucination labels. Each response in the remaining fold is scored from its within-response statistics using these fixed values.

Compared with the closeness in detection performance, the advantage of CASDET in computational cost is more prominent. CASDET takes 100–313 s to complete the whole evaluation set on RAGTruth and 90–267 s on HalluRAG. Relative to the six baselines that require additional computation, it is 2.3–134× faster on RAGTruth and 2.0–90× faster on HalluRAG, while its detection time remains in the same order of magnitude as Perplexity and LN-Entropy, which require only a single forward pass. This gap follows from the designs of the baselines: SelfCheckGPT and EigenScore require multiple sampled generations; RAGAs and RefChecker require an external judge or verification model; ReDeEP searches attention heads and Transformer layers on a parameter-selection set; LUMINA requires one extra forward pass per response under the random-document condition. CASDET needs none of these. It performs a single forward pass on the evaluated responses and additionally computes one local-control FFN output per layer, with no parameter-selection set and no pre-selected attention heads or Transformer layers. CASDET thus attains detection performance close to the stronger methods at a substantially lower cost.

To validate the independent scoring scheme of Subsection 3.5, we perform five-fold cross-fitting within each dataset–LLM combination. With the other four folds as an unlabeled reference set, σ_l and $r(\cdot)$ are fixed. Each response in the remaining fold is scored from its within-response statistics using these fixed values. The procedure is repeated over five random fold assignments. Table 3 reports the per-combination results. Across the seven combinations, the maximum absolute changes in AUROC, AUPRC, and PCC relative to full-batch scoring are 0.0025, 0.0053, and 0.0024, respectively. The Spearman correlation between the two score orderings is at least 0.9983. This shows that full-batch scoring is not required by CASDET. Within-set calibration requires no separate reference set and is the default for batch evaluation. For single-response scoring, σ_l and $r(\cdot)$ can be fixed on an unlabeled reference set, after which each response is scored independently with essentially unchanged detection performance and score ordering.

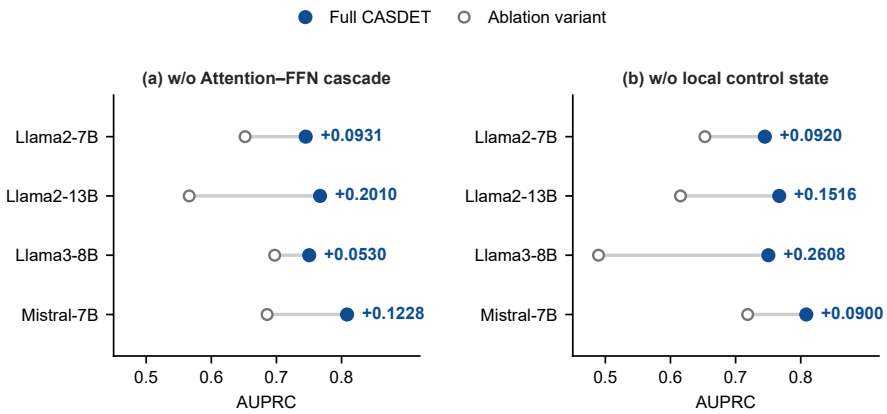


Fig. 3 Ablation study on RAGTruth. AUPRC comparison between the full CASDET and two ablated variants: (a) removing the Attention–FFN cascade and (b) removing the local control state. The annotated value in each row gives the AUPRC gain of the full CASDET over the ablated variant

4.4 Ablation study

The main results above establish the overall effectiveness of CASDET. This subsection analyzes the contribution of three key designs of CASDET: the Attention–FFN cascade relation, the local control state, and the three influence statistics. The first two modify the signal itself and are ablated on RAGTruth with AUPRC as the evaluation metric. The third only varies the combination of the three statistics computed from the same layer-wise results, while its evaluation uses the three metrics averaged over RAGTruth and HalluRAG. In every case, all other settings are kept identical to the full CASDET.

Impact of the Attention–FFN cascade. CASDET uses the Attention–FFN cascade relation directly as the basis for RAG hallucination detection. To examine the role of the subsequent FFN computation, we construct a control setting that considers only the direct effect of attention. It quantifies the direct influence of the external-context contribution, after being written into the residual stream by attention, on the prediction of the current generated token. It no longer considers the influence on the expression and use of parametric knowledge in the subsequent FFN. Figure 3a compares the AUPRC of this setting and of the full CASDET on the four LLMs. Against the attention-only control, the full pipeline gains 0.0530–0.2010 AUPRC on every LLM, 0.1175 on average. This influence on the subsequent FFN's expression and use of parametric knowledge thus carries information beyond the direct contribution of attention. Incorporating it markedly improves detection performance.

Impact of the local control state. CASDET constructs, at the FFN input, a local control state that removes only the current layer's external-context contribution. To examine its role, we omit this construction and quantify directly with the normal FFN output. The Logit-Lens-based quantification of Subsection 3.3 and the hallucination risk score are kept unchanged. Removing the local control state then degrades detection on every LLM (Figure 3b). The full CASDET exceeds the ablated setting by 0.0900–0.2608 AUPRC, 0.1486 on average. The local control state is thus essential for the separated signal to discriminate hallucinations.

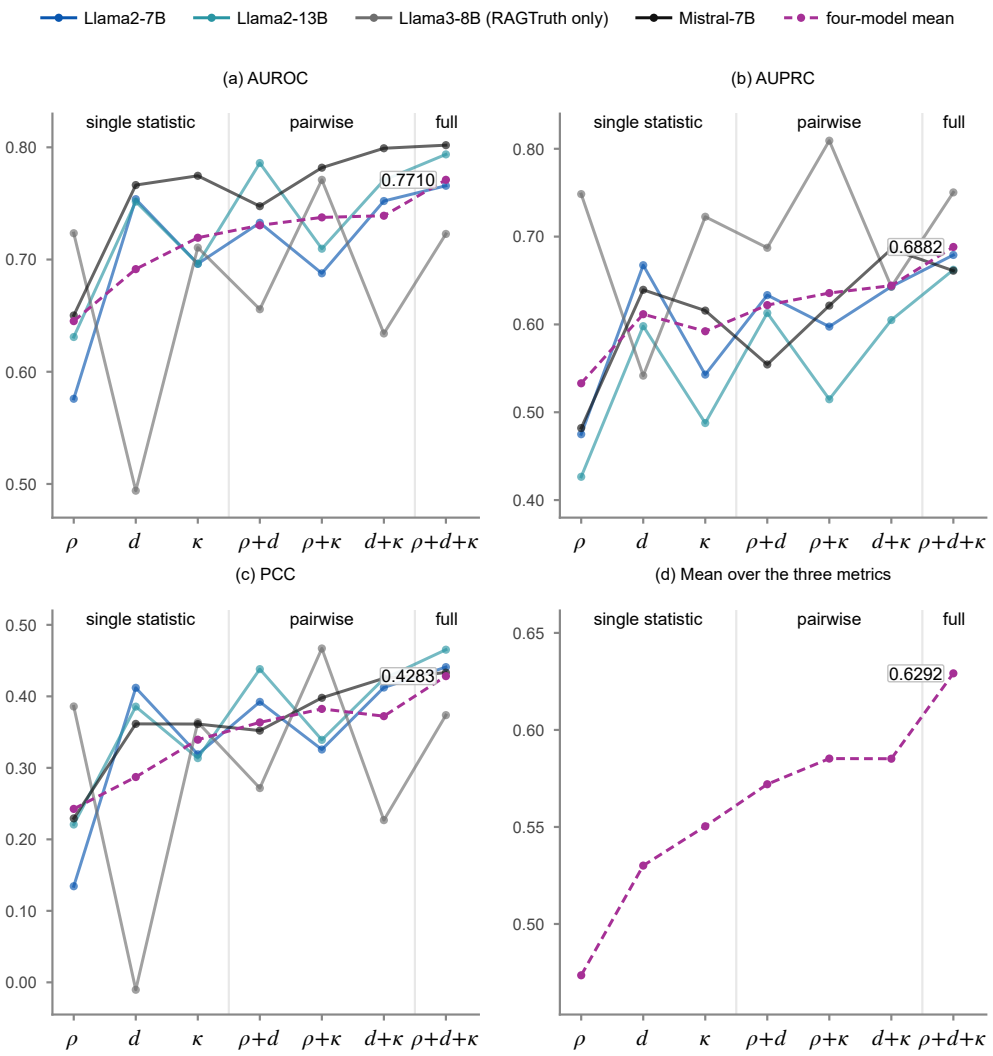


Fig. 4 Impact of the influence statistics. (a)–(c) AUROC, AUPRC, and PCC of each statistic combination. Llama2-7B, Llama2-13B, and Mistral-7B are averaged over RAGTruth and HalluRAG. Llama3-8B is shown on RAGTruth only. The solid lines show the individual LLMs and the magenta dashed line marks the mean of the four LLM curves. (d) The magenta dashed curve of (a)–(c), averaged over the three metrics

Roles of the three influence statistics. CASDET defines influence breadth κ, influence depth d, and influence intensity ρ on the layer-wise results of Eq. (9). The three are combined into the response-level hallucination risk score by the weighted sum of Eq. (16). This subsection examines the three statistics through an ablation over all combinations: each statistic alone, the three pairwise combinations, and the full combination. The hallucination risk score is recomputed over the same layer-wise results on both benchmarks.

Figure 4 plots AUROC, AUPRC, and PCC across the statistic combinations. Detection performance improves consistently as the statistics are added. Averaged over the four LLMs, every pairwise combination outperforms both of its constituent statistics on all three metrics. The full combination ranks first among all combinations on every metric, including panel (d), which averages the three metrics. Every statistic used alone, in contrast, is outperformed by every combination. The full combination also stays within 0.0232 of the best configuration on every metric of both benchmarks. The monotone pattern indicates that the three statistics are complementary and jointly effective, substantiating the combined score of Eq. (16).

The ablation confirms that each design choice is necessary for detection performance. Removing the Attention–FFN cascade or the local control state consistently weakens detection on every LLM, and the full combination of the three statistics performs best across both benchmarks.

4.5 Sensitivity analysis of the weights

Having established that the three statistics are individually useful, we now ask whether their relative weights should be tuned. The three weights of Eq. (16) are the only hyperparameters of CASDET. We discretize the weight simplex in steps of 1/12, yielding 91 nonnegative settings that sum to 1, including equal weights (1/3 each). For each setting, we recompute the score from the same layer-wise results and evaluate AUROC, AUPRC, and PCC on both benchmarks (Figure 5). No additional model forward pass is required. We compare the equal-weight default with the best setting on each benchmark and metric, and test whether the best setting on one benchmark transfers to the other.

Panels 5(a)–(f) average the detection performance over the models within each benchmark. The position of the optimum is unstable. The setting with the highest detection performance changes with the benchmark and the metric. On RAGTruth, the peaks of all three metrics lie where the weight of influence depth is close to zero. On HalluRAG they shift to depth dominance, that is, in the opposite direction. Panels 5(g) and 5(h) give the per-model view of the same scan. Placed in the space of the three metrics, the equal-weight setting of every model lies within the top-performing region. Equal weights, by contrast, stay close to optimal in every case (the peak improvement over equal weights is only 0.0071–0.0537), even though the tuned optima point in no consistent direction.

Cross-benchmark transfer is of greater practical concern. Weights calibrated on one benchmark lose performance on the other. Figure 5(i) quantifies this trade-off. Applying the optimal weights of RAGTruth to HalluRAG decreases the metrics by up to 0.1549, and the reverse transfer decreases them by 0.0355–0.0568. In other words, the gain from tuning the weights is at most about 0.05, while the cost of choosing the wrong calibration benchmark can reach 0.15. Since the optimal weights on the two benchmarks shift in no consistent direction, Eq. (16) retains equal weights. This introduces no hyperparameter that needs to be determined, relies on no parameter-selection data, and avoids the risk of cross-benchmark transfer.

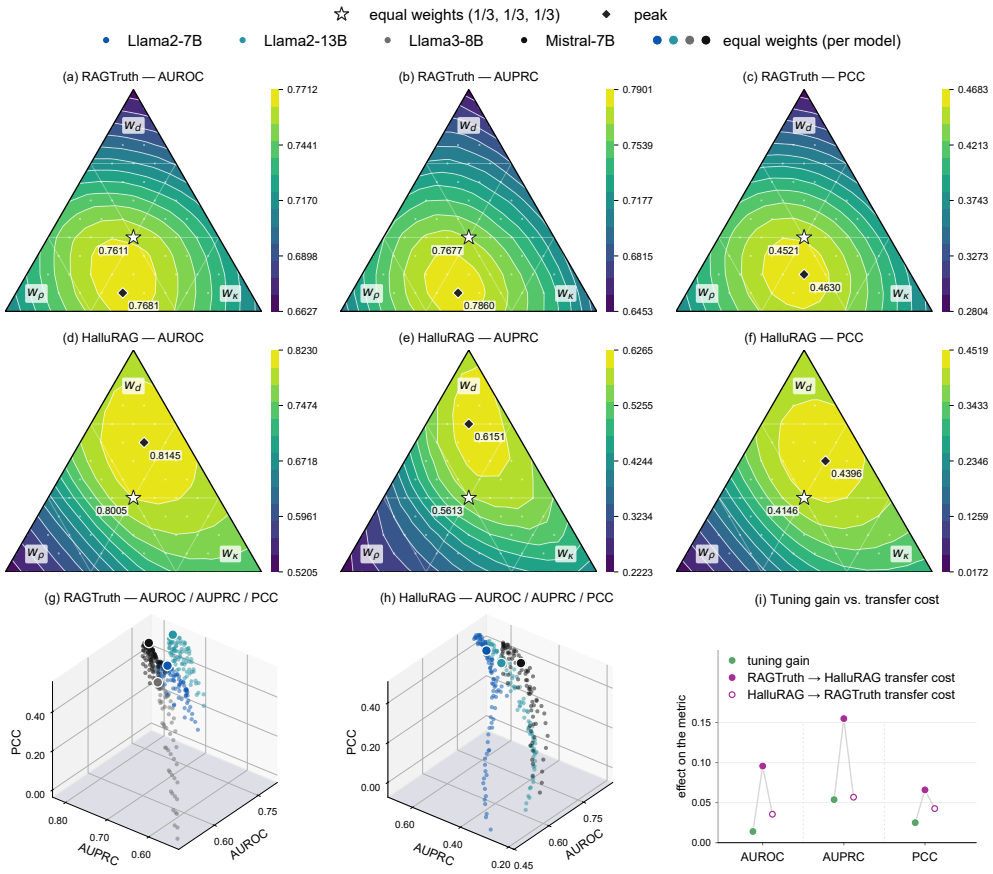


Fig. 5 Detection performance over the weight simplex of influence breadth (κ), influence depth (d), and influence intensity (ρ), where each point is a valid weight setting. (a)–(f) Model-averaged performance, reported per benchmark (rows) and metric (columns). Filled contours show the performance over the 91 scanned settings of the 1/12 grid (small dots). The star marks equal weights and the diamond the peak of the scan. (g)–(h) Per-model view: every scanned setting of each model for RAGTruth and HalluRAG, respectively, placed in the space of its three metric scores. The large dots mark each model at equal weights. (i) Tuning gain versus cross-benchmark transfer cost. For each metric, the green marker gives the gain of the benchmark-specific optima over equal weights (the larger of the two benchmarks), and the magenta markers give the loss when the optimal weights of one benchmark are applied to the other

5 Discussion

This section discusses the conjecture in Section 2. The experiments in Section 4 support it. The hallucination risk score separates non-hallucinated from hallucinated responses, and because the score decreases with influence breadth, influence depth, and influence intensity, this separation means that the influence of external-context introduction on the subsequent FFN provides stronger positive support to the current token in non-hallucinated responses than in hallucinated responses. The conjecture requires only

a systematic difference. The experiments further show that the difference has this direction.

This support rests on the ablation in Subsection 4.4. When only the direct effect of attention is quantified, detection performance decreases on every LLM. When the local control state is removed, detection performance also decreases. The first result shows that the influence through the subsequent FFN is necessary for the separation. The second shows that isolating the current layer's external-context contribution is necessary for this influence to be observable. The separation is therefore attributable to the influence of external-context introduction on the subsequent FFN, not to the presence of external context in the residual stream.

This result distinguishes the influence of external-context introduction on the subsequent FFN from the separate utilization strengths measured by LUMINA and ReDeEP. These methods estimate the utilization strength of external context and parametric knowledge separately, then fuse the two estimates. They do not measure the influence of external-context introduction on the expression and use of parametric knowledge within the same block. The comparison in Subsection 4.3 indicates that this influence provides discriminative information beyond the two utilization strengths. The conjecture therefore concerns an influence that is related to, but distinct from, the separate utilization strengths.

Two questions remain open. The experiments are correlational with respect to hallucination. They show that the influence of external-context introduction on the subsequent FFN differs between non-hallucinated and hallucinated responses, but not that increasing this influence would remove hallucination. Testing the latter would require an intervention on the influence itself. The scope is also limited to the setting of the conjecture. The external context is relevant to the query and contains the answer. Whether this influence can be used for mitigation, and whether it extends to inaccurate or insufficient contexts, are questions for future work.

6 Conclusion

This paper studied Attention–FFN cascade influence in RAG generation and proposed CASDET, a training-free detector for RAG hallucinations. CASDET separates the influence of the current layer's external-context contribution on the subsequent FFN with a local control state, quantifies it with the Logit Lens, and combines influence breadth, depth, and intensity into a response-level hallucination risk score without thresholds or attention-head/layer selection. On RAGTruth and HalluRAG with four open-source LLMs, CASDET is competitive with the stronger baselines, while its detection time is comparable to that of single-forward uncertainty methods and substantially lower than that of the remaining baselines. The ablation shows that removing either the Attention–FFN cascade or the local control state degrades detection on every LLM, and that the three statistics are complementary. These results support the conjecture that the influence of external-context introduction on the expression and use of parametric knowledge differs systematically between non-hallucinated and hallucinated responses. They further indicate that measuring this influence directly provides a practical basis for training-free RAG hallucination detection.

CASDET applies when the external context is relevant to the query and contains the correct information needed to answer it. When the context is inaccurate, irrelevant, or insufficient, CASDET does not guarantee its detection performance. Future work will extend CASDET to these settings, explore the influence signal for hallucination mitigation through decoding-time intervention, and cover multi-turn and multilingual scenarios.

Declarations

Funding. The authors declare that no funding was received for this work.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and consent to participate. Not applicable. This study did not involve human participants, human data, animals, or biological materials requiring ethical approval.

Consent for publication. Not applicable.

Data availability. The datasets used in this study are publicly available: RAGTruth at https://github.com/ParticleMedia/RAGTruth and HalluRAG at https://github.com/F4biian/HalluRAG.

Materials availability. Not applicable.

Code availability. The source code for CASDET is available at https://github.com/Qiyang-Gu/CASDET and archived in the Zenodo repository at https://doi.org/10.5281/zenodo.23468986.

Author contributions. Qiyang Gu conceived and designed the study, curated the data, implemented the code, conducted the experiments, analyzed the results, and wrote the original draft. Fei Zhang and Pengyu Guo conceptualized and supervised the study and revised the manuscript. Min Tang, Fei Wang, and Hui Chen participated in the investigation, validation, and manuscript revision. All authors read and approved the final manuscript.

References

[1] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. 2025;43(2):1-55. https://doi.org/10.1145/3703155.

[2] Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. In: Findings of the Association for Computational Linguistics: EMNLP; 2021. p. 3784-803. https://doi.org/10.18653/v1/2021.findings-emnlp.320.

[3] Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. arXiv preprint arXiv:2312.10997. 2023. https://doi.org/10.48550/arXiv.2312.10997.

[4] Niu C, Wu Y, Zhu J, Xu S, Shum K, Zhong R, et al. 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); 2024. p. 10862-78. https://doi.org/10.18653/v1/2024.acl-long.585.

[5] Magesh V, Surani F, Dahl M, Suzgun M, Manning CD, Ho DE. Hallucination-free? Assessing the reliability of leading AI legal research tools. J Empir Leg Stud. 2025;22(2):216-42. https://doi.org/10.1111/jels.12413.

[6] Xu R, Qi Z, Guo Z, Wang C, Wang H, Zhang Y, et al. Knowledge conflicts for LLMs: a survey. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2024. p. 8541-65. https://doi.org/10.18653/v1/2024.emnlp-main.486.

[7] Sun Z, Zang X, Zheng K, Xu J, Zhang X, Yu W, et al. ReDeEP: detecting hallucination in retrieval-augmented generation via mechanistic interpretability. Proc Int Conf Learn Represent. 2025.

[8] Yeh S, Li S, Mallick T. LUMINA: detecting hallucinations in RAG system with context–knowledge signals. Proc Int Conf Learn Represent. 2026.

[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998-6008.

[10] Geva M, Schuster R, Berant J, Levy O. Transformer feed-forward layers are key-value memories. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2021. p. 5484-95. https://doi.org/10.18653/v1/2021.emnlp-main.446.

[11] Geva M, Caciularu A, Wang K, Goldberg Y. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2022. p. 30-45. https://doi.org/10.18653/v1/2022.emnlp-main.3.

[12] Ren J, Luo J, Zhao Y, Krishna K, Saleh M, Lakshminarayanan B, et al. Out-of-distribution detection and selective generation for conditional language models. Proc Int Conf Learn Represent. 2023.

[13] Malinin A, Gales M. Uncertainty estimation in autoregressive structured prediction. Proc Int Conf Learn Represent. 2021.

[14] Manakul P, Liusie A, Gales M. SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2023. p.

9004-17. https://doi.org/10.18653/v1/2023.emnlp-main.557.

[15] Chen C, Liu K, Chen Z, Gu Y, Wu Y, Tao M, et al. INSIDE: LLMs' internal states retain the power of hallucination detection. Proc Int Conf Learn Represent. 2024.

[16] Es S, James J, Espinosa-Anke L, Schockaert S. 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; 2024. p. 150-58. https://doi.org/10.18653/v1/2024.eacl-demo.16.

[17] Hu X, Ru D, Qiu L, Guo Q, Zhang T, Xu Y, et al. RefChecker: reference-based fine-grained hallucination checker and benchmark for large language models. arXiv preprint arXiv:2405.14486. 2024. https://doi.org/10.48550/arXiv.2405.14486.

[18] Elhage N, Nanda N, Olsson C, Henighan T, Joseph N, Mann B, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread. 2021. Available from: https://transformer-circuits.pub/2021/framework/index.html [cited 2026 Oct 12].

[19] Ferrando J, Sarti G, Bisazza A, Costa-jussà MR. A primer on the inner workings of transformer-based language models. arXiv preprint arXiv:2405.00208. 2024. https://doi.org/10.48550/arXiv.2405.00208.

[20] Kobayashi G, Kuribayashi T, Yokoi S, Inui K. Incorporating residual and normalization layers into analysis of masked language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2021. p. 4547-68. https://doi.org/10.18653/v1/2021.emnlp-main.373.

[21] Vig J, Gehrmann S, Belinkov Y, Qian S, Nevo D, Singer Y, et al. Investigating gender bias in language models using causal mediation analysis. Adv Neural Inf Process Syst. 2020;33:12388-401.

[22] Meng K, Bau D, Andonian A, Belinkov Y. Locating and editing factual associations in GPT. Adv Neural Inf Process Syst. 2022;35:17359-72. https://doi.org/10.52202/068431-1262.

[23] Zhang F, Nanda N. Towards best practices of activation patching in language models: metrics and methods. Proc Int Conf Learn Represent. 2024.

[24] Wu K, Wu E, Zou J. ClashEval: quantifying the tug-of-war between an LLM's internal prior and external evidence. Adv Neural Inf Process Syst. 2024;37:33402-22. Datasets and Benchmarks Track. https://doi.org/10.52202/079017-1053.

[25] nostalgebraist. Interpreting GPT: the logit lens. LessWrong. 2020. Available from: https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens [cited 2026 Oct 12].

[26] Belrose N, Ostrovsky I, McKinney L, Furman Z, Smith L, Halawi D, et al. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112. 2023. https://doi.org/10.48550/arXiv.2303.08112.

[27] Page EB. Ordered hypotheses for multiple treatments: a significance test for linear ranks. J Am Stat Assoc. 1963;58(301):216-30. https://doi.org/10.1080/01621459.1963.10500843.

[28] Azaria A, Mitchell T. The internal state of an LLM knows when it's lying. In: Findings of the Association for Computational Linguistics: EMNLP; 2023. p. 967-76. https://doi.org/10.18653/v1/2023.findings-emnlp.68.

[29] Chuang YS, Xie Y, Luo H, Kim Y, Glass JR, He P. DoLa: decoding by contrasting layers improves factuality in large language models. Proc Int Conf Learn Represent. 2024.

[30] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288. 2023. https://doi.org/10.48550/arXiv.2307.09288.

[31] Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783. 2024. https://doi.org/10.48550/arXiv.2407.21783.

[32] Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de las Casas D, et al. Mistral 7B. arXiv preprint arXiv:2310.06825. 2023. https://doi.org/10.48550/arXiv.2310.06825.

[33] Ridder F, Schilling M. The HalluRAG dataset: detecting closed-domain hallucinations in RAG applications using an LLM's internal states. arXiv preprint arXiv:2412.17056. 2024. https://doi.org/10.48550/arXiv.2412.17056.

[34] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432.