SPOT THE DIFFERENCE: DETECTING HALLUCINATIONS THROUGH CONTEXT-HEAD INTERVENTION AND SPARSE AUTOENCODER FEATURE CONTRASTS

Anonymous authors
Paper under double-blind review

ABSTRACT

RAG aims to enhance factuality in LLMs, yet these models frequently generate unfaithful outputs that contradict or exceed the provided context. While the internal representations of the LLM often contain features sufficient to flag hallucinations, the additive nature of the residual stream allows ingrained parametric knowledge to progressively outweigh contextual information during inference. In this work, we investigate this internal conflict, termed the *Knowledge-Context Gap* (KCG), in the sparse autoencoder (SAE) feature space. We show that jointly modeling paired SAE features provides complementary signals for hallucination detection, though constructing the paired representations via prompt-level manipulation incurs computational overhead and representational drift. To address these challenges, we propose RAGap a mechanistic framework that characterizes KCG by isolating the LLM's parametric knowledge directly within the residual stream. By identifying and suppressing specialized *context heads* through differentiable masking, RAGap enables an efficient characterization of the latent divergence between parametric knowledge and contextual information. Leveraging SAEs for interpretable feature extraction, our method consistently outperforms state-of-the-art hallucination detectors across multiple benchmarks. Our findings provide a novel mechanistic perspective on RAG faithfulness and offer a practical solution for reliable, inference-time hallucination detection. Our code is publicly available at `https://anonymous.4open.science/r/ragap-5062`.

1 INTRODUCTION

Retrieval-augmented generation (RAG) improves the factuality of LLMs by anchoring responses in an external context (Guu et al., 2020). However, models frequently produce unfaithful outputs by contradicting sources or adding unsupported details (Magesh et al., 2025; Gao et al., 2023; Maynez et al., 2020; Rahman et al., 2025), which limits their utility in critical domains (Huang et al., 2025; Zakka et al., 2024). While supervised detectors (Bao et al., 2024; Tang et al., 2024a) and LLM-as-a-judge frameworks (Zheng et al., 2023; Li et al., 2024a) have been explored to mitigate these hallucinations, they are often constrained by heavy data dependency, high computational costs, or evaluative blind spots.

Previous studies have examined how contextual information and parametric knowledge accumulate in the residual stream through attention and feed-forward updates (Zhao et al., 2025; Sun et al., 2025). Later feed-forward updates can allow parametric knowledge to outweigh contextual evidence, leading to unfaithful generation (Ma et al., 2026). This motivates examining internal activations for hallucination detection (Han et al., 2024; Sun et al., 2025; Zhou et al., 2025), particularly signals reflecting the interaction between contextual information and parametric knowledge.

Existing approaches characterize this interaction through different detection signals. ReDeEP (Sun et al., 2025) estimates external-context and parametric-knowledge utilization through attention-based semantic alignment and distributional changes across feed-forward networks. LUMINA (Yeh et al., 2026) measures sensitivity to context replacement and tracks the evolution of token predictions across layers. Both approaches summarize knowledge-context interactions through utilization scores. To extract feature-level signals from high-dimensional activations, RAGLens (Xiong et al., 2026) employs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of representative hallucination detectors by signal extraction and knowledge–context interaction modeling. Interventions refer to detection-time operations.

Method	Input	Intervention	Readout	Interaction modeling
ReDeEP	Unchanged	None	Attention/FFN	Utilization scores
LUMINA	Modified	Context replacement	Token distributions	Input/layer contrasts
RAGLens	Unchanged	None	SAE activations	Not explicitly modeled
RAGap	Unchanged	**Context-head masking**	**Paired** SAE activations	**SAE feature interactions**

sparse autoencoders (SAEs) to decompose internal states into interpretable features. However, it extracts SAE features from the overall hidden representations without explicitly modeling knowledge-context interactions. Table 1 compares representative methods in terms of signal extraction and knowledge-context interaction modeling.

In this work, we use SAE-based feature extraction to construct the conflict between contextual information and parametric knowledge, which we term the *Knowledge-Context Gap* (KCG). We model this gap by combining the encoded SAE features from context-conditioned and context-free states with their differences and interactions. However, removing context changes the computation throughout the network and may alter the parametric knowledge activated by the original input.

To address this issue, we propose **RAGap**, which constructs paired context-conditioned and context-suppressed states by locally suppressing context-routing heads identified through differentiable masking. Retaining the original input and reusing pre-intervention computation limits representational drift and avoids an additional full forward pass. Experiments demonstrate effective hallucination detection across multiple benchmarks and lower latency than prompt-level context removal.

Our main contributions are as follows:

- We characterize KCG through paired SAE features and show that jointly modeling context-conditioned and context-free representations provides complementary detection signals beyond either representation alone.
- We analyze prompt-level construction of paired representations and identify two limitations: the computational cost of an additional forward pass and representational drift caused by context removal.
- We propose **RAGap**, which constructs paired context-conditioned and context-suppressed states through localized context-head intervention and jointly models their SAE feature interactions for efficient hallucination detection.

2 Related Work

RAG Faithfulness & Hallucination Detection. RAG enhances LLM factuality by grounding responses in external context (Lewis et al., 2020; Guu et al., 2020; Shuster et al., 2021; Wei et al., 2025). However, faithfulness errors persist when models contradict or exceed provided evidence (Niu et al., 2024; Sun et al., 2025; Zhang et al., 2025; Elchafei & Abu-Elkheir, 2025). To address these, early researchers utilized fine-tuned classifiers (Manakul et al., 2023; Tang et al., 2024a) or LLM-as-a-judge frameworks (Zheng et al., 2023; Turpin et al., 2023). These approaches often involve high computational costs or lack transparency regarding the model's internal decision process. While internal representations offer a more direct alternative, they are frequently limited by the opacity of hidden states (Elchafei & Abu-Elkheir, 2025). Consequently, sparse autoencoders have emerged to extract interpretable features through sparsity-inducing bottlenecks, providing a more transparent signal for hallucination detection (Bricken et al., 2023; Shu et al., 2025; Shi et al., 2025).

Mechanistic Interpretability & Knowledge Conflicts Mechanistic interpretability reverse-engineers Transformer computations into human-understandable algorithms (Elhage et al., 2021; Wang et al., 2023). This framework has enabled researchers to locate factual associations in MLP layers (Meng et al., 2023) and leverage attention patterns to elicit truthful generations (Li et al.,

2024b). A critical challenge identified in recent surveys is the knowledge conflict between parametric knowledge and contextual information (Xu et al., 2024a). Evidence suggests that Transformers employ specialized, non-overlapping subsets of attention heads to route these streams, where context heads process external prompts and parametric circuits handle internal memory (Zhao et al., 2025). Within this paradigm, hallucinations are viewed as the byproduct of unresolved representational conflicts occurring in the residual stream (Suresh et al., 2025).

3 KNOWLEDGE-CONTEXT GAP

We present experiments in this section to characterize the *Knowledge-Context Gap* (KCG) in the SAE feature space. The results show that jointly modeling context-conditioned and context-free SAE features provides complementary signals for hallucination detection.

3.1 EXPERIMENTAL SETUP

Data. To analyze the KCG across various RAG scenarios, we utilize two representative benchmarks: RAGTruth (Niu et al., 2024) and TofuEval (Tang et al., 2024b). RAGTruth includes three subtasks: summarization (Summary), data-to-text generation (Data2txt), and question answering (QA). Formally, we represent each original data instance as a tuple $\mathcal{I}^+ = (\mathcal{C}, q, y, \ell)$, where $\mathcal{C}$ denotes the retrieved context, q is the query, y is the model-generated response, and $\ell \in \{0, 1\}$ is the faithfulness label. To isolate the model's parametric knowledge from the influence of contextual information, we construct a counterpart $\mathcal{I}^- = (\emptyset, q, y, \ell)$ by directly removing the retrieved passages $\mathcal{C}$ from the input prompt. Dataset details are provided in Appendix B.1.

Feature Extraction Inspired by Xiong et al. (2026), we define a pipeline that applies to both $\mathcal{I}^+$ and $\mathcal{I}^-$. For simplicity, we omit the superscripts as both $\mathcal{I}^+$ and $\mathcal{I}^-$ follow the same pipeline. Formally, for each instance $\mathcal{I}$, we denote $\Phi_l(\cdot)$ as the mapping that returns the layer-l hidden states. We extract the hidden states for each token y_t in the generated response y as $h_t = \Phi_l(y_t, q, \mathcal{C})$. These hidden states are transformed via a pretrained SAE encoder $\mathcal{E}$ into sparse features. By applying channel-wise max pooling over all token activations, we obtain the instance-level representation:

$$\bar{h} = [\max_t \mathcal{E}(h_t)_1, \ldots, \max_t \mathcal{E}(h_t)_K] \in \mathbb{R}^K. \tag{1}$$

To identify the most relevant signals, we quantify the mutual information (MI) between each pooled feature $\bar{h}_k$ ($k = 1, \ldots, K$) and the faithfulness label ℓ:

$$I(\bar{h}_k; \ell) = \int_{\mathbb{R}} \sum_{\ell \in \{0,1\}} p(\bar{h}_k, \ell) \log_2 \frac{p(\bar{h}_k, \ell)}{p(\bar{h}_k)p(\ell)} d\bar{h}_k. \tag{2}$$

We then rank features by MI to select the top K' dimensions, whose indices form the set:

$$\mathcal{S} = \arg\max_{|\mathcal{S}|=K'} \sum_{k \in \mathcal{S}} I(\bar{h}_k; \ell), \tag{3}$$

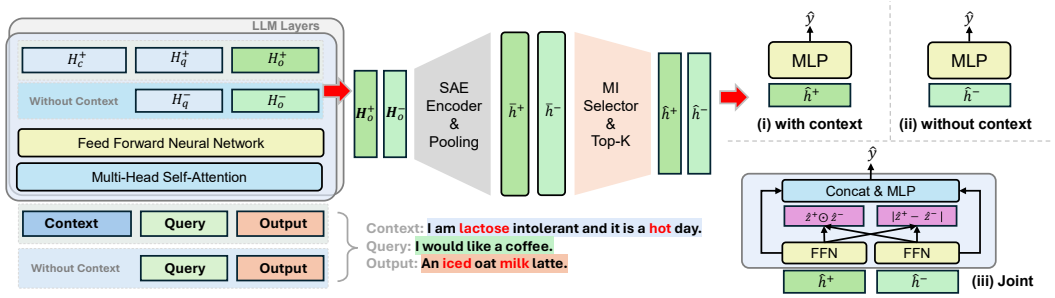
yielding the final feature vector $\hat{h} \in \mathbb{R}^{K'}$ as the subvector of $\bar{h}$ restricted to indices $\mathcal{S}$. This standardized pipeline ensures that our characterization of KCG is based on semantically meaningful activations that implicitly capture interactions between the generated answer and the retrieved content.

Hallucination Detection Following the objective of characterizing KCG as a pivotal indicator for hallucination, we design three contrastive detection settings: *With-Context*, *Without-Context*, and *Joint*. We define a unified encoder $\Psi_\omega(x) = \rho(Wx + b)$ with activation $\rho(\cdot)$. For the two single-input settings, the detection model f_θ is formulated as:

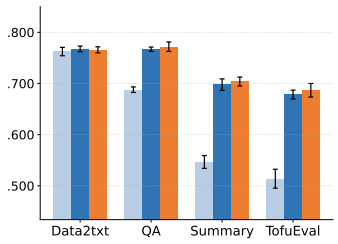
$$\hat{\ell} = \sigma(g_\phi(\Psi_\omega(\hat{h}))), \quad \hat{h} \in \{\hat{h}^+, \hat{h}^-\} \tag{4}$$

where $\hat{h}^+$ and $\hat{h}^-$ are the SAE features from $\mathcal{I}^+$ and $\mathcal{I}^-$, respectively, $g_\phi(\cdot)$ is a lightweight MLP classifier, and $\sigma(\cdot)$ is the sigmoid function. To explicitly characterize the interaction between internal and external signals, the *Joint* setting employs dual encoders Ψ_ω^+ and Ψ_ω^- to construct a fused representation $\mathbf{z}_{\text{joint}} \in \mathbb{R}^{4D}$:

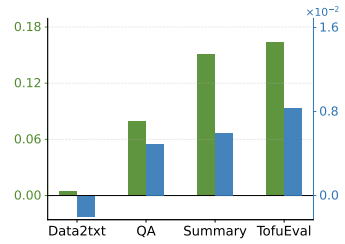
$$\mathbf{z}_{\text{joint}} = [z^+, z^-, |z^+ - z^-|, z^+ \odot z^-] \tag{5}$$



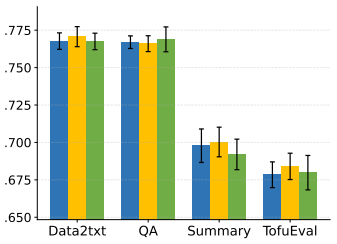
(a) An overview of the experimental setup. **(i) with context**: conditioned on contextual information $\mathcal{C}$; **(ii) without context**: reflecting parametric knowledge where $\mathcal{C} = \emptyset$; and **(iii) Joint**: explicitly capturing the KCG. We employ macro F_1 as the primary metric to evaluate hallucination detection performance.



(b) The F_1 scores are employed as a proxy to evaluate the discriminative capability of features under all settings. **Observation:** Across all benchmarks, the *joint* setting consistently achieves the highest performance. While scores are nearly identical for *Data2txt*, the *joint* approach significantly outperforms the *w/ ctx.* baseline in other tasks **where the KCG is more pronounced.**



(c) Green and blue bars represent the performance gap between *w/ ctx.* and *w/o ctx.* and the relative gain of the *joint* setting over the *w/ ctx.* baseline, respectively. **Observation:** Across all evaluated benchmarks, a more negative performance gap is consistently accompanied by a larger improvement in hallucination detection under the *Joint* setting.



(d) The F_1 scores compare the *w/ ctx.* baseline against variants that incorporate interaction signals: the absolute difference between features and the element-wise product of features. **Observation:** Across all four evaluated benchmarks the addition of either the absolute difference or the element-wise product consistently leads to a higher F_1 score compared to the *w/ ctx.* baseline.

Figure 1: Experimental results for **Llama-2-7B**. Findings are reported as the mean across 10 independent trials with distinct random seeds, ensuring high reproducibility and statistical stability. The results support the value of jointly modeling paired SAE features for hallucination detection across the evaluated RAG subtasks.

where $z^+ = \Psi_\omega^+(\hat{h}^+)$ and $z^- = \Psi_\omega^-(\hat{h}^-)$. The final prediction is derived as $\hat{\ell}_{joint} = \sigma(g_\phi(\mathbf{z}_{\text{joint}}))$. This formulation enables the classifier to directly sense the divergence between the model's internal belief and its context-conditioned output. All models are optimized via binary cross-entropy loss $\mathcal{L} = -\mathbb{E}[\ell \log \hat{\ell} + (1-\ell)\log(1-\hat{\ell})]$ and evaluated using macro F_1. Implementation details are provided in Appendix C.

Language Models We experiment with Qwen3-0.6B/4B (Yang et al., 2025) and 3 LLaMA models (Touvron et al., 2023): Llama-3.2-1B, Llama-3-8B, Llama-2-7B.

3.2 Experimental Results

Figure 1 illustrates the hallucination detection performance across various RAG subtasks. Figure 4 in Appendix D shows consistent patterns across additional language models.

Finding 1: KCG is a significant signal for faithfulness. The *Joint* detection setting consistently achieves the highest macro F_1 scores across nearly all evaluated RAG subtasks, surpassing both the *With-Context* and *Without-Context* baselines. While the magnitude of the improvement varies across

tasks, the consistent gains indicate that the interaction between the model's internal parametric knowledge and the external contextual information provides complementary discriminative information beyond either source alone.

Finding 2: KCG efficacy is contingent on contextual reliance. The performance gain of the *Joint* setting is linked to the performance disparity between the *With-Context* and *Without-Context* configurations. In tasks such as *Data2txt* where this gap is minimal, the *Joint* model yields negligible improvement over the *With-Context* baseline, as the context-free signal offers little complementary evidence when the model's behavior is largely invariant to context ablation. Conversely, as the discrepancy between the two settings becomes more pronounced, the gain from *Joint* modeling also increases. These results indicate that the diagnostic value of KCG is derived from capturing the shift in model behavior when contextual information is ablated. This suggests that the signal becomes most informative when the model exhibits a strong reliance on context, allowing the *Joint* framework to sense the latent conflict between ingrained parametric knowledge and external evidence.

3.3 The Limitations of Prompt-level Manipulation

While the results validate KCG as a fundamental indicator of faithfulness, the current characterization relies on prompt-level manipulation to contrast *With-Context* and *Without-Context* states. This approach faces two primary limitations:

Computational Inefficiency Characterizing KCG requires obtaining both $\hat{h}^+$ and $\hat{h}^-$, necessitating two independent forward passes for each detection. This redundant inference increases latency and resource consumption, which is often prohibitive for real-time RAG deployment.

Representational Drift FFN layers are often interpreted as a linear associative memory, functioning as key-value storage for information retrieval (Geva et al., 2021), where parametric knowledge $\boldsymbol{v}$ is retrieved via input keys $\boldsymbol{k}$ triggered by the pre-FFN residual state $\boldsymbol{z}^l$:

$$\boldsymbol{h}^l = \boldsymbol{z}^l + \underbrace{\boldsymbol{W}^l_{\text{out}} \overbrace{\sigma(\boldsymbol{W}^l_{\text{in}}\gamma(\boldsymbol{z}^l))}^{\boldsymbol{k}}}_{\boldsymbol{v}}. \tag{6}$$

Explicitly removing context tokens $\mathcal{C}$ drastically alters z^l, which can prevent the precise activation of the keys k originally primed by the full prompt. Such representational drift indicates that $\hat{h}^-$ deviates from the internal parametric knowledge that is actually active during context-conditioned generation. Appendix A provides a formal analysis of this representational drift.

4 RAGap: Mechanistic Characterization of KCG

In this section, we present **RAGap**, a framework designed to characterize KCG. We first establish the role of specific attention heads in routing external contextual information (Section 4.1). We then describe our methodology for identifying these heads using a differentiable masking approach (Section 4.2). Finally, we outline the integrated framework for utilizing these internal signals to detect faithfulness (Section 4.3).

4.1 Contextual Information in the Residual Stream

KCG is at the core of our work. Here we first introduce how contextual information is integrated into the LLM's internal representations.

In the Transformer architecture, the residual stream serves as a superposition of parametric knowledge and contextual information (Sun et al., 2025). Recent studies (Jin et al., 2024; Niu et al., 2025) indicate that context utilization is mediated by a specialized subset of attention heads, known as *context heads*. Let $\boldsymbol{h}^{l-1} \in \mathbb{R}^d$ be the output of the previous layer. The sequential update at layer l can be decomposed by partitioning the attention heads into the set of context heads $\mathcal{S}_{ctx}$ and the

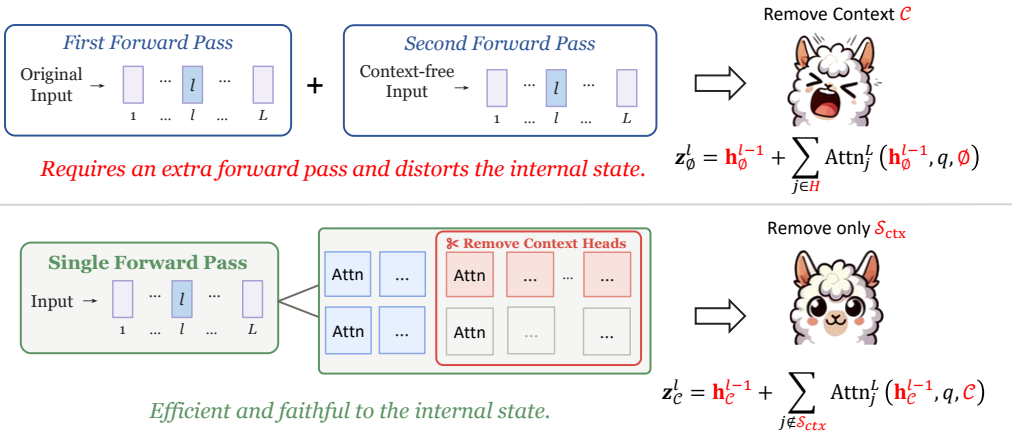


Figure 2: Comparison between RAGap and prompt-level manipulation. RAGap is both efficient and faithful to the internal state.

remaining heads:

$$\boldsymbol{z}^l = \boldsymbol{h}^{l-1} + \boldsymbol{c}^l + \sum_{j \notin \mathcal{S}_{ctx}} \text{Attn}_j^l(\boldsymbol{h}^{l-1}), \quad \boldsymbol{h}^l = \boldsymbol{z}^l + \text{FFN}(\boldsymbol{z}^l), \tag{7}$$

where $\boldsymbol{c}^l = \sum_{j \in \mathcal{S}_{ctx}} \text{Attn}_j^l(\boldsymbol{h}^{l-1})$ denotes the contribution of context heads. An effective approach to overcoming the limitations discussed in Section 3.3 is **to make the contextual information $\boldsymbol{c}^l$ disappear within the residual stream.** This localized intervention ensures that the activation manifold required for FFN-based knowledge retrieval (Eq. 6) remains intact, allowing for an efficient characterization of KCG without the collateral damage of prompt-level manipulation.

Therefore, we identify the functional circuit of context heads and apply a targeted masking mechanism to suppress $\boldsymbol{c}^l$. We will detail the methodology for discovering these context heads in the next subsection.

4.2 Identifying Context Heads via Differentiable Masking

Differentiable Masking Inspired by prior work (Yu et al., 2025; Niu et al., 2025), we employ a gradient-based discovery approach that selectively activates or deactivates each attention head across all layers via a learnable binary mask $m_j \in \{0, 1\}$. To ensure differentiability, we utilize the Gumbel-Sigmoid distribution with a straight-through estimator (STE) (Bengio et al., 2013). Let $\text{sg}[\cdot]$ denote the stop-gradient operator. The continuous latent variable s_j and the discrete mask m_j are defined as:

$$s_j = \sigma\left(\frac{l_j - \log \frac{\log \mathcal{U}_1}{\log \mathcal{U}_2}}{\tau}\right), \quad m_j = \text{sg}[\mathbb{I}(s_j > 0.5) - s_j] + s_j, \tag{8}$$

where l_j is the learnable logit and $\mathcal{U}_1, \mathcal{U}_2 \sim \text{Uniform}(0, 1)$ introduce stochasticity for exploration.

Data Construction We construct discovery samples from factual triplets $f = (s, r, o)$ sourced from the LRE dataset (Hernandez et al., 2024). For each sample, we select two distinct facts $f_1 = (s_1, r, o_1)$ and $f_2 = (s_2, r, o_2)$ where $s_1 \neq s_2$. The model is provided with an input $\mathbf{x} = \langle \mathcal{C}, q \rangle$, where the context $\mathcal{C}$ is a statement of f_1 and the query q is a prompt for f_2. We define y_c and y_r as the first differing tokens of o_1 and o_2. Details of the LRE dataset are provided in Appendix B.2.

Optimization Objective Our objective is to identify context heads whose activation promotes the contextual information of y_c while overshadowing the parametric knowledge y_r. This is achieved by minimizing the following discovery loss $\mathcal{L}_{disc}$:

$$\mathcal{L}_{disc} = \hat{\ell}(y_r) - \hat{\ell}(y_c) + \lambda \cdot \frac{1}{|H|} \sum_{j=1}^{|H|} \sigma(l_j), \tag{9}$$

where $\hat{\ell}(y)$ represents the log-probability of token y under the masked model. Minimizing Eq. 9 forces the optimizer to retain masks ($m_j \approx 1$) only for the essential heads that facilitate the transmission of contextual information into the residual stream.

4.3 Characterizing KCG via Masked Intervention

To characterize the KCG, RAGap performs a targeted mechanistic intervention using the masks optimized in Section 4.2. Specifically, we derive the context-suppressed representation $\hat{h}^- = \text{SAE}(\hat{\Phi}_l^-(y_t, q, \mathcal{C}))$, where $\hat{\Phi}_l^-$ denotes the mapping that returns the layer-l hidden states under the intervention that mutes context heads ($m_j = 0, \forall j \in \mathcal{S}_{ctx}$). By integrating this high-fidelity signal into the joint detection framework (Eq. 5), the classifier identifies hallucinations by sensing the latent divergence between the standard context-conditioned state $\hat{h}^+$ and the mechanistically suppressed state $\hat{h}^-$. Figure 2 shows the difference between RAGap and prompt-level manipulation.

4.4 Inference-time Complexity

We evaluate the inference cost required to obtain representations prior to SAE encoding. Let $L_+ = |\mathcal{C}| + |q| + |y|$ and $L_- = |q| + |y|$ be the sequence lengths. The complexity of a full forward pass is defined as $\mathcal{T}_{\text{fwd}}(L) = \mathcal{O}(N(L^2 d + L d d_{\text{ff}}))$, where N, d, and d_{ff} denote the number of layers, hidden size, and FFN dimension, respectively. Prompt-level manipulation requires an additional full pass $\mathcal{T}_{\text{fwd}}(L_-) = \mathcal{O}(N(L_-^2 d + L_- d d_{\text{ff}}))$. In contrast, RAGap leverages the conditioned pass $\mathcal{T}_{\text{fwd}}(L_+)$, which is the base inference cost already incurred by the RAG system during normal generation. By applying static masks at the intervention layer, RAGap avoids redundant N-layer computation and only necessitates a single-layer FFN re-computation:

$$\mathcal{T}_{RAGap} = \mathcal{T}_{\text{fwd}}(L_+) + \mathcal{O}(L_+ d d_{\text{ff}}). \tag{10}$$

Since the additional overhead is localized to one layer rather than N layers, RAGap achieves significantly lower latency than prompt-level manipulation.

5 Experiment

5.1 Experimental Settings

To evaluate the robustness of RAGap across diverse RAG scenarios, we utilize two categories of benchmarks. Primary evaluation is conducted on RAGTruth (Niu et al., 2024) and Dolly (Accurate Context) (Hu et al., 2024), which provide human-annotated faithfulness labels for outputs generated by Llama2-7B and 13B. To further assess the architectural generalizability of our mechanistic signals, we extend our analysis to AggreFact (Tang et al., 2023) and TofuEval (Tang et al., 2024b). Following standard evaluation protocols in the field (Sun et al., 2025; Xiong et al., 2026), we employ balanced accuracy (Acc) and macro F_1 as our primary metrics.

We compare RAGap with representative detectors based on (i) prompt engineering (e.g., Friel & Sanyal, 2023), (ii) model uncertainty (e.g., Manakul et al., 2023), or (iii) LLM internal representations (e.g., Sun et al., 2025). We also include a fine-tuning baseline, "Llama2-13B(LR)", following existing work (Sun et al., 2025). For a fair comparison with methods that analyze internal signals during LLM generation, we evaluate all detectors on Llama2-7B and Llama2-13B using the corresponding samples generated by these models in RAGTruth and Dolly. Notably, as both RAGap and the relevant representation-based baselines (e.g., RAGLens) utilize SAE, we reproduce these results using identical SAE architectures and pretrained weights. Table 2 lists the specific baselines, with details in Appendix B. Implementation details are provided in Appendix C.

5.2 Performance on RAG Hallucination Detection

Table 2 summarizes the performance of RAGap on RAGTruth and Dolly under the Llama2-7B/13B backbones. RAGap consistently achieves the best overall performance across datasets, backbones, and evaluation metrics. Compared with prompting-based, uncertainty-based, and internal representation-based detectors, RAGap shows clear advantages in both ranking-oriented and classification-oriented metrics. Notably, it also outperforms the strongest prior internal-representation baseline, RAGLens,

Table 2: Performance comparison of different hallucination detection methods on RAGTruth and Dolly. The best results are highlighted in **bold**.

Method	RAGTruth (Llama2-7B)			Dolly (Llama2-7B)			RAGTruth (Llama2-13B)			Dolly (Llama2-13B)		
	AUC	Acc	F_1	AUC	Acc	F_1	AUC	Acc	F_1	AUC	Acc	F_1
Prompt	–	0.6700	0.6720	–	0.6200	0.5476	–	0.7300	0.6899	–	0.6700	0.5823
Llama2-13B(LR)	–	0.6350	0.6572	–	0.6043	0.6616	–	0.7044	0.6725	–	0.5545	0.6664
LMvLM	–	0.6940	0.7365	–	0.6550	0.7702	–	0.5956	0.7684	–	0.6800	0.7000
FAcTScore	0.5428	0.5333	0.6719	0.4813	0.5354	0.6849	0.5294	0.4533	0.6239	0.4389	0.4646	0.5954
FactCC	0.4976	0.5022	0.4589	0.6169	0.5758	0.5882	0.4753	0.4800	0.4121	0.6496	0.6162	0.5250
ChainPoll	0.6738	0.6841	0.7006	0.6593	0.6200	0.5581	0.7414	0.7378	0.7370	0.7070	0.6800	0.6004
RAGAS	0.7290	0.6822	0.6667	0.6648	0.6560	0.6392	0.7541	0.7080	0.6987	0.6412	0.6480	0.5306
TruLens	0.6510	0.6821	0.6658	0.6264	0.6800	0.6567	0.7073	0.6756	0.7063	0.6622	0.5700	0.3944
RefCheck	0.6912	0.6467	0.6736	0.6494	0.6100	0.5412	0.7897	0.7200	0.7823	0.6621	0.5700	0.3944
P(True)	0.7093	0.5648	0.6549	0.6191	0.5344	0.5095	0.8496	0.6266	0.7038	0.6422	0.5260	0.5240
SelfCheckGPT	–	0.5844	0.4642	–	0.5300	0.3188	–	0.5844	0.4642	–	0.5300	0.3188
LN-Entropy	0.5912	0.5620	0.6850	0.6074	0.5656	0.6261	0.5912	0.5620	0.6850	0.6074	0.5656	0.6261
Energy	0.5619	0.5088	0.6657	0.6074	0.5656	0.6261	0.5619	0.5088	0.6657	0.6074	0.5656	0.6261
Focus	0.6233	0.5533	0.6522	0.6783	0.6212	0.6545	0.7888	0.6000	0.6758	0.7067	0.6500	0.6567
Perplexity	0.5091	0.5333	0.6749	0.6825	0.6363	0.7097	0.5091	0.5333	0.6749	0.6825	0.6363	0.7097
EigenScore	0.6045	0.5422	0.6682	0.6786	0.6596	0.7241	0.6640	0.5267	0.6637	0.7214	0.6211	0.7200
SEP	0.7143	0.6187	0.7048	0.6067	0.6060	0.7023	0.8098	0.7288	0.7799	0.7093	0.6800	0.6923
SAPLMA	0.7107	0.5155	0.6502	0.6500	0.6084	0.6653	0.8029	0.5488	0.6923	0.7088	0.6100	0.6605
ITI	0.6714	0.5667	0.6496	0.5494	0.5800	0.6281	0.8501	0.6177	0.6850	0.6530	0.5583	0.6712
ReDeEP	0.7458	0.6822	0.7190	0.7949	0.7373	0.7833	0.8244	0.7889	0.7587	0.8420	0.7070	0.7603
RAGLens	0.8413	0.7576	0.7636	0.8764	0.7778	0.8070	0.8624	0.7831	0.7828	0.9188	0.8500	0.8496
RAGap (Ours)	**0.8810**	**0.7996**	**0.8004**	**0.9308**	**0.8682**	**0.8677**	**0.8677**	**0.7895**	**0.7908**	**0.9309**	**0.8578**	**0.8572**

Table 3: Ablation study of RAGap on Llama-3-8B. We report performance and relative drop compared to the full model.

Method	RAGTruth			AggreFact			TofuEval		
	AUC	Acc	F_1	AUC	Acc	F_1	AUC	Acc	F_1
Full	**.850**	**.772**	**.775**	**.777**	**.679**	**.685**	**.716**	**.633**	**.636**
w/o $\mathcal{S}_{ctx}$	.847 (-.003)	.768 (-.004)	.769 (-.006)	.769 (-.008)	.670 (-.009)	.677 (-.008)	.710 (-.006)	.632 (-.001)	.632 (-.004)
w/o z^+	.842 (-.008)	.753 (-.019)	.755 (-.020)	.764 (-.013)	.665 (-.014)	.672 (-.013)	.704 (-.012)	.605 (-.028)	.607 (-.029)
w/o z^-	.846 (-.005)	.759 (-.013)	.765 (-.010)	.774 (-.003)	.675 (-.004)	.681 (-.004)	.709 (-.007)	.618 (-.015)	.619 (-.017)
w/o KCG	.845 (-.005)	.760 (-.012)	.765 (-.010)	.777 (-.000)	.670 (-.009)	.676 (-.009)	.711 (-.005)	.622 (-.012)	.624 (-.012)

demonstrating that explicitly characterizing the KCG yields a more discriminative signal for RAG hallucination detection than using standard hidden representations alone.

To further evaluate whether the improvement generalizes beyond Llama2 backbones, Figure 3 compares RAGap with RAGLens across four models and three datasets. RAGap consistently improves both Acc and F_1 on RAGTruth, AggreFact, and TofuEval. While the magnitude of the improvement varies across models and datasets, the consistent gains suggest that jointly modeling standard and context-suppressed SAE features provides complementary detection signals beyond those captured by RAGLens from a single context-conditioned state. These results support the effectiveness of explicitly characterizing the KCG across different backbone families, model scales, and hallucination detection datasets.

5.3 Ablation Studies

To assess the effectiveness of each component and the necessity of mechanistic intervention in RAGap, we conducted a series of ablation studies. We evaluate four variants: (1) **w/o** $\mathcal{S}_{ctx}$, which replaces the targeted masking of the context head set $\mathcal{S}_{ctx}$ with masking all attention heads in the intervention layer; (2) **w/o** z^+ and (3) **w/o** z^-, which set the standard contextualized representation and the mechanistically suppressed representation, respectively, to zero within the joint vector in Eq. 5; (4) **w/o KCG**, where z^- is replaced by a duplicate of z^+ to eliminate the KCG signal.

The results on Llama-3-8B are presented in Table 3. Results for the other LLMs are provided in Appendix F. We make the following observations: (1) Replacing RAGap with *w/o* $\mathcal{S}_{ctx}$ leads to consistent performance drops. This confirms that masking the entire layer induces excessive representational drift, while targeting $\mathcal{S}_{ctx}$ precisely isolates the parametric knowledge by only muting context heads. (2) Setting either z^+ or z^- to zero results in significant degradation across all tasks. This indicates that z^- alone is insufficient; the detector requires the contrast between the LLM's

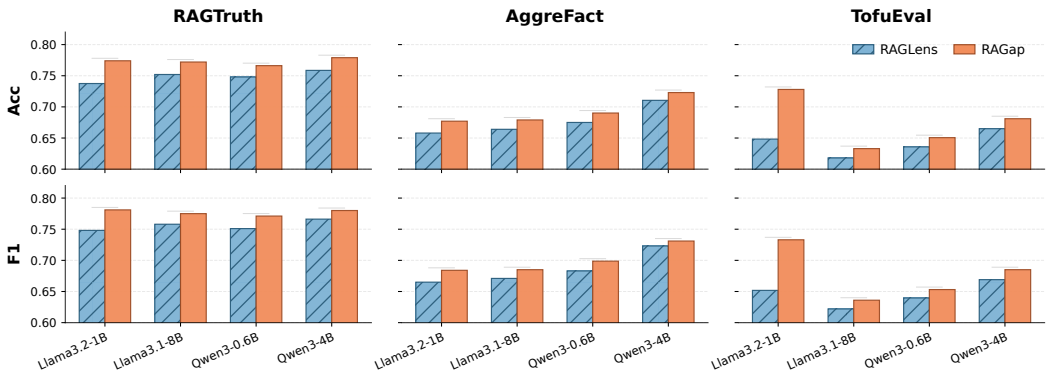


Figure 3: Comparison between RAGLens and RAGap across four models on RAGTruth, AggreFact, and TofuEval, measured by Acc and F_1.

Table 4: Stage-wise and end-to-end detection latency on Dolly. Times are reported in milliseconds per example as means $\pm$ standard deviations over 3 seeds $\times$ 2 folds. Additional encoding processes the context-removed input. Latency reductions are relative to prompt-level manipulation.

Stage	Llama-2-7B		Llama-3.2-1B	
	Prompt-level	RAGap	Prompt-level	RAGap
Primary encoding	118.04 ± 0.33	117.82 ± 0.23	44.64 ± 0.38	44.42 ± 0.16
Additional encoding	48.33 ± 0.31	–	25.29 ± 0.20	–
Feature processing	0.53 ± 0.01	0.59 ± 0.02	0.69 ± 0.03	0.64 ± 0.02
MLP prediction	0.01 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.02 ± 0.00
End-to-end	173.86 ± 1.47	$\mathbf{126.35 \pm 2.06}$	78.15 ± 2.39	$\mathbf{53.15 \pm 2.02}$
Latency reduction	–	**27.3%**	–	**32.0%**

internal belief (z^-) and the context-conditioned state (z^+) to effectively detect hallucinations. (3) The *w/o KCG* variant fails to achieve comparable performance despite having identical dimensionality. This demonstrates that the interaction signals ($|z^+ - z^-|$ and $z^+ \odot z^-$) provide the discriminative power.

Additional hyperparameter sensitivity results are provided in Appendix G.

5.4 Efficiency Analysis

We evaluate detection latency on Dolly, comparing RAGap with prompt-level manipulation (Section 3.3). Table 4 reports both stage-wise runtime and end-to-end detection latency.

As shown in Table 4, RAGap reduces end-to-end detection latency by 27.3% on Llama-2-7B and 32.0% on Llama-3.2-1B. The primary encoding cost remains comparable between the two methods, while the main efficiency gain comes from avoiding the additional full forward pass required by prompt-level manipulation. Feature processing and MLP prediction introduce only minor overhead.

6 Conclusion

In this work, we characterize the *Knowledge-Context Gap* in the SAE feature space and demonstrate the value of jointly modeling paired SAE features for RAG hallucination detection. To address the inefficiency and representational drift of prompt-level manipulation, we propose RAGap, a mechanistic framework that isolates KCG by suppressing *context heads* through differentiable masking. Our method enables efficient and interpretable hallucination detection that consistently outperforms state-of-the-art baselines across diverse benchmarks.

REFERENCES

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

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. HHEM-2.1-Open, 2024. URL https://huggingface.co/vectara/hallucination_evaluation_model.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013. URL https://arxiv.org/abs/1308.3432.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. Lm vs lm: Detecting factual errors via cross examination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12621–12640, 2023.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410.

Passant Elchafei and Mervat Abu-Elkheir. Hallucination detectives at semeval-2025 task 3: Span-level hallucination detection for llm-generated answers. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pp. 601–606, 2025.

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, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 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*, pp. 150–158, 2024.

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

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *EMNLP (1)*, pp. 5484–5495. Association for Computational Linguistics, 2021.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *arXiv preprint arXiv:2405.14486*, 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting Off the Head Ends the Conflict: A Mechanism for Interpreting and Mitigating Knowledge Conflicts in Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.70.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 9332–9346, 2020.

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

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2024b. URL https://arxiv.org/abs/2306.03341.

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

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. Chatgpt as a factual inconsistency evaluator for text summarization. *arXiv preprint arXiv:2303.15621*, 2023.

Xuhua Ma, Richong Zhang, and Zhijie Nie. Corect: Context-aware logit contrast for hidden state rectification to resolve knowledge conflicts, 2026. URL https://arxiv.org/abs/2602.08221.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025.

Andrey Malinin and Mark J. F. Gales. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 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)*, pp. 10862–10878, 2024.

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, and Amir H. Abdi. Llama see, llama do: A mechanistic perspective on contextual entrainment and distraction in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16218–16239, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.791. URL https://aclanthology.org/2025.acl-long.791/.

Subhey Sadi Rahman, Md Adnanul Islam, Md Mahbub Alam, Musarrat Zeba, Md Abdur Rahman, Sadia Sultana Chowa, Mohaimenul Azam Khan Raiaan, and Sami Azam. Hallucination to truth: A review of fact-checking and factuality evaluation in large language models. *arXiv preprint arXiv:2508.03860*, 2025.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=kJUS5nD0vPB.

Wei Shi, Sihang Li, Tao Liang, Mingyang Wan, Guojun Ma, Xiang Wang, and Xiangnan He. Route sparse autoencoder to interpret large language models. *arXiv preprint arXiv:2503.08200*, 2025.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*, 2025.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803, 2021.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=ztzZDzgfrh.

Praneet Suresh, Jack Stanley, Sonia Joseph, Luca Scimeca, and Danilo Bzdok. From noise to narrative: Tracing the origins of hallucinations in transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=xd7cXU2P4G.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11626–11644, 2023.

Liyan Tang, Philippe Laban, and Greg Durrett. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, 2024a.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4455–4480, 2024b.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Truera. Trulens: Evaluation and tracking for llm experiments and ai agents. https://github.com/truera/trulens, 2025. Version 2.3.1 (latest release as of September 2025).

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P1qhkp8gQT.

Guangzhi Xiong, Zhenghao He, Bohan Liu, Sanchit Sinha, and Aidong Zhang. Toward faithful retrieval-augmented generation with sparse autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hgBZP67BkP.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey, 2024a. URL https://arxiv.org/abs/2403.08319.

Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. Qa-lora: Quantization-aware low-rank adaptation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=WvFoJccpo8.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Samuel Yeh, Sharon Li, and Tanwi Mallick. Lumina: Detecting hallucinations in rag system with context–knowledge signals. In *The Fourteenth International Conference on Learning Representations*, 2026.

Lei Yu, Jingcheng Niu, Zining Zhu, Xi Chen, and Gerald Penn. Sheaf discovery with joint computation graph pruning and flexible granularity, 2025. URL https://arxiv.org/abs/2407.03779.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068, 2024.

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

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pp. 1–46, 2025.

Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Understanding parametric and contextual knowledge reconciliation within large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=76cFMRgEzQ.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. Hademif: Hallucination detection and mitigation in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=VwOYxPScxB.

APPENDIX TABLE OF CONTENTS

To facilitate navigation, we organize the supplementary materials as follows:

- **Appendix A: Formal Analysis of Representational Drift**
  Formal analysis of representational drift under prompt-level context removal and localized context-head intervention.
- **Appendix B: Dataset and Baseline Details**
  Hallucination detection datasets, the LRE dataset for context-head discovery, and baseline descriptions, including the CoT prompt and scoring procedure.
- **Appendix C: Implementation Details**
  SAE configurations, detector training hyperparameters, feature selection, and computing infrastructure.
- **Appendix D: Additional Results on the Knowledge-Context Gap**
  Additional SAE-based KCG characterization results across model families and scales.
- **Appendix E: Generalization across Datasets and Tasks**
  Zero-shot transfer evaluation across datasets and RAG subtasks.
- **Appendix F: Additional Ablation Results**
  Component ablations on Llama-3.2-1B and Qwen3-4B across three hallucination detection benchmarks.
- **Appendix G: Hyperparameter Analysis**
  Sensitivity to the selected feature count and the encoder and fusion dimensions.
- **Appendix H: Detailed Efficiency Measurements**
  Stage-wise and end-to-end detection latency, with setup costs reported separately.
- **Appendix I: Limitations and Research Statements**
  Limitations, broader impacts, ethics, reproducibility, and AI use.

A Formal Analysis of Representational Drift

We provide a formal treatment of the **representational drift** described in Section 3.3.

A.1 Modeling the Pre-FFN Residual State

Let $\boldsymbol{h}^{l-1} \in \mathbb{R}^d$ be the output of layer $l-1$. The state entering the FFN at layer l, denoted as $\boldsymbol{z}^l$, is the sum of the layer input and the aggregate attention output. In the context-conditioned setting ($\mathcal{I}^+$), we decompose the attention output into contributions from Context Heads ($\mathcal{S}_{ctx}$) and the remaining heads ($H \setminus \mathcal{S}_{ctx}$):

$$\boldsymbol{z}_+^l = \boldsymbol{h}_+^{l-1} + \underbrace{\sum_{j \in \mathcal{S}_{ctx}} \text{Attn}_j^l(\boldsymbol{h}_+^{l-1})}_{\boldsymbol{c}^l} + \sum_{j \notin \mathcal{S}_{ctx}} \text{Attn}_j^l(\boldsymbol{h}_+^{l-1}), \tag{11}$$

where $\boldsymbol{c}^l$ represents the external context signals written into the residual stream by specialized context heads.

A.2 Collateral Drift vs. Targeted Intervention

When context is removed via **prompt-level ablation** ($\mathcal{I}^-$), the resulting state $\boldsymbol{z}_-^l$ suffers from a cumulative deformation across all preceding layers. We define the total representational drift as $\boldsymbol{\Delta} = \boldsymbol{z}_+^l - \boldsymbol{z}_-^l$. Crucially, this drift is not merely the absence of context $\boldsymbol{c}^l$:

$$\boldsymbol{\Delta} = \boldsymbol{c}^l + \boldsymbol{\delta}_{collateral}, \tag{12}$$

where $\boldsymbol{\delta}_{collateral}$ represents the non-linear, cumulative shift in the residual stream caused by the altered self-attention patterns in layers 1 to $l-1$. Because prompt-level ablation changes the KV-cache and attention scores globally, $\boldsymbol{\delta}_{collateral}$ pushes the representation away from the model's learned activation manifold.

A.3 Contextual Priming and Knowledge Attenuation

The FFN retrieves parametric knowledge $\boldsymbol{v}^l$ based on whether $\boldsymbol{z}^l$ falls within the activation manifold of its keys $\boldsymbol{k}_i$. The activation of the i-th unit is $a_i(\boldsymbol{z}) = \sigma(\langle \boldsymbol{k}_i, \gamma(\boldsymbol{z}) \rangle + b_i)$.

Hypothesis 1 (Knowledge Attenuation via Collateral Drift). *Certain internal knowledge keys $\boldsymbol{k}_i$ are "contextually primed," requiring the hidden state to remain on the conditioned trajectory to trigger activation. For prompt-level ablation, the following condition holds for these keys:*

$$\langle \boldsymbol{k}_i, \gamma(\boldsymbol{z}_+^l) \rangle > \tau \quad \text{but} \quad \langle \boldsymbol{k}_i, \gamma(\boldsymbol{z}_+^l - (\boldsymbol{c}^l + \boldsymbol{\delta}_{collateral})) \rangle < \tau. \tag{13}$$

In this case, the threshold violation is primarily driven by $\boldsymbol{\delta}_{collateral}$, which causes a catastrophic mismatch between the distorted input and the static FFN keys $\boldsymbol{k}_i$, leading to the failure of parametric knowledge retrieval.

A.4 The RAGap Distinction: Preserving the Manifold

In contrast, RAGap performs a **localized intervention** within the same forward pass. By masking only the output of $\mathcal{S}_{ctx}$ while maintaining the conditioned state $\boldsymbol{h}_+^{l-1}$ and the attention patterns of non-context heads, RAGap constructs:

$$\boldsymbol{z}_{RAGap}^l = \boldsymbol{z}_+^l - \boldsymbol{c}^l. \tag{14}$$

Since $\boldsymbol{z}_{RAGap}^l$ is derived directly from the original conditioned path, it satisfies $\boldsymbol{\delta}_{collateral} \approx 0$. This ensures that the state $\boldsymbol{z}_{RAGap}^l$ stays within the model's activation manifold.

Consequently, while the "contextual boost" $\boldsymbol{c}^l$ is removed, the structural integrity of the representation is preserved, allowing the FFN to still activate the parametric keys $\boldsymbol{k}_i$ that are intrinsically supported by the query q. This mechanistic isolation allows RAGap to characterize the KCG without the confounding noise of global representational drift.

B Dataset and Baseline Details

B.1 Hallucination Detection Datasets

We evaluate our method on four hallucination detection datasets, described below.

- **RAGTruth** (Niu et al., 2024), a large-scale dataset containing nearly 18,000 LLM-generated responses in retrieval-augmented settings, covering summarization, data-to-text generation, and question answering.
- **Dolly** (Hu et al., 2024), a dataset for evaluating the faithfulness of model-generated responses to relevant and accurate context documents under the "Accurate Context" setting.
- **AggreFact** (Tang et al., 2023), a benchmark for evaluating factual consistency in summarization, pairing source documents with model-generated summaries and human-annotated factuality labels.
- **TofuEval** (Tang et al., 2024b), a benchmark for topic-focused dialogue summarization, featuring a MeetingBank subset with multi-turn meeting transcripts and topic-specific summaries annotated with sentence-level factuality labels.

B.2 LRE Dataset Specifications

The LRE dataset serves as the foundational source for constructing our experimental prompts (Hernandez et al., 2024). This benchmark contains 47 distinct relations across 10,000 instances, encompassing factual associations, commonsense knowledge, implicit biases, and linguistic knowledge. Originally derived by filtering noisy triples from PARAREL (Elazar et al., 2021), the dataset provides a structured framework for evaluating how models retrieve and prioritize specific relational information.

B.3 Baselines

We benchmark RAGap against a diverse set of representative hallucination detection methods. For clarity, we group all baselines into three main categories, detailed below.

Prompting-based Detectors. This category captures hallucination signals by leveraging the generative and reasoning capabilities of LLMs to produce token-level decisions. It encompasses various prompting strategies, such as prompt engineering (Friel & Sanyal, 2023), multi-agent collaboration (Cohen et al., 2023), as well as supervised fine-tuning. Following Sun et al. (2025), we include a fine-tuned Llama2-13B(LR) baseline, where the detector is trained on RAGTruth using LoRA (Xu et al., 2024b). The full list of baseline detectors in this category includes: Prompt (Niu et al., 2024), Llama2-13B(LR) (Sun et al., 2025), LMvLM (Cohen et al., 2023), FActScore (Min et al., 2023), FactCC (Kryściński et al., 2020), ChainPoll (Friel & Sanyal, 2023), RAGAS (Es et al., 2024), TruLens (Truera, 2025), RefCheck (Hu et al., 2024), and P(True) (Kadavath et al., 2022).

Uncertainty-based Detectors. These methods assess hallucination likelihood based on the uncertainty of LLM outputs, typically measured via sampled tokens or the distribution of output logits prior to token generation. The complete list of detectors in this category includes: SelfCheckGPT (Manakul et al., 2023), LN-Entropy (Malinin & Gales, 2021), Energy (Liu et al., 2020), Focus (Zhang et al., 2023), and Perplexity (Ren et al., 2023).

Internal Representation-based Detectors. These approaches probe the internal representations of the LLM, such as hidden states, attention patterns, or other intermediate representations, to identify unfaithful generations. By analyzing these internal model dynamics, these detectors aim to capture subtle signals associated with hallucination that may not be reflected in output tokens or logits. Methods in this category include: EigenScore (Chen et al., 2024), SEP (Han et al., 2024), SAPLMA (Azaria & Mitchell, 2023), ITI (Li et al., 2024a), ReDeEP (Sun et al., 2025) and RAGLens (Xiong et al., 2026).

B.4 Implementation of the CoT Baseline

For the zero-shot baseline, we utilize Chain-of-Thought (CoT) prompting (Wei et al., 2022) to elicit reasoning before a final judgment. Following Luo et al. (2023), we employ the following template for all evaluated LLMs:

Hallucination Detection Prompt

Instruction Does the following [Response] contain any hallucinations relative to the [Context]? A hallucination is defined as any claim that **contradicts** the context or is **not explicitly supported** by it (extrinsic information not in the context is also considered a hallucination).

Context [Context]

Response [Response]

Output Briefly explain your reasoning, then answer **yes** (there is a hallucination) or **no** (it is faithful).

To calculate the AUROC, we extract the logits of the tokens **yes** and **no** at the final decision step, denoted as l_{yes} and l_{no} respectively. The final hallucination score is then computed as the normalized probability:

$$s = \frac{\exp(l_{\text{yes}})}{\exp(l_{\text{yes}}) + \exp(l_{\text{no}})}. \tag{15}$$

This continuous score $s \in [0, 1]$ represents the model's internal confidence in the presence of a hallucination.

C Implementation Details

C.1 Sparse Autoencoders

While a trained SAE is required for detection in RAGap, the rapid development of the SAE community (Shu et al., 2025) has led to the public release of many pre-trained SAEs for widely used LLMs. Moreover, SAEs are typically trained on general-purpose corpora rather than task-specific datasets, allowing for their reuse in analyses beyond the scope of this work. To demonstrate the efficiency and practicality of our pipeline, we utilize publicly available SAEs whenever possible, showing that our method does not require resource-intensive, specially tuned SAEs for effective performance. Below, we list the specific SAEs used in our experiments:

- **Llama2-7B:** We use the pretrained SAE from `https://huggingface.co/yuzhaouoe/Llama2-7b-SAE`, which includes SAEs for multiple layers. Specifically, we select the SAE trained on "layers.15", with an expansion factor of 32 and Top-K activation ($K = 192$).
- **Llama2-13B:** As no public SAE is available for Llama2-13B, we train our own using the `sparsify` package[1] with default settings. The SAE is trained on "layers.15", with an expansion factor of 16 and Top-K activation ($K = 16$).
- **Llama3.2-1B:** We use the pretrained SAE from `https://huggingface.co/EleutherAI/sae-Llama-3.2-1B-131k`, which covers multiple layers. For results in Section 3, we select the SAE trained on "layers.8.mlp".
- **Llama3-8B:** We use the pretrained SAE from `https://huggingface.co/EleutherAI/sae-llama-3-8b-32x`. For Section 3, we select the SAE trained on "layers.17". The SAEs have an expansion factor of 32 and Top-K activation ($K = 192$).
- **Qwen3-4B:** Similarly, we train our own SAEs for Qwen3-4B. For Section 3, we select the SAE trained on "layers.17". The SAEs have an expansion factor of 32 and Top-K activation ($K = 16$).
- **Qwen3-0.6B:** Similarly, we train our own SAEs for Qwen3-0.6B. For Section 3, we select the SAE trained on "layers.17". The SAEs have an expansion factor of 32 and Top-K activation ($K = 16$).

[1] `https://github.com/EleutherAI/sparsify`

C.2 Training and Feature Selection

To ensure the reproducibility of our results, we provide a detailed summary of the experimental configurations in Table 5. For all neural-based detectors, we utilize the AdamW optimizer with a learning rate of 10^{-5} and a weight decay of 10^{-4}. All models are trained with a batch size of 256 for a maximum of 1000 epochs, using early stopping with a patience of 50 rounds and a tolerance $\Delta = 10^{-5}$ monitored on the validation split. Regarding feature engineering, we select the top $K' = 1000$ SAE features based on their mutual Information with the faithfulness labels.

Table 5: Comprehensive experimental hyperparameters and protocol settings.

Category	Hyperparameter	Value
Shared Neural Settings	Optimizer	AdamW
	Feature selection size K'	1000
	Learning rate	10^{-5}
	Weight decay	10^{-4}
	Batch size	256
	Max training epochs	1000
MLP Classifier Settings	dimension	512
MLP Fusion Settings	dimension	512

C.3 Computing Infrastructure

Experiments are conducted on a server equipped with an Intel Xeon Gold 6248R and four A100-PCIE-40GB GPUs.

D Additional Results on the Knowledge-Context Gap

Figure 4 extends the KCG characterization in Section 3 to four additional models. The results support the findings discussed in Section 3.2 across model families and scales.

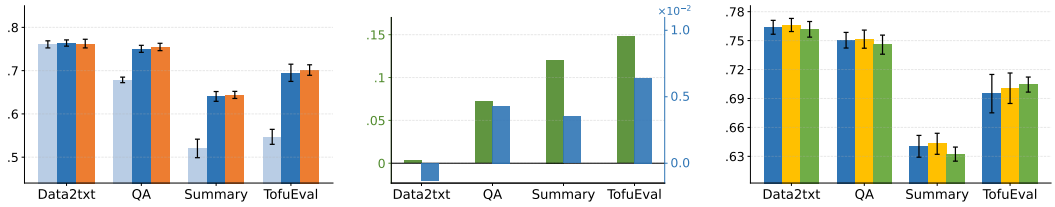
E Generalization across Datasets and Tasks

Table 6 evaluates the zero-shot transferability of RAGap across different distributions and RAG subtasks. RAGap consistently outperforms the CoT baseline in unseen domains. Specifically, detectors trained on RAGTruth or the Summary task exhibit the strongest cross-domain stability, suggesting that the mechanistically characterized KCG captures a universal, task-agnostic signal of unfaithfulness. The successful transfer between distinct tasks (e.g., Summary to QA) confirms that by isolating context heads, RAGap identifies a core representational conflict in the residual stream rather than overfitting to dataset-specific linguistic artifacts.

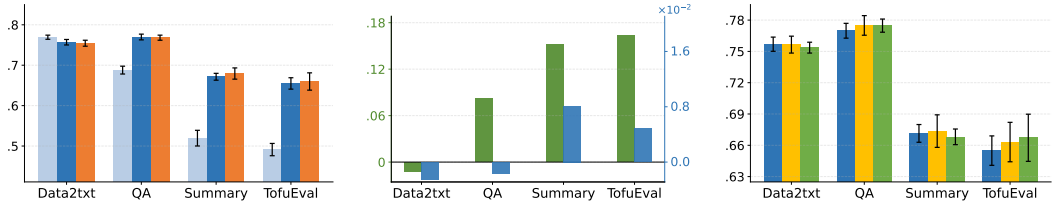
Table 6: Generalization across datasets and tasks. "None" denotes CoT. All scores are AUROC.

Train\Test	RAGTruth	AggreFact	TofuEval
Llama-2-7B			
None	.484	.574	.556
RAGTruth	**.881**	.773	.769
AggreFact	.600	**.794**	.624
TofuEval	.777	.555	**.774**
Llama-3.2-1B			
None	.422	.530	.493
RAGTruth	**.854**	.761	.751
AggreFact	.658	**.781**	.619
TofuEval	.776	.663	**.785**

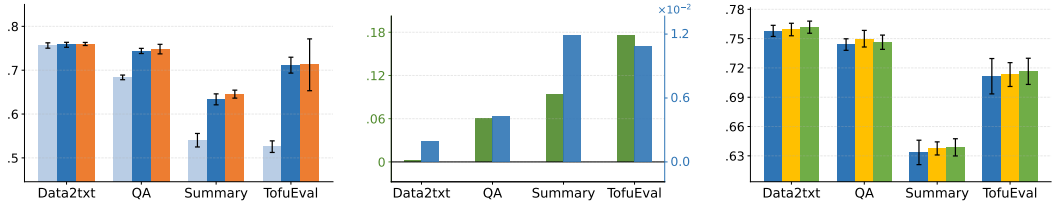
Train\Test	Summary	QA	Data2txt
Llama-2-7B			
None	.492	.485	.495
Summary	**.804**	.819	.595
QA	.715	**.887**	.684
Data2txt	.556	.644	**.843**
Llama-3.2-1B			
None	.453	.456	.404
Summary	**.699**	.807	.584
QA	.625	**.868**	.583
Data2txt	.529	.581	**.837**



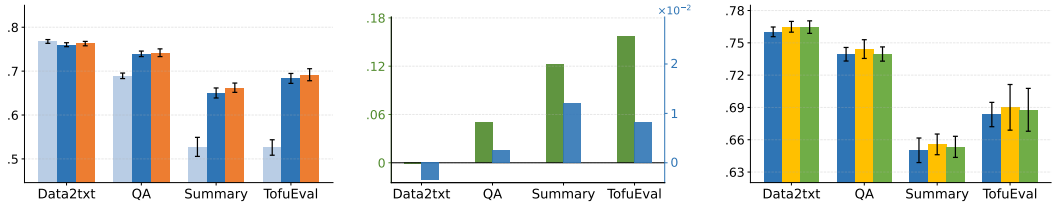
(a) **Llama-3.2-1B**



(b) **Llama-3-8B**



(c) **Qwen3-0.6B**



(d) **Qwen3-4B**

Figure 4: KCG characterization results for additional models. Each row corresponds to one model, with three panels following the plotting scheme of Figure 1. All results are averaged over 10 independent trials with distinct random seeds.

Table 7: Ablation study of RAGap on Llama-3.2-1B. We report performance and relative drop compared to the full model.

Method	RAGTruth			AggreFact			TofuEval		
	AUC	Acc	F_1	AUC	Acc	F_1	AUC	Acc	F_1
Full	**.854**	**.774**	**.781**	**.781**	**.677**	**.684**	**.785**	**.728**	**.733**
w/o $\mathcal{S}_{ctx}$	.850 (-.004)	.771 (-.003)	.773 (-.007)	.780 (-.001)	.672 (-.005)	.679 (-.004)	.760 (-.025)	.691 (-.038)	.701 (-.032)
w/o z^+	.847 (-.007)	.760 (-.014)	.765 (-.016)	.779 (-.002)	.677 (-.000)	.684 (-.000)	.768 (-.017)	.688 (-.041)	.687 (-.046)
w/o z^-	.845 (-.009)	.770 (-.004)	.777 (-.004)	.775 (-.006)	.675 (-.002)	.683 (-.001)	.782 (-.002)	.701 (-.027)	.710 (-.023)
w/o KCG	.848 (-.006)	.767 (-.007)	.773 (-.007)	.772 (-.009)	.675 (-.002)	.682 (-.002)	.780 (-.005)	.724 (-.004)	.730 (-.003)

Table 8: Ablation study of RAGap on Qwen3-4B. We report performance and relative drop compared to the full model.

Method	RAGTruth			AggreFact			TofuEval		
	AUC	Acc	F_1	AUC	Acc	F_1	AUC	Acc	F_1
Full	**.858**	**.779**	**.780**	**.813**	**.723**	**.731**	**.783**	**.681**	**.685**
w/o $\mathcal{S}_{ctx}$	.854 (-.004)	.777 (-.002)	.779 (-.001)	.807 (-.006)	.716 (-.007)	.721 (-.010)	.766 (-.017)	.666 (-.015)	.670 (-.015)
w/o z^+	.850 (-.008)	.776 (-.003)	.779 (-.001)	.808 (-.005)	.712 (-.011)	.720 (-.011)	.766 (-.017)	.666 (-.015)	.672 (-.013)
w/o z^-	.852 (-.006)	.771 (-.008)	.772 (-.008)	.811 (-.002)	.715 (-.008)	.722 (-.009)	.773 (-.009)	.670 (-.011)	.676 (-.009)
w/o KCG	.853 (-.005)	.765 (-.014)	.766 (-.014)	.803 (-.011)	.701 (-.022)	.708 (-.023)	.779 (-.004)	.675 (-.006)	.678 (-.007)

F Additional Ablation Results

To further demonstrate the cross-architecture robustness of our approach, we provide additional ablation results for Llama-3.2-1B and Qwen3-4B across three RAG benchmarks in Table 7 and Table 8. The results consistently verify that targeted mechanistic intervention on context heads is superior to full-layer masking by minimizing representational drift, and confirm that the interaction signals characterizing KCG are universally critical for reliable hallucination detection across different model families and parameter scales.

G Hyperparameter Analysis

We analyze the sensitivity of RAGap to three key hyperparameters: the feature count K', the encoder dimension D of Ψ_ω, and the fusion dimension D_{fuse} of the classification head g_ϕ. As shown in Figure 5, performance generally improves as K' increases, peaking at approximately $K' = 512$ across most benchmarks. This suggests that a higher number of SAE features provides a more granular characterization of the KCG.

Regarding the architectural dimensions, D and D_{fuse} control the projection capacity of the feature encoders and the complexity of the interaction fusion, respectively. We observe that RAGap exhibits low sensitivity to these parameters, maintaining stable macro F_1 scores as these dimensions vary from 96 to 768. This stability indicates that the interaction signals encoded in $\mathbf{z}_{\text{joint}}$ are sufficiently discriminative, allowing the model to achieve consistent detection performance without requiring extensive tuning of the hidden layers.

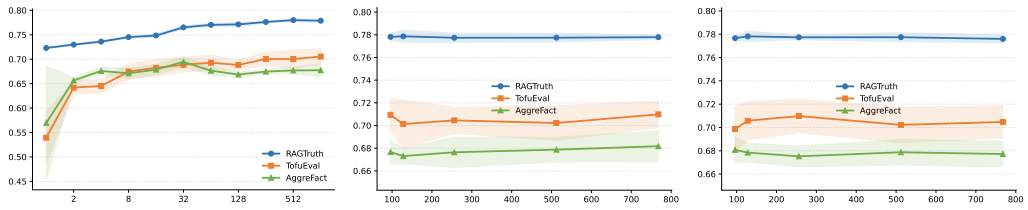
H Detailed Efficiency Measurements

Table 9 provides the stage-wise latency measurements underlying Section 5.4. *Prompt-level* denotes prompt-level context removal. *Encode1* and *Encode2* denote the first and second LLM encoding stages, respectively; *Select* and *Predict* denote feature processing and lightweight MLP inference. The end-to-end runtime is measured directly, excludes one-time setup, and is reported separately from the individual stage timings.

I Limitations and Research Statements

I.1 Limitations

Despite the contributions of this work, several limitations should be acknowledged.



(a) Feature count K' (b) Encoder dimension D (c) Fusion dimension D_{fuse}

Figure 5: Parameter sensitivity analysis of RAGap on RAGTruth, TofuEval, and AggreFact.

Table 9: Stage-wise latency on Dolly in milliseconds per example, reported as mean $\pm$ standard deviation over 3 seeds $\times$ 2 folds. Setup is listed separately and excluded from Runtime. A dash indicates that no second full LLM forward pass is required.

Stage	Llama-2-7B		Llama-3.2-1B	
	Prompt-level	RAGap	Prompt-level	RAGap
Setup	6.25 ± 0.00	6.25 ± 0.00	1.13 ± 0.00	1.13 ± 0.00
Encode1	118.04 ± 0.33	117.82 ± 0.23	44.64 ± 0.38	44.42 ± 0.16
Encode2	48.33 ± 0.31	–	25.29 ± 0.20	–
Select	0.53 ± 0.01	0.59 ± 0.02	0.69 ± 0.03	0.64 ± 0.02
Predict	0.01 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.02 ± 0.00
Runtime	173.86 ± 1.47	126.35 ± 2.06	78.15 ± 2.39	53.15 ± 2.02

Generalizability across Diverse Model Architectures. Our evaluation primarily focuses on a specific set of mainstream Large Language Models. Consequently, the findings might not fully capture the unique hallucination patterns or behaviors inherent in other architectures, such as Mixture-of-Experts (MoE) models or those specifically fine-tuned for niche domains. Future work is required to validate the robustness of our approach across a more exhaustive range of model families.

Specificity to Hallucination Detection. This research is strictly scoped to the identification and mitigation of hallucinations. While this is a pivotal challenge in LLM reliability, our method does not extend to other critical alignment dimensions, including preventing toxic outputs, protecting against adversarial jailbreaking, and ensuring the logical consistency of multi-turn reasoning. A holistic safety framework would need to integrate our detection mechanism with other specialized safety filters.

Dependence on Benchmark Quality. The empirical results are contingent upon the quality and diversity of the datasets used for testing. As hallucinations can be highly context-dependent, our current evaluation may not account for all edge cases in real-world, open-domain conversations where the definition of "truth" can be fluid or temporally sensitive.

Granularity of Hallucination Localization. While RAGap demonstrates high efficacy in instance-level hallucination detection, it does not currently localize specific hallucinated spans or tokens within a generated response. In complex RAG scenarios involving multi-sentence outputs, a hallucination may be confined to a single unsupported entity or clause, while the rest of the generation remains faithful. Since our current framework aggregates features across the entire sequence to characterize the KCG, it lacks the granularity required for fine-grained attribution. Future work could explore the temporal dynamics of context head suppression at each decoding step, potentially enabling token-level diagnostics to assist in targeted iterative refinement of LLM outputs.

I.2 Broader Impacts

The RAGap framework yields several positive societal impacts by advancing the interpretability and reliability of RAG systems. By providing a mechanistic understanding of how context heads resolve or succumb to knowledge conflicts, our work enables developers to build safer AI assistants that are less prone to generating plausible but false information. This is particularly beneficial in

domains where factual integrity is paramount, such as digital journalism, automated research tools, and educational platforms.

However, we must also acknowledge potential negative impacts. The technical insights regarding context head vulnerabilities could be exploited by sophisticated actors to design adversarial prompts that mimic factual patterns while evading RAGap detection. Furthermore, widespread deployment of such detection tools might inadvertently lead to a "false sense of security" among users, who may reduce their critical scrutiny of model outputs. We encourage the research community to view RAGap as one component within a broader, multi-layered strategy for AI alignment and verification rather than a standalone solution for truthfulness.

I.3 Ethics Statement

This work utilizes standard public benchmark datasets and does not involve the collection of new personal data. The proposed RAGap framework is intended as a general-purpose research contribution to enhance AI reliability, rather than for harmful applications such as surveillance, deception, or weaponization.

By explicitly constructing mechanistic conflicts across context heads, our method provides a diagnostic tool to mitigate hallucinations, thereby promoting the responsible deployment of Large Language Models. We are not aware of any violations regarding privacy, consent, or copyright. However, we acknowledge that the pretrained models and benchmarks analyzed throughout this study may inherit existing societal biases. Such biases should be carefully considered during downstream applications of our detection mechanism. Additionally, while RAGap assists in identifying non-factual content, users should avoid over-reliance on automated tools for high-stakes decision-making.

I.4 Reproducibility Statement

Datasets and baseline methods are documented in Appendix B, with implementation details and experimental hyperparameters in Appendix C and Table 5. The formal analysis is included in Appendix A. The anonymous code repository is available at `https://anonymous.4open.science/r/ragap-5062`.

I.5 AI Use Statement

We used AI-assisted tools to correct grammatical errors in the manuscript. The authors take full responsibility for the final content of the paper.