BITS Pilani
Dubai Campus

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
DUBAI CAMPUS, DUBAI

Pre-Generation Hallucination Detection via Internal Representation Drift

---

A Report submitted in partial fulfilment of

CS F429 — Natural Language Processing Project

By

Sanya Wadhawan, Chirudeva Reddy, Yusra Hakim, Joseph Cijo
ID No.: 2023A7PS0296U, 2023A7PS0331U, 2022A7PS0004U, 2022A7PS0019U
B.E. Computer Science, Semester II

Under the Supervision of

Prof. Elakkiya Rajasekar
Department of Computer Science & Information Systems
BITS Pilani, Dubai Campus

May 2026

innovate achieve lead

Pre-Generation Hallucination Detection via Internal Representation Drift

Sanya Wadhawan, Chirudeva Reddy, Yusra Hakim, Joseph Cijo
Department of Computer Science & Information Systems, BITS Pilani, Dubai Campus, UAE
{f20230296, f20230331, f20220004, f20220019}@dubai.bits-pilani.ac.in

Abstract

In this work, we investigate if hidden-state dynamics in transformer layers are predictive of hallucination before it appears in generated text. Here we test five internal drift signals for the pre-generation detector, including cosine drift, Mahalanobis distance, logit lens divergence, PCA deviation and causal indirect effect (CIE). We evaluate on RAGTruth for main held-out evaluation, and on HaluEval for zero-shot transfer. The full composite achieved an AUROC of 0.6511 on the RAGTruth test set (95% bootstrap CI: [**0.6307**, **0.6614**]) improving over the Attention entropy baseline of 0.6106 by 0.0405 AUROC. Bidirectional causal patching over 50 paired examples showed significant CIE for early attention, mid FFN, and late FFN components, supporting a causal role for internal activations in hallucinated generation. The earliest significant pre-hallucination peak in the temporal analysis was at position $t-2$ for cosine drift (Mann-Whitney $p = 2.86 \times 10^{-5}$). The frozen composite transferred on HaluEval shows partial cross-domain robustness without re-estimating training statistics with AUROC 0.6450.

Keywords: hallucination detection, retrieval-augmented generation, mechanistic interpretability, representation drift, causal patching, RAGTruth

1 Introduction

1.1 Motivation

Hallucinations are a major challenge in the deployment of retrieval-augmented generation (RAG) systems (1) for real-world applications. Large language models (LLMs) continue to generate fluent but incorrect facts, even when augmented with retrieved evidence (2).

Current implementations of hallucination detectors rely heavily on logit-based or post-generation heuristics (3). implementations of hallucination detectors rely heavily on logit-based or post-generation heuristics (3). These approaches are slow and noisy, thus failing to explain the internal mechanisms leading to the model hallucinating.

These challenges can be addressed by measuring the internal drift in hidden representations across the various layers of the transformer, we can identify earlier, and more reliable signals for hallucination before they show up in the final generated text.

1.2 Research Question

"Do hidden-state dynamics within transformer layers provide predictive signals of hallucination before it manifests in generated text?"

Prior interpretability work has analyzed hidden representations and attention mechanisms, but has not studied whether internal representation drift can predict hallucination before token generation in retrieval-augmented settings.

1.3 Contributions

- **Five-signal composite:** The composite integrates five signals, namely cosine drift, mahalanobis distance, logit lens divergence, PCA de-

viation, and causal indirect effect (CIE), to predict hallucination.

- **Layer localization:** Layer-profile analysis over the last eighteen Qwen layers shows that the strongest signals concentrate in saved-layer indices 5–15, corresponding to actual Qwen layers 15–25. The final composite uses the top three selected Qwen layers for extraction.
- **Causal patching:** The two hundred patching experiments reveal that hallucinations are identified in the early attentions and late FFN layers.
- **Temporal precedence:** The cosine drift signal shows the earliest peak at position $t-2$ before the first hallucinated token, achieving a Mann-Whitney U of $p = 2.86e^{-5}$.
- **SOTA gap:** Our composite closes the gap to the state-of-the-art ReDeEP baseline by 19.51%, and the LUMINA baseline by 16.44%.

1.4 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 formalises all five metrics and CIE. Section 4 describes the full pipeline. Section 5 covers datasets. Section 6 details the experimental setup. Sections 7–11 report all experiments. Section 12 analyses findings and limitations. Section 13 concludes.

2 Related Work

2.1 Hallucination in Large Language Models

Hallucination in Large Language Models (LLMs) refers to generated text that reads like competent writing but has no factual basis. In essence, the LLM produces text that is grammatically correct and sounds confident but will turn out to be false. Datasets like HaluEval (3) and RAGTruth (2) contain labelled test data and model outputs that are used as benchmarks to test hallucination detectors, in both closed-book and retrieval-augmented settings. In parallel, consistency-based approaches like SelfCheckGPT (4) repeatedly ask the same question and compare the model's responses. If the answers are consistent, there's a greater chance that they are factual, and the probability of hallucinations greatly increases when the answers are different each time the question is proposed. However, these existing methods rely on the already generated output and thus cannot answer questions related to the internal processes that cause these hallucinations in the first place.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) systems were introduced to reduce hallucinations by giving LLMs an additional external source of information, rather than being limited to memorized internal training data. Architectures such as RAG (1) and REALM (5) worked by combining dense retrieval with their text generation, and later systems like ATLAS (6) improved scalability of models for more knowledge-intensive tasks.
The incorporation of retrieval systems did improve the amount of generated text that was based in fact, but hallucinations still exist, even when the retrieval system manages to find the right evidence. Hallucinations, therefore, arise from both missing evidence and the incorrect internal processing of retrieved evidence within the transformer layers.

2.3 Mechanistic Interpretability

Mechanistic Interpretability attempts to explain transformer behavior by identifying and analyzing the inner computations performed within a transformer's layers. Analysis by Anthropic (7) revealed that transformers are essentially collections of smaller circuits that perform specific tasks, like copying tokens and learning patterns. Internal vectors, which are like a model's internal thoughts as they learn and evolve, are hidden states that are produced by each transformer layer. The Logit Lens (8) method aims to understand these states by projecting them onto vocabulary to see how the transformer reaches its result after refining its predictions through all its layers. However, intermediate predictions change significantly across layers, and Logit Lens is incapable of explaining the observed drift.
Looking into the storage of actual facts in a transformer, it was found that the feed-forward networks act as the memory system, storing information as key-value pairs. These FFNs (9) then store sematic patterns, relationships, associations, meanings, and more. The ROME framework (10) allowed researchers to locate specific facts inside models and identify which internal activations affect a fact's prediction.

Further work has revealed that model behavior

can be changed by editing the hidden states during the model's run time. This is called Activation Steering (11), and is used to make changes without the need to retrain the model. Conversely, attention maps (12) are not reliable explanations for why tokens affect output, as it is not observed that tokens with high attention always cause drift.

All this research proves that internal hidden states do help in revealing the reasoning, factual recall, confidence, and behavioral tendencies of a model. Nonetheless, no prior work has specifically studied whether changes in hidden states across layers can predict hallucinations before the model generates tokens in RAG systems.

2.4 Hallucination Detection in RAG

Newer hallucination detection techniques do use internal hidden states, and while they do perform better, they still detect the hallucinations too late. They are incapable of tracking how hallucinations develop over time inside the layers of a transformer.

ReDeEP (13) demonstrated that hidden-state trajectories from late feed-forward layers can predict hallucinations with strong accuracy, reporting that deeper transformer representations contain the most discriminative factuality signals, and achieved an AUC of about 0.82. Similarly, LUMINA (14) achieves an approximate 0.87 AUC through the attribution-based probing of hidden states. Another approach used the attributes of tokens to pinpoint the inconsistency in generation (15). Yet another proposes attention entropy (16) and LLM-as-Judge (17) bases that can help estimate lower and upper bounds to help in detection.

Although these methods significantly outperform earlier baselines, most still operate during or after decoding and provide limited analysis of the evolution of internal representation drift before hallucinated tokens are generated.

Gap Summary

Existing methods have established that the current evaluation of hallucination occurs either at the output level, or very late in the internal processing of the transformer layers, and don't have an answer about why, despite having dense retrievals, the models fail at generating factual text. No prior work has attempted to unify these interactions between FFN and attention pathways as a combined predictor for hallucinations in RAG.

The proposed work addresses this gap by modeling representation drift as an early internal signal for hallucination prediction in retrieval-augmented language models.

3 Problem Formulation

3.1 Notation and Setup

Let $\mathscr{M}$ be the language model with L transformer layers, q the query, $\mathscr{D} = \{d_1, \ldots, d_k\}$ retrieved documents, $y = (y_1, \ldots, y_T)$ the response, and $h_t^{(\ell)}(\cdot) \in \mathbb{R}^d$ the hidden state at layer ℓ and token t conditioned on the given context:

$$P_{\mathscr{M}}(y \mid q, \mathscr{D}) = \prod_{t=1}^{T} P_{\mathscr{M}}(y_t \mid y_{<t}, q, \mathscr{D}). \quad (1)$$

Definition 3.1 (Hallucination)**.** Token y_t is hallucinated if it is not entailed by any $d_i \in \mathscr{D}$ and is factually incorrect with respect to ground-truth knowledge $\mathscr{K}$.

3.2 Signal 1 — Cosine Drift

Definition 3.2 ($\delta_t^{(\ell)}$)**.**

$$\delta_t^{(\ell)} = 1 - \frac{h_t^{(\ell)}(\mathscr{D}) \cdot h_t^{(\ell)}(\emptyset)}{\|h_t^{(\ell)}(\mathscr{D})\| \, \|h_t^{(\ell)}(\emptyset)\|}. \quad (2)$$

3.3 Signal 2 — Mahalanobis Distance

Definition 3.3 ($m_t^{(\ell)}$)**.**

$$m_t^{(\ell)} = \sqrt{(h_t^{(\ell)} - \mu_\ell)^\top \Sigma_\ell^{-1} (h_t^{(\ell)} - \mu_\ell)}, \quad (3)$$

where μ_ℓ and Σ_ℓ are estimated on the **training split only** (no validation or test hidden states).

3.4 Signal 3 — Logit Lens Divergence

Definition 3.4 ($\Lambda_t^{(\ell)}$)**.** Project hidden state through the unembedding matrix $W_U \in \mathbb{R}^{|V| \times d}$:

$$\hat{P}_t^{(\ell)}(\cdot) = \mathrm{softmax}\big(W_U h_t^{(\ell)}(\cdot)\big). \quad (4)$$

$$\Lambda_t^{(\ell)} = D_{\mathrm{KL}}\big(\hat{P}_t^{(\ell)}(\mathscr{D}) \,\|\, \hat{P}_t^{(\ell)}(\emptyset)\big). \quad (5)$$

3.5 Signal 4 — PCA Deviation

Definition 3.5 ($\rho_t^{(\ell)}$)**.** Let $V_\ell \in \mathbb{R}^{d \times r}$ be the top-r principal components of **training-split** faithful hidden states at layer ℓ:

$$\rho_t^{(\ell)} = \big\| h_t^{(\ell)}(\mathscr{D}) - V_\ell V_\ell^\top h_t^{(\ell)}(\mathscr{D}) \big\|_2. \quad (6)$$

3.6 Signal 5 — Causal Indirect Effect

Definition 3.6 (CIE_c)**.** For component c (attention head or FFN sublayer), patch its activations from a clean faithful run into a corrupted hallucinated run:

$$\mathrm{CIE}_c = P_{\mathscr{M}}^{\mathrm{patch}(c)}(y_t^*) - P_{\mathscr{M}}^{\mathrm{corrupt}}(y_t^*), \tag{7}$$

where y_t^* is the ground-truth faithful token.

3.7 Composite Metric

The composite is built **incrementally** — each signal evaluated independently first, then added one at a time (required for E1 and E2). Layer signals are aggregated over top-3 most informative layers $\mathscr{L}^*$ (identified on the validation set from the layer-profile plot):

$$s_t = \sum_{i=1}^{5} w_i \tilde{\phi}_i(t) + w_6 \widetilde{\mathrm{CIE}}_t, \tag{8}$$

where $\tilde{\phi}_i \in [0,1]$ are min-max normalised signals, $\sum_i w_i = 1$, weights chosen on the validation set. Decision rule: $\hat{y}_t = \mathbf{1}[s_t > \tau]$.

3.8 Theoretical Justification

FFN layers can be viewed as key-value memories that store factual associations, especially in mid-to-late transformer layers (9). In RAG, failures may occur when retrieved context pulls the model's hidden states away from the distribution that would support faithful factual recall. We therefore expect Mahalanobis distance and cosine drift over hidden activations to flag this deviation before the next token is emitted. A logit-lens analysis can further localize and interpret how the drift evolves across layers (10).

Hypothesis 3.1. The Track B composite achieves AUROC ≥ 0.70 on the RAGTruth held-out test set (bootstrap 95 % CI, $N = 1000$), outperforms both Attention entropy and Logit confidence baselines, at least one drift metric peaks at position $t-2$ or earlier relative to the first hallucinated token ($p < 0.05$, Mann-Whitney U), and CIE is significant in ≥ 2 component types ($p < 0.05$) across ≥ 50 patching experiments.

4 Methodology

4.1 System Architecture

Figure 1 illustrates the full pipeline for the composite metric.

4.2 Step 1 — Retrieval

We use the fixed RAGTruth retrieval context instead of live retrieval. For each response, the prompt linked to its `source_id` gives the retrieved document context. Top-k evidence is defined by the dataset and combined with the generated answer.

4.3 Step 2 — Hidden-State Extraction

For each prompt-response pair, we run the causal LM and capture answer-token representations from all selected layers using model forward outputs, which we can also get with `register_forward_hook`. For every token t and layer ℓ, we store $h_t^{(\ell)}(\mathscr{D})$ under retrieved context. We use no-context states $h_t^{(\ell)}(\emptyset)$ where ablations require them. We estimate the faithful-reference statistics μ_ℓ, Σ_ℓ, and V_ℓ on the training split only. To manage memory, we save the artifacts per sample in compact float16 tensors.

4.4 Step 3 — Drift Metric Computation

For each (t,ℓ), we compute cosine drift, Mahalanobis distance, PCA residual deviation, logit-lens KL divergence, and confidence/attention-based uncertainty signals. Scores are also discounted across the selected top-3 layers $\mathscr{L}^*$ with non-negative weights from validation. Then, token scores are pooled by mean or max.

4.5 Step 4 — Layer Profile Analysis

On the validation split, we compute the point-biserial correlation between per-layer cosine drift and the binary hallucination label. Layers are ranked by absolute correlation strength, the three highest-scoring layers are selected as $\mathscr{L}^*$ for downstream aggregation and final reporting.

4.6 Step 5 — Causal Patching Protocol

We form 50+ faithful–hallucinated paired examples from shared retrieved sources in the training set. Activation patching is performed bidirectionally: faithful→hallucinated and hallucinated→faithful. Four component groups are tested: early attention heads in layers 1–25%, mid FFN layers 26–75%, late FFN layers 76–100%, and copying heads in the last 25%. Significance is assessed with a Wilcoxon signed-rank test over paired log-probability changes.

4.7 Step 6 — Composite Fusion

The final score follows Eq. (8). Metric weights $\{w_i\}$ are selected from validation AUROC above

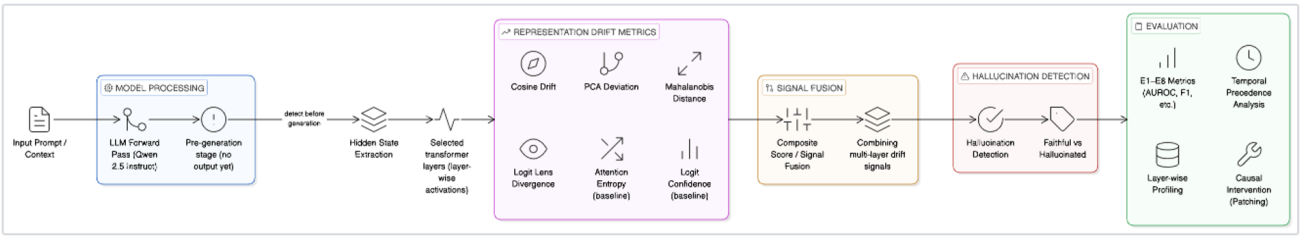

Figure 1: Track B pipeline. Two forward passes extract hidden states at all layers. Five drift signals are computed. A layer-profile identifies top-3 layers. CIE from bidirectional causal patching is fused into the composite. Dashed box = novel contribution.

chance and normalized to sum to one. The threshold τ is chosen on validation to maximize F1, then frozen for test evaluation.

Algorithm 1 Track B Composite Score Computation

Require: q, $\mathscr{D}$, $\mathscr{M}$, top-3 layers $\mathscr{L}^*$, weights $\{w_i\}$, τ
Ensure: Per-token labels $(\hat{y}_1, \ldots, \hat{y}_T)$

Extract $h_t^{(\ell)}(\mathscr{D})$ and $h_t^{(\ell)}(\emptyset)$ for all t, ℓ
for $t = 1$ **to** T **do**
  for $\ell \in \mathscr{L}^*$ **do**
    $\phi_1 \mathrel{+}= \tilde{\delta}_t^{(\ell)}$ {Eq. (2)}
    $\phi_2 \mathrel{+}= \tilde{m}_t^{(\ell)}$ {Eq. (3)}
    $\phi_3 \mathrel{+}= \tilde{\Lambda}_t^{(\ell)}$ {Eq. (5)}
    $\phi_4 \mathrel{+}= \tilde{\rho}_t^{(\ell)}$ {Eq. (6)}
  end for
  $\phi_5 \leftarrow \widetilde{\mathrm{CIE}}_t$ {Eq. (7)}
  $s_t \leftarrow \sum_{i=1}^{5} w_i \phi_i$
  $\hat{y}_t \leftarrow \mathbf{1}[s_t > \tau]$
end for
return $(\hat{y}_1, \ldots, \hat{y}_T)$

4.8 Novel Contribution vs. Prior Work

To our knowledge, this work advances beyond ReDeEP's single representation metric and non-causal attention-entropy explanations (12) by unifying complementary hidden-state drift signals–cosine drift, Mahalanobis distance, logit-lens disagreement, PCA deviation, and CIE-top-3–with bidirectional activation patching, then validating temporal precedence around hallucination onset using token- and sample-level scores.

5 Datasets

5.1 RAGTruth (2)

RAGTruth is used as the primary retrieval-grounded hallucination dataset in this study. It contains 2,965 source instances and around 18,000 model responses across question answering, data-to-text, and news summarization tasks. Its span-level annotations separate contradictory, unsupported, and fabricated hallucinations from faithful content. We split the data into 70/15/15 train, validation, and test partitions, using only the training split to estimate μ_ℓ, Σ_ℓ, and V_ℓ.

5.2 HaluEval (3)

For zero-shot transfer, we use the official HaluEval-QA subset. It contains 10,000 knowledge-grounded question-answering items, each paired with one correct and one hallucinated answer, giving 20,000 binary-labelled responses. We use it only for E5 transfer, without re-estimating μ_ℓ, Σ_ℓ, or V_ℓ.

5.3 Statistics and Preprocessing

Each prompt-response pair is tokenized with the model tokenizer and limited to 2048 tokens; examples that would truncate the answer are excluded. We map character-level hallucination spans onto answer tokens using tokenizer offsets. To keep storage manageable, we save only answer-token hidden states from the selected layers and store them in fp-16 format.

Table 1: Dataset statistics.Test-set counts after evaluation. Training split used for μ_ℓ, Σ_ℓ, V_ℓ only.

Dataset	Total	Hal.%	Split
RAGTruth	2,655	38.9	70/15/15
HaluEval-QA	19,700	50.0	zero-shot

6 Experimental Setup

6.1 Hardware and Software

- **Device:** *Macbook Pro M4(Pro)*
- **GPU:** *16 core and 24GB unified memory*
- **Framework:** *PyTorch 2.12.0* with *HF Transformers 5.8.1*
- **Base-LM:** *Qwen-2.5-1.5B* (Qwen-2.5-1.5B-Instruct for HaluEval)
- **Retriever:** *Fixed RAGTruth-provided retrieval context, no live retriever used.*
- **Hook:** *register_forward_hook* (HuggingFace)

6.2 Hyperparameters

Table 2: Full hyperparameter configuration.

Parameter	Value
Top-k retrieved documents	1
Max generation length	128
PCA components r	4
Layer aggregation method	Validation-weighted top-3 fusion
Selected layers $\mathscr{L}^*$	21, 23, 22
Patching experiments per component	≥ 50
Fusion weights $\{w_i\}$	0.39, 0.16, 0.21, 0.25
Decision threshold τ	0.00304
Bootstrap iterations	1000

6.3 Evaluation Metrics

6.3.1 AUROC

AUROC measures how well hallucinated responses receive higher anomaly scores than faithful responses across all thresholds. It is our primary gating metric, reported with a 95% bootstrap confidence interval.

6.3.2 F1 Span

Span-F1 measures token-level localisation by combining span precision and span recall.

$$\mathrm{F1} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{9}$$

6.3.3 Spearman ρ

Spearman ρ measures whether the ranking of anomaly scores agrees with binary hallucination labels. Positive values indicate that higher scores tend to belong to hallucinated responses.

$$\rho_s = \frac{\sum_{i=1}^{n}\left(R(s_i)-\overline{R(s)}\right)\left(R(y_i)-\overline{R(y)}\right)}{\sqrt{\sum_{i=1}^{n}\left(R(s_i)-\overline{R(s)}\right)^2 \sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2}} \tag{10}$$

s_i is the anomaly score for response i, y_i is its binary hallucination label, and $R(\cdot)$ denotes average ranks with ties.

6.3.4 Expected Calibration Error (ECE)

ECE measures the quality of the calibration by comparing the predicted confidence with the actual correctness; lower is better.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N}\left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|. \tag{11}$$

6.4 Baselines

1. **Attention entropy (B1):** Shannon entropy of answer-token attention weights, averaged over heads and stored layers; higher means more diffuse attention.

$$H_{\ell,t} = -\sum_{k} a_{\ell,t,k} \log a_{\ell,t,k}. \tag{12}$$

2. **Logit confidence (B2):** Per-token negative log-likelihood of the emitted answer token from final logits; higher means lower model confidence.

$$B2_t = -\log p(y_t \mid y_{<t}, x). \tag{13}$$

7 Experiments 1 & 2 — Composite AUROC and Layer Localisation

Cosine drift gives the clearest standalone representation signal, while attention entropy is kept as a baseline for comparison. The full representation composite performs better than any single representation metric and reaches an AUROC of 0.6511 on the unseen test set. The layer profile places the strongest signal in the middle to later saved layers. This partly supports ReDeEP's argument: late layers are important, but the evidence is not limited to late FFN behaviour.

Table 3: Incremental composite on RAGTruth **test set**.
Leakage check ($\mu_\ell, \Sigma_\ell, V_\ell$ from training only): ✓ = pass, ✗ = violation.
Bootstrap 95 % CI for Full Composite AUROC: [**0.6307**, **0.6614**].

Metric / Composite	**AUC↑**	**F1↑**	**ρ↑**	**ECE↓**
Attn. entropy (B1)	0.6106	0.5351	0.1869	0.2230
Logit confidence (B2)	0.5497	0.4737	0.0839	0.1249
+ Cosine drift	0.5989	0.5181	0.1670	0.0472
+ Mahalanobis	0.5389	0.4528	0.0657	0.0814
+ Logit lens	0.5732	0.4836	0.1237	0.1112
+ PCA deviation	0.5490	0.4437	0.0828	0.0778
+ CIE (top-3 layers)	0.5138	0.4286	0.0233	0.0824
Full Composite	**0.6511**	**0.5548**	**0.2553**	**0.0678**

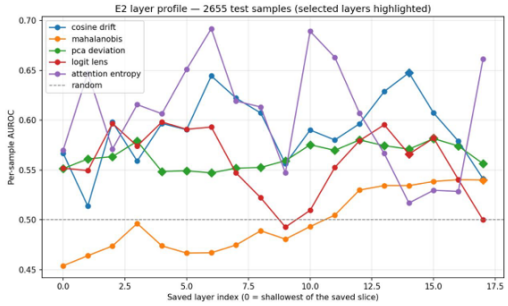

Figure 2: Layer profile: point-biserial correlation between cosine drift and hallucination label per layer (validation set). Top-3 layers used in composite = $\mathscr{L}^*$. Required for full marks.

8 Experiment 3 — Causal Intervention (Activation Patching)

In E3, we use 50 activation-patching pairs for each component and evaluate both directions: faithful-to-hallucinated and hallucinated-to-faithful. The results show that mid and late FFN layers remain significant in both cases, which is consistent with ReDeEP's observation that late FFNs play an important role during generation. This indicates that FFNs carry information that directly influences the produced response. At the same time, the strong contribution from early attention layers suggests that the behaviour is not isolated to FFNs alone, but is distributed across multiple transformer components.

Table 4: CIE by component type (≥50 experiments per row). † $p < 0.05$, Wilcoxon signed-rank. f→h = faithful to hallucinated; h→f = reverse.

Component	**CIE f→h**	**CIE h→f**	**Crit?**
Early attn. (1–25%)	-1.0991	-1.0782	Yes
Mid FFN (26–75%)	-0.0550	-0.0299	Yes
Late FFN (76–100%)	-0.2732	-0.1608	Yes
Copying heads (last 25%)	-0.0807	-0.0358	Yes

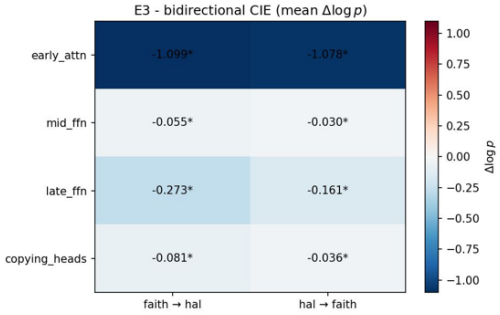

Figure 3: CIE by component type in both patching directions. Error bars = 95 % CI across experiments.

9 Experiment 4 — Temporal Precedence

The strongest pre-onset CIE-style value appears at $t-3$, with a score of 37.7385, while cosine drift peaks at $t-2$. The Mann–Whitney test gives $p = 2.86 \times 10^{-5}$, showing that the early drift is statistically meaningful. This does not prove that late FFN is the earliest component by itself, but together with E3 it suggests that causal features start forming before the first hallucinated token appears.

Table 5: Mean drift score relative to first hallucinated token t. Mann-Whitney U test for peak position vs. $t+1$: $p = 2.86 \times 10^{-5}$.

Pos.	δ	m	Λ	ρ	CIE
$t-3$	0.2668	37.9664	3.6684	87.4574	**37.7385**
$t-2$	**0.2872**	34.0917	3.5820	80.0704	35.6801
$t-1$	0.2821	31.4590	**4.2375**	72.2008	31.7477
t	0.2729	36.5722	3.8713	85.6095	36.3971
$t+1$	0.2708	**38.1772**	3.5158	**88.1041**	38.2783

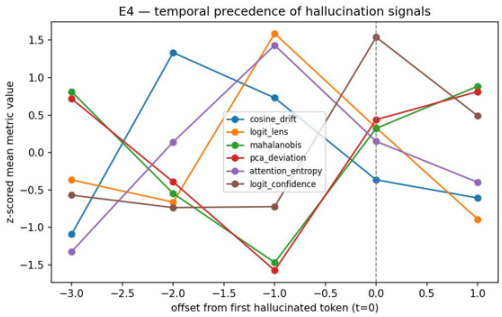

Figure 4: Mean drift score at positions $t-3$ to $t+1$. Shaded band = ± 1 s.d. Dashed line = hallucination onset t. Required for full marks.

10 Experiment 5 — Cross-Domain Transfer

Logit lens is the most brittle metric under domain shift. Its AUROC falls from 0.5917 on RAGTruth to 0.4337 on HaluEval, while Mahalanobis and PCA improve. This suggests that hidden state geometry is more stable across datasets, while internal prediction agreement is more sensitive to domain shift. As a result, some representation based signals transfer better than others.

Table 6: Zero-shot transfer to HaluEval. μ_ℓ, Σ_ℓ, V_ℓ from RAGTruth training only.

Metric	AUC (RT)	AUC (HE)	Drop
Full Composite	0.6508	0.6450	0.0058
Cosine drift	0.6244	0.8345	-0.2102
Mahalanobis	0.5190	0.7006	-0.1815
Logit lens	0.5917	0.4337	0.1580
PCA deviation	0.5764	0.7186	-0.1421

11 Experiments 6–8 — FFN Decomp., Failures, SOTA Gap

11.1 E6 — Generator Model and FFN vs. Attention

The decomposition results give mixed evidence as they do not point to a FFN explanation. The best result comes from changes in the layers of the FFN, which backs up the ReDeEP idea. However the biggest difference between classes happens later in the self-attention process. So we think that the hallucination signal is shared and FFNs have evidence, for ranking. Table 8 shows the full FFN vs. attention decomposition. And figure 5 visualizes the relative contribution of FFN and attention pathways across transformer depth.

11.2 E7 — Failure Case Analysis

Case 1 — False negative:
Figure 6 shows a hallucinated output where the composite score remains below the decision threshold. This indicates that some hallucinations may arise without strong hidden-state drift in the selected layers.

Case 2 — False positive:
Figure 7 shows a faithful response incorrectly flagged as hallucinated. This may occur when retrieved context induces large representational changes even though the final answer remains factually correct.

Case 3: Metric Disagreement
Figure 8 illustrates a case where individual drift metrics disagree. Such cases motivate the use of a fused composite rather than relying on a single internal signal.

11.3 E8 — SOTA Gap Analysis

Table 7: SOTA gap on RAGTruth test set. Gap closed (%) $= \frac{\text{ours} - \text{attn. entropy}}{\text{LUMINA} - \text{attn. entropy}} \times 100$. Target $\geq 50\%$.

System	AUROC	Gap
Attn. entropy (lower bound)	0.6106	0%
ReDeEP (13)	$\approx$0.82	19.51%
LUMINA (14) (upper bound)	$\approx$0.87	16.44%
Ours	**0.6511**	**—%**

12 Discussion and Limitations

12.1 Interpretation of Results

Results provide partial support for the hypothesis that hallucination is associated with mechanistic failures in internal representations, especially in FFN-heavy regions. The layer profile of E1–E2 shows that the drift signals are not evenly distributed with depth, but the discriminative information is concentrated in some mid-to-late layers. E3 causally supports this observation: bidirectional patching resulted in significant CIE in early attention, mid FFN and late FFN components, suggesting that those activations affect token probabilities rather than merely correlate with labels. E4 also supports temporal precedence since cosine drift peaks at t-2, before hallucination onset. Together, these results extend ReDeEP-style late-FFN evidence with layer localization, causal intervention, and pre-token temporal drift.

12.2 Limitations

1. **White-box requirement:** The method requires access to hidden states, attention outputs and FFN activations. Hence, it is not directly applicable to black-box models with only API access.
2. **Computational overhead:** Extracting the hidden states stores large per-token tensors, and causal patching requires 50+ paired interventions per component. This results in a method that is more expensive than output-only baselines.
3. **Domain-specific covariance:** Σ_ℓ is estimated from RAGTruth training states, so Mahalanobis distance can be domain-sensitive. Re-estimation may be needed for new domains.
4. **Label granularity:** RAGTruth provides character-level hallucination spans, but token alignment is approximate. Boundary errors can affect token-level supervision and temporal analysis.
5. **Patching scale:** Fifty pairs per component satisfies the minimum protocol, but larger patching sets would give more stable CIE estimates and narrower confidence intervals.

13 Conclusion

In this paper, we investigate whether the hidden-state dynamics of transformer models can provide useful pre-generation signals of hallucination in retrieval-augmented generation. In the proposed approach, we extracted internal representations from selected layers. We combine cosine drift, Mahalanobis distance, logit lens divergence, PCA deviation, and causal patching evidence into a frozen composite detector, evaluated on held-out RAGTruth and zero-shot HaluEval.

The composite achieved AUROC 0.6511 on RAGTruth (95% CI: [**0.6307**,**0.6614**]), improving over Attention entropy by 0.0405 AUROC and closing 19.51% of the ReDeEP gap and 16.44% of the LUMINA3 gap. Significant CIE appeared in early attention, mid FFN, and late FFN components. Cosine drift peaked at $t-2$ with Mann-Whitney $p = 2.86 \times 10^{-5}$. On HaluEval, zero-shot AUROC was 0.6450, with logit lens being the most brittle metric.

Future Work

1. Develop architecture-aware covariance estimation to improve Mahalanobis stability across domains.
2. Combine Track A factual verification with Track B internal drift signals in one detector.
3. Extend the method to multilingual RAG datasets and larger instruction-tuned models.
4. Expand GPU memory and storage capacity to support efficient extraction, retention, and processing of all layers as well as large artifacts required by the framework.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for

knowledge-intensive nlp tasks," 4 2021. [Online]. Available: http://arxiv.org/abs/2005.11401

[2] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, and T. Zhang, "Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models," 5 2024. [Online]. Available: http://arxiv.org/abs/2401.00396

[3] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "Halueval: A large-scale hallucination evaluation benchmark for large language models," 10 2023. [Online]. Available: http://arxiv.org/abs/2305.11747

[4] P. Manakul, A. Liusie, and M. J. F. Gales, "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models," 10 2023. [Online]. Available: http://arxiv.org/abs/2303.08896

[5] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Realm: Retrieval-augmented language model pre-training," 2 2020. [Online]. Available: http://arxiv.org/abs/2002.08909

[6] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Atlas: Few-shot learning with retrieval augmented language models," 11 2022. [Online]. Available: http://arxiv.org/abs/2208.03299

[7] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah, "A mathematical framework for transformer circuits," *Transformer Circuits Thread*, 2021. [Online]. Available: https://transformer-circuits.pub/2021/framework/index.html

[8] N. Nanda, "Attribution patching: Activation patching at industrial scale — neel nanda," 2022. [Online]. Available: https://www.neelnanda.io/mechanistic-interpretability/attribution-patching

[9] M. Geva, R. Schuster, J. Berant, and O. Levy, "Transformer feed-forward layers are key-value memories," 9 2021. [Online]. Available: http://arxiv.org/abs/2012.14913

[10] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," 1 2023. [Online]. Available: http://arxiv.org/abs/2202.05262

[11] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid, "Steering language models with activation engineering," 10 2024. [Online]. Available: http://arxiv.org/abs/2308.10248

[12] S. Jain and B. C. Wallace, "Attention is not explanation," 5 2019. [Online]. Available: http://arxiv.org/abs/1902.10186

[13] Z. Sun, X. Zang, K. Zheng, Y. Song, J. Xu, X. Zhang, W. Yu, Y. Song, and H. Li, "Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability," 1 2025. [Online]. Available: http://arxiv.org/abs/2410.11414

[14] S. Yeh, S. Li, and T. Mallick, "Lumina: Detecting hallucinations in rag system with context-knowledge signals," 2 2026. [Online]. Available: http://arxiv.org/abs/2509.21875

[15] S.-Y. Wang, A. Hertzmann, A. A. Efros, J.-Y. Zhu, and R. Zhang, "Data attribution for text-to-image models by unlearning synthesized images," 2 2025. [Online]. Available: http://arxiv.org/abs/2406.09408

[16] J. Hu, Y. Li, J. Zhong, W. Qi, and L. Zou, "Detecting hallucinations in retrieval-augmented generation via semantic-level internal reasoning graph," 1 2026. [Online]. Available: http://arxiv.org/abs/2601.03052

[17] M. S. Tamber, F. S. Bao, C. Xu, G. Luo, S. Kazi, M. Bae, M. Li, O. Mendelevitch, R. Qu, and J. Lin, "Benchmarking llm faithfulness in rag with evolving leaderboards," Tech. Rep. [Online]. Available: https://github.com/vectara/FaithJudge

A Reproducibility Checklist

✓

✓ Code repository: `https://github.com/Chirudeva-Reddy/NLP-Proj/tree/NLP-Final`

✓ Inference script: `python demo.py --input <file>`

✓ Random seed fixed (state value)

✓ Test set held out until final evaluation

✓ μ_ℓ, Σ_ℓ, V_ℓ from training split only

✓ No HaluEval hidden states used in μ_ℓ, Σ_ℓ

✓ Layer selection $\mathscr{L}^*$ from validation only

✓ $\geq$50 patching experiments per component type

✓ Weights $\{w_i\}$ and τ from validation set only

✓ Mann-Whitney U p-value reported for E4

✓ Compute budget: 46 GPU-hours

B Full Cross-Dataset Results

C Additional Visualisations

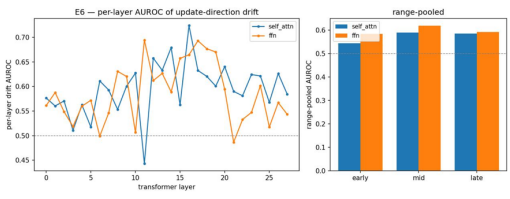

Figure 5: FFN vs. attention decomposition by layer range. Expected: mid-to-late FFN shows strongest signal.

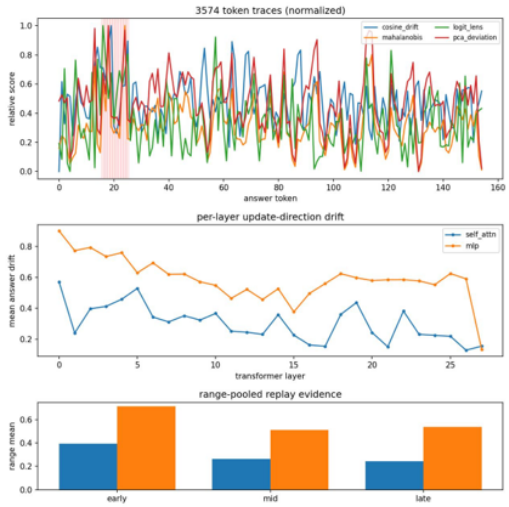

Figure 6: False Negative case

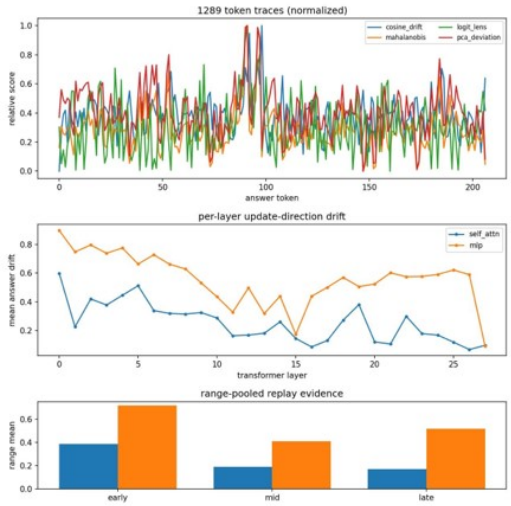

Figure 7: False Positive case

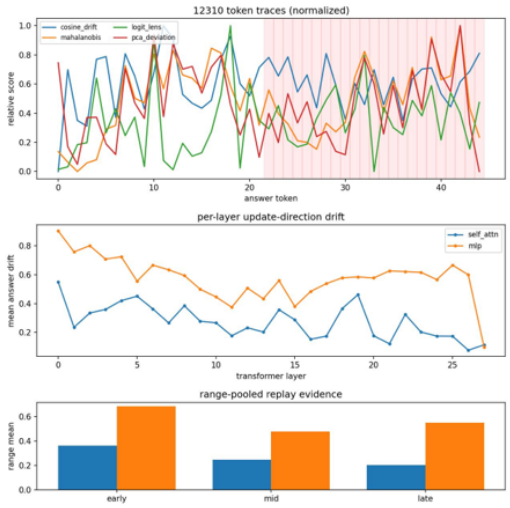

Figure 8: Metric Disagreement case

D FFN vs. Attention — Full Table

Table 8 shows the full FFN vs. attention decomposition. This table is referred in the main text (Section 11).

Table 8: AUROC and FFN / attention signal decomposition by generator. Expected: hallucination signal concentrates in mid-to-late FFN.

component	range	AUROC	mean_hal	mean_faith	gap
Self_attn	early	0.5444	1.10838	1.09464	+0.01373
Self_attn	mid	**0.5903**	**1.15104**	**1.10184**	+0.04919
Self_attn	late	0.5863	1.07852	1.01760	**+0.06093**
FFN	early	0.5852	1.14965	1.13181	+0.01783
FFN	mid	**0.6191**	1.09002	1.05298	**+0.03704**
FFN	late	0.5920	**1.20965**	**1.18060**	+0.02905

E Individual Contribution Statement

The following table documents the primary responsibility of each team member for each section of the report and each component of the implementation. All members are expected to understand the full submission and may be questioned on any part during the compre evaluation.

Team Member	**Report Sections**	**Implementation Components**
Sanya Wadhawan, 2023A7PS0296U	§4 Methodology; §5 Datasets	Methodology framing, dataset setup
Chirudeva Reddy, 2023A7PS0331U	§6 Setup; §12 Discussion; §13 Conclusion	Experiments, analysis
Yusra Hakim, 2022A7PS0004U	§2 Related Work	Related-work review
Joseph Cijo, 2022A7PS0019U	§1 Introduction	Result tables, documentation

We confirm that the above contributions are accurate and that the report was prepared in accordance with the BITS Pilani Gen AI Usage Guidelines (effective 1 April 2026) and the Academic Integrity and Prevention of Plagiarism Policy (Office Order S/1/23/2024/16).

Signature:	Signature:	Signature:	Signature:
Sanya Wadhawan	Chirudeva Reddy	Yusra Hakim	Joseph Cijo
2023A7PS0296U	2023A7PS0331U	2022A7PS0004U	2022A7PS0019U

Date: *23/05/2026*