# FROM DETECTION TO MITIGATION: NEURON ATTRIBUTION FOR CONTEXTUAL HALLUCINATION IN RAG

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Retrieval-augmented language models may ignore externally provided evidence when it conflicts with parametric knowledge, leading to contextual hallucination. We introduce NADM (Neuron Attribution for Detection and Mitigation), a unified framework that uses response-conditioned neuron attribution as a shared internal signal for contextual hallucination detection and mitigation. NADM represents a generated response by computing gradient-based attribution from internal neurons to the logits of its realized tokens, and trains a simple linear classifier on the resulting attribution patterns to distinguish context-faithful from contextually hallucinated generation. Across multiple datasets and model scales, the NADM detector achieves strong discrimination between context-faithful and contextually hallucinated generation. We further examine the causal role of the key neurons identified by the detector and find that, although they participate in context–parametric competition, transferring their activations alone reproduces only a small fraction of the model's context-versus-parametric preference difference. This motivates using attribution as a readout rather than directly manipulating the selected neurons. At inference time, NADM samples alternative responses, scores them with the same frozen detector, and selects alternatives with stronger evidence of contextual faithfulness. This yields an inference-time intervention that requires no update to the base-model parameters and does not assume that direct manipulation of the selected neurons is sufficient to reproduce the desired source preference. Together, NADM connects detection and mitigation through the same response-conditioned attribution signal.

## 1 INTRODUCTION

Retrieval-augmented language models are expected to answer according to evidence supplied at inference time rather than blindly reproduce facts memorized during pretraining (Lewis et al., 2020). This expectation becomes especially important when retrieved evidence updates, overrides, or contradicts parametric knowledge. Under such context–parametric conflicts, a model may either follow the external evidence or fall back to memorized knowledge (Longpre et al., 2021; Xie et al., 2024; Xu et al., 2024). We refer to the former behavior as *context-faithful* and the latter as *parametric*. This distinction is conditioned on the supplied context rather than on factual correctness with respect to the external world. Figure 1 illustrates this failure mode: the same query may lead either to grounded generation or to contextual hallucination depending on which information source dominates.

Reliable RAG therefore requires two complementary capabilities: *detecting* when a generated answer is unsupported by relevant retrieved evidence and *mitigating* such behavior when the model is inclined to follow conflicting parametric knowledge. Existing detectors can be broadly organized by whether they rely on output-level or model-internal signals. Output-level methods use self-evaluation, sampling consistency, semantic uncertainty, or explicit support checking against retrieved evidence (Kadavath et al., 2022; Manakul et al., 2023; Farquhar et al., 2024; Es et al., 2024; Hu et al., 2024). Internal-state methods instead exploit hidden representations, attention dynamics, or mechanistically motivated features (Chen et al., 2024; Kossen et al., 2024; Sun et al., 2025; Gao et al., 2026). In parallel, contextual faithfulness has been improved through prompting, alignment objectives, and test-time manipulation of internal computation (Zhou et al., 2023; Bi et al., 2025; Zhang et al., 2025;



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.

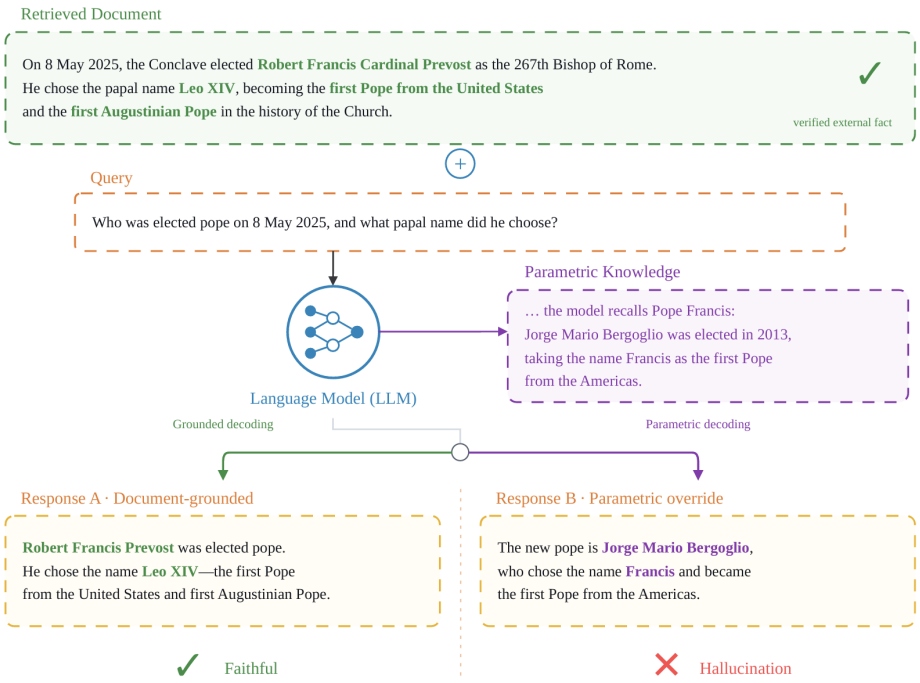


Figure 1: Context–parametric conflict can yield either a context-faithful response or a parametric override that constitutes contextual hallucination.

Li et al., 2023; Jin et al., 2024; Li et al., 2025). These lines of work are usually developed separately: detectors estimate whether a response is reliable, whereas mitigation methods alter generation once conflict is present. This separation raises a natural question: can the same internal signal that reliably detects context-unfaithful behavior also support inference-time mitigation?

We introduce NADM (Neuron Attribution for Detection and Mitigation), a unified framework built around *gradient-based neuron attribution* as such a signal. Specifically, for each generated token NADM measures how fixed FFN neuron coordinates contribute to its pre-softmax logit and aggregates these attribution patterns over the evaluated response. Our premise is not that a small set of neurons stores a shared semantic concept of "faithfulness." Questions, contexts, and answers in our setting are semantically heterogeneous, so such a shared semantic interpretation is neither required nor assumed. Instead, we treat the *pattern of attribution* over fixed neuron identities as a predictive internal signal: the signal lies in whether the distribution of neural contributions to realized predictions changes systematically between context-faithful and contextually hallucinated generation. Because neuron identities are preserved across examples, a coordinate can be useful for detection when its relative contribution exhibits a stable statistical relationship with response state, even when the semantic content of those examples differs substantially. Building on prior work on neuron-level attribution and localization (Dai et al., 2022), we use these response-conditioned attribution patterns as features for a lightweight classifier. The representation is tied to the model's realized output trajectory, rather than to hidden-state geometry alone, and preserves the identities of the contributing neurons across examples.

Strong predictive separation, however, does not imply that the selected neurons independently determine the model's information-source choice. We therefore conduct causal intervention analyses on the neurons selected by the NADM detector. Donor-to-receiver activation transfer shows that the identified neurons participate in context–parametric competition, but transferring the full selected neuron set reproduces only a limited fraction of the unpatched donor–receiver preference difference. Additional temporal interventions show the same pattern when activations are transferred throughout answer generation. We therefore interpret the selected neurons as a predictive readout that is causally involved in source preference but insufficient on its own to reproduce that preference.

This distinction matters for mitigation. Several recent methods directly manipulate localized internal components, including attention heads, FFNs, or activation directions associated with contextual and parametric information (Li et al., 2023; Jin et al., 2024; Sun et al., 2025; Li et al., 2025). Such approaches can be effective, but direct manipulation is most naturally justified when the localized representation provides a sufficiently stable control handle across inputs. That assumption need not hold uniformly: causal localization can depend on the intervention definition (Zhang & Nanda, 2024), components may jointly support contextual and parametric information (Li et al., 2025), and circuit structure can vary across prompt formulations (Franco et al., 2026). A predictive detector has a weaker requirement: its internal signature must remain informative about response state without being a complete or invariant causal mediator.

For mitigation, NADM preserves the model's unmodified computation and reuses its detector to *select among candidate generations that the model can already produce*. Rather than forcing a desired internal activation pattern, the intervention lets the base model generate alternative trajectories and uses the detector only to compare their attribution signatures. At inference time, we sample alternative candidate responses and score each complete response with the same frozen detector. Candidates with stronger evidence of contextual faithfulness can then replace the greedy baseline generation. This turns the detector into an internal candidate scorer and yields a weight-update-free mitigation strategy without assuming that direct manipulation of the selected neurons is sufficient to reproduce the desired source preference.

Overall, NADM uses the same response-conditioned attribution signal first to identify context-unfaithful generation and then to guide candidate selection toward more context-faithful responses. In this way, NADM links detection and mitigation while avoiding direct activation manipulation of the selected neurons.

## 2 Related Work

### 2.1 Contextual hallucination detection in RAG

Hallucination detection methods can be broadly grouped by whether they rely on output-level or model-internal signals. Output-level approaches estimate reliability from generated responses or their agreement with external evidence: P(True) elicits self-assessed answer correctness (Kadavath et al., 2022), SelfCheckGPT measures consistency across stochastic generations (Manakul et al., 2023), semantic entropy quantifies uncertainty over semantically equivalent generations (Farquhar et al., 2024), and RAGAS and RefChecker assess whether generated claims are supported by retrieved evidence (Es et al., 2024; Hu et al., 2024). Internal-state methods instead exploit signals from the model's own computation. INSIDE/EigenScore uses the eigenspectrum of internal representations (Chen et al., 2024), SEP predicts semantic uncertainty from hidden states (Kossen et al., 2024), and ReDeEP combines signals from Knowledge FFNs and Copying Heads (Sun et al., 2025). NADM follows the internal-signal direction but uses *response-conditioned neuron attribution to realized-token logits* as its detection representation, rather than output consistency, hidden-state geometry, attention dynamics, or prespecified mechanistic components.

### 2.2 Mitigating context–parametric conflict

Methods for mitigating context–parametric conflict either update model parameters or intervene only at inference time. Context-DPO improves contextual reliance through preference optimization (Bi et al., 2025). Among methods that leave the base-model weights unchanged, context-aware decoding (CAD) contrasts next-token distributions with and without context (Shi et al., 2024), while COIECD adaptively modifies decoding using contextual information-entropy signals to identify and resolve knowledge conflicts (Yuan et al., 2024). Internal-intervention approaches include Inference-Time Intervention (ITI), which shifts selected attention-head activations along learned directions (Li et al., 2023), and AARF, which modulates the contributions of Copying Heads and Knowledge FFNs (Sun et al., 2025). NADM instead preserves the model's internal computation during candidate generation and reuses its attribution-based detector to select among model-generated responses.

# 3 DETECTING CONTEXTUAL HALLUCINATION

We first evaluate the detection component of NADM and ask whether neuron attribution provides a reliable internal signal for contextual hallucination. Rather than judging a generation only from its surface form or hidden-state geometry, we characterize how fixed internal units contribute to the tokens in the generated response. We then train a lightweight detector over these response-level attribution patterns and evaluate it across datasets and model families.

## 3.1 METHOD

Let $x$ denote the model input containing retrieved evidence and a query, and $r = (y_1, \ldots, y_T)$ the generated response.

Assume a fixed set of $D$ differentiable scalar internal units shared across examples. For each response token $y_t$, let $a_{j,t}$ be the activation of unit $j$ when predicting $y_t$ from $(x, y_{<t})$, and let $z_t(y_t)$ be the corresponding pre-softmax logit. We define Gradient $\times$ Activation attribution as

$$A_{j,t} = a_{j,t} \frac{\partial z_t(y_t)}{\partial a_{j,t}}. \tag{1}$$

$A_{j,t}$ is a local first-order attribution score, interpretable as an approximation to the target-logit change associated with zeroing the activation.

To prevent positions with larger total attribution magnitude from dominating, we normalize positive attribution independently at each prediction step:

$$p_{j,t} = \frac{\max(A_{j,t}, 0)}{\sum_{j'=1}^{D} \max(A_{j',t}, 0) + \epsilon}, \tag{2}$$

where $\epsilon > 0$ ensures numerical stability. We retain positive attribution to characterize units that locally support the realized token.

We then aggregate attribution over the response:

$$\phi_j(r, x) = \frac{1}{T} \sum_{t=1}^{T} p_{j,t}, \tag{3}$$

yielding

$$\boldsymbol{\phi}(r, x) = [\phi_1(r, x), \ldots, \phi_D(r, x)] \in \mathbb{R}^D. \tag{4}$$

The representation is *response-conditioned*, since attribution targets the logits of realized tokens along their autoregressive trajectory, and *identity-preserving*, since each coordinate refers to the same internal unit across examples. We therefore use the attribution pattern as a distributed predictive signal rather than interpreting individual neurons as semantic "faithfulness" units.

For each model–dataset setting, $\boldsymbol{\phi}(r, x)$ is fed to an $\ell_1$-regularized logistic regression classifier, which forms the NADM detector. Throughout short-form detection, PARAMETRIC responses form the positive class. For SQuAD2, we construct an analogous pair of references: the dataset-provided answer supported by the passage serves as the context-faithful reference, while the base model's closed-book answer to the same question serves as the parametric reference. Context-conditioned generations are labeled against these two references using the same short-answer matching rule. The $\ell_1$ penalty induces sparse predictive coordinates while the base language model remains frozen. We define the negative classifier logit as the contextual-faithfulness score $S_F(r, x)$, with larger values indicating stronger evidence of context-supported generation. The same response-level score is reused by the NADM reranker in Section 5.

## 3.2 EXPERIMENTAL SETUP

**Datasets and models.** We evaluate on ConFiQA-QA/MR/MC (Bi et al., 2025) and 6,000 examples sampled from the answerable subset of SQuAD 2.0 (Rajpurkar et al., 2018). ConFiQA explicitly constructs context–parametric conflicts. SQuAD2 is not natively a context–parametric conflict

benchmark; we use it as an auxiliary naturally grounded setting and derive model-specific parametric references from closed-book responses, as detailed in Appendix A.1. Experiments use Llama-2-7B-Chat, Llama-2-13B-Chat (Touvron et al., 2023), and Mistral-7B-Instruct (Jiang et al., 2023); all base-model parameters remain frozen.

**Baselines and evaluation.** For short-form detection, we compare against P(True), RefChecker, SelfCheckGPT, EigenScore, SEP, and ReDeEP (Kadavath et al., 2022; Hu et al., 2024; Manakul et al., 2023; Chen et al., 2024; Kossen et al., 2024; Sun et al., 2025). We use repeated stratified five-fold cross-validation with five repeats and report AUROC as the primary metric together with Macro-F1. Detailed dataset construction, baseline implementations, and cross-validation procedures are provided in Appendix A.

### 3.3 Results and Analysis

Table 1 reports results on ConFiQA-QA/MR/MC and SQuAD2. Across the 12 model–dataset settings, the NADM detector achieves the highest AUROC in every case. Its Macro-F1 ranks first in five settings and second in the remaining seven; on Mistral-7B-Instruct, NADM is best on both metrics for all four datasets.

Table 1: Contextual-hallucination detection on short-form benchmarks. Best results within each model–dataset block are shown in **bold**, and second-best results are underlined.

| | ConFiQA-QA | | ConFiQA-MR | | ConFiQA-MC | | SQuAD2 | |
|---|---|---|---|---|---|---|---|---|
| Method | AUROC | Macro-F1 | AUROC | Macro-F1 | AUROC | Macro-F1 | AUROC | Macro-F1 |
| *Llama-2-7B-Chat* | | | | | | | | |
| P(True) | 0.5221 | 0.7171 | 0.5453 | 0.6949 | 0.5496 | 0.8068 | 0.6629 | 0.6800 |
| RefChecker | 0.8276 | 0.7708 | 0.6517 | 0.7740 | 0.6570 | 0.7729 | 0.5660 | 0.6617 |
| SelfCheckGPT | 0.5916 | 0.6782 | 0.5397 | 0.7062 | 0.4586 | 0.4986 | 0.6260 | 0.6558 |
| EigenScore | 0.5890 | 0.5254 | 0.6522 | 0.6059 | 0.6340 | 0.6321 | 0.6239 | 0.6942 |
| SEP | 0.9337 | **0.9169** | 0.9386 | **0.9364** | 0.9408 | **0.9308** | 0.8185 | **0.8158** |
| ReDeEP | 0.7455 | 0.6433 | 0.7313 | 0.6847 | 0.6792 | 0.7528 | 0.6271 | 0.5505 |
| **NADM** | **0.9877** | 0.9120 | **0.9769** | 0.9154 | **0.9782** | 0.8862 | **0.8529** | 0.7361 |
| *Llama-2-13B-Chat* | | | | | | | | |
| P(True) | 0.5463 | 0.6426 | 0.5569 | 0.6194 | 0.5573 | 0.7109 | 0.6488 | 0.6817 |
| RefChecker | 0.7889 | 0.7113 | 0.6955 | 0.6645 | 0.4944 | 0.6018 | 0.5245 | 0.6583 |
| SelfCheckGPT | 0.5589 | 0.5500 | 0.5195 | 0.5404 | 0.4795 | 0.5058 | 0.6366 | 0.6792 |
| EigenScore | 0.5781 | 0.5658 | 0.5028 | 0.4890 | 0.5022 | 0.4826 | 0.5748 | 0.6358 |
| SEP | 0.9325 | 0.8855 | 0.9479 | **0.8968** | 0.9581 | **0.9127** | 0.8337 | **0.8025** |
| ReDeEP | 0.6614 | 0.6677 | 0.6766 | 0.6382 | 0.6955 | 0.7306 | 0.6436 | 0.5938 |
| **NADM** | **0.9747** | **0.8902** | **0.9486** | 0.8687 | **0.9670** | 0.8759 | **0.8920** | 0.7946 |
| *Mistral-7B-Instruct* | | | | | | | | |
| P(True) | 0.5538 | 0.6935 | 0.5682 | 0.6724 | 0.5615 | 0.7486 | 0.6712 | 0.6849 |
| RefChecker | 0.8246 | 0.7481 | 0.7118 | 0.6849 | 0.6084 | 0.6798 | 0.5847 | 0.6625 |
| SelfCheckGPT | 0.6037 | 0.6892 | 0.5476 | 0.7151 | 0.5012 | 0.5583 | 0.6418 | 0.6710 |
| EigenScore | 0.6124 | 0.6018 | 0.5261 | 0.5027 | 0.5198 | 0.5416 | 0.5964 | 0.6468 |
| SEP | 0.9568 | 0.9287 | 0.9547 | 0.9391 | 0.9604 | 0.9386 | 0.8251 | 0.7246 |
| ReDeEP | 0.7589 | 0.6715 | 0.7422 | 0.6968 | 0.7167 | 0.7741 | 0.6598 | 0.6127 |
| **NADM** | **0.9919** | **0.9519** | **0.9864** | **0.9521** | **0.9860** | **0.9532** | **0.8373** | **0.7507** |

Across the short-form experiments, the attribution readout remains effective across all three model configurations spanning two model families and beyond explicitly constructed conflicts, including SQuAD2. The consistent ranking across architectures and datasets supports treating the attribution pattern as a distributed predictive signal rather than evidence that individual neurons encode a semantic notion of "faithfulness."

Prediction alone does not establish causal sufficiency. We therefore next test whether the NADM-selected neurons participate causally in context–parametric competition and whether transferring them alone is sufficient to reproduce the corresponding source preference.

Table 2: Natural activation transfer with the frozen 134-neuron set. Values are donor-directed changes in the context-minus-parametric logit margin $M$; brackets, when shown, denote 95% confidence intervals. The first transfer moves activations between context and prior-knowledge instructions; the second moves activations from either source-instructed run into the standard prompt.

| Neuron set | Context instruction ↔ Prior-knowledge instruction | Either source instruction → Standard prompt |
|---|---|---|
| NADM-selected neurons | $0.2203$ $[0.1812, 0.2623]$ | $0.1088$ $[0.0889, 0.1294]$ |
| Random matched neurons | $0.0029$ | $0.0032$ |
| Layer-matched high-attribution neurons | $0.0388$ $[0.0222, 0.0551]$ | $0.0226$ $[0.0138, 0.0316]$ |

# 4 Causal Role of NADM-Selected Neurons

High detection accuracy does not reveal whether the neurons selected by the NADM detector merely predict response state or also participate in the underlying computation. We therefore test two progressively stronger claims on Llama-2-7B-Chat using ConFiQA-QA: whether changing these neurons can causally shift the model's preference between retrieved context and parametric knowledge, and whether they are sufficient to reproduce that preference on their own. We analyze 134 neurons whose attribution features repeatedly receive nonzero, sign-consistent coefficients from the $\ell_1$-regularized detector fitted on the training split. Their identities are fixed before evaluation on 228 held-out questions. Full selection criteria, prompt construction, statistics, and additional causal analyses are provided in Appendix B.

## 4.1 Donor-to-Receiver Activation Transfer Establishes Causal Involvement

To create controlled runs with known source preferences, we vary only the source instruction while holding the context, question, and candidate answers fixed. One prompt explicitly instructs the model to answer from the retrieved context, while the other instructs it to ignore the context and rely on prior knowledge. We also evaluate the original baseline prompt, which contains no source-specific instruction. These modified instructions are used only for the causal intervention analysis.

We intervene at the first token position at which the context-supported and parametric candidate answers diverge. This is the earliest position where the model's logits directly distinguish the two answers. Let $v_C$ and $v_P$ denote the two candidate tokens and define the source-preference margin as

$$M = z(v_C) - z(v_P), \tag{5}$$

so larger values indicate stronger preference for the context-supported answer. We call the run from which activations are copied the *donor* and the run whose selected activations are replaced the *receiver*; the receiver prompt and all unpatched activations remain unchanged.

We first transfer activations bidirectionally between the context-instructed and prior-knowledge-instructed runs. We then transfer activations from either source-instructed run into the standard baseline prompt. Effects are sign-aligned so that positive values always indicate movement toward the donor's source preference. Table 2 shows clear donor-directed shifts for the NADM-selected neurons, whereas an exactly layer- and size-matched random neuron set has effects close to zero. A second control uses 134 non-selected neurons with high mean positive normalized attribution on the training responses, matched by layer to the selected set; its effects are substantially smaller.

These interventions establish causal involvement beyond predictive correlation: changing only the NADM-selected neurons systematically moves the receiver toward the donor's source preference. We next ask whether this causal effect is large enough for the neurons to reproduce that preference on their own.

## 4.2 Causal Involvement Does Not Imply Sufficiency

Before intervention, the context-instructed and prior-knowledge-instructed runs differ substantially in the same source-preference margin. If the selected neurons were sufficient on their own to reproduce the model's source preference, replacing all of their activations with donor values should recover a substantial fraction of this unpatched donor–receiver difference.

The mean unpatched margin difference is $3.1530$ between the context- and prior-knowledge-instructed runs, and $1.5809$ between either source-instructed run and the standard prompt after direction alignment. Relative to these full-model differences, transferring all $134$ selected neurons recovers only

$$\frac{0.2203}{3.1530} \approx 7.0\%, \qquad \frac{0.1088}{1.5809} \approx 6.9\%. \tag{6}$$

Thus, the selected neurons have a reproducible causal effect but reproduce only a small fraction of the source-preference difference expressed by the full model. These ratios quantify recovery of the behavioral margin under this intervention and should not be interpreted as the fraction of information about source preference stored in the selected neurons.

We additionally examine where this effect appears during answer generation using a separate protocol on $276$ held-out questions. Patching the $76$ positive-coefficient neurons throughout teacher-forced answer prediction produces significant donor-directed shifts in both transfer directions, but the average aligned shift is approximately $0.0400$ relative to an unpatched margin difference of $5.0244$, or about $0.8\%$ recovery. Patching the same neurons only at the final prompt position yields negligible first-token effects. The two temporal interventions use different behavioral endpoints; full protocols and statistics are reported in Appendix B.3.

Overall, the NADM-selected neurons participate causally in context–parametric competition, but transferring their naturally occurring activations alone is insufficient to reproduce the donor's source preference. Additional low-dimensional analyses in Appendix B.4 characterize how the causal effect is organized within the selected neurons. This motivates using attribution as a predictive readout for candidate selection rather than directly overwriting the selected neurons.

## 5 Mitigating Contextual Hallucination

Motivated by the causal results above, NADM reuses its attribution-based detector as an internal scorer over candidate responses rather than directly manipulating the selected neurons. The frozen base model generates alternative responses, and the same response-conditioned attribution signal used by the NADM detector ranks these candidates by contextual faithfulness. This yields a weight-update-free mitigation strategy without assuming that direct manipulation of the selected neurons is sufficient to reproduce the desired source preference.

### 5.1 Method

The NADM reranker operates over candidate responses generated by the frozen base model. For an input $x$, let

$$\mathcal{C}(x) = \{r^{(0)}, r^{(1)}, \ldots, r^{(K)}\} \tag{7}$$

denote candidate responses produced by the frozen base model, where $r^{(0)}$ is the Vanilla (greedy) baseline candidate. We compute the same response-level attribution representation $\boldsymbol{\phi}(r^{(k)}, x)$ defined in Section 3.1 and score each candidate with the frozen contextual-faithfulness score $S_F$.

Candidates are sampled from the model's own predictive distribution by branching from the top-$M$ first-token alternatives and drawing $N$ sampled continuations from each branch. The Vanilla baseline candidate is always retained in the candidate set. All reported mitigation experiments use $M = 5$ and $N = 3$; duplicate sampled responses are removed before scoring. Candidate construction uses neither contextual nor parametric reference answers nor test labels. Full sampling details are provided in Appendix C.2.

The highest-scoring candidate is

$$r^* = \arg\max_{r \in \mathcal{C}(x)} S_F(r, x). \tag{8}$$

We quantify its score improvement over the Vanilla baseline candidate as

$$\Delta S_F = S_F(r^*, x) - S_F(r^{(0)}, x). \tag{9}$$

The selected response is

$$r_{\text{final}} = \begin{cases} r^*, & \Delta S_F \geq \tau, \\ r^{(0)}, & \Delta S_F < \tau, \end{cases} \tag{10}$$

where the acceptance margin $\tau$ is selected on a disjoint calibration set and frozen before test evaluation.

Table 3: Mitigation on ConFiQA and SQuAD2. Columns compare Vanilla decoding, four inference-time baselines, and the NADM reranker. Best results within each model–dataset–metric row are shown in **bold**, and second-best results are underlined. Colored arrows indicate the preferred metric direction. For **NADM**, the small colored $\Delta$ at the lower right of each value is the percentage-point change relative to Vanilla decoding.

| Dataset | Metric | Vanilla | CAD | COIECD | ITI | AARF | **NADM** |
|---|---|---|---|---|---|---|---|
| *Llama-2-7B-Chat* | | | | | | | |
| ConFiQA-QA | Faithful ↑ | 85.28 | 88.64 | 87.07 | 84.44 | 87.80 | **91.80** $\Delta$+6.52 |
| | Parametric ↓ | 14.41 | 10.30 | 11.88 | 15.04 | 11.36 | **5.68** $\Delta$-8.73 |
| | $F/(F+P)$ ↑ | 85.55 | 89.59 | 87.99 | 84.88 | 88.55 | **94.17** $\Delta$+8.62 |
| ConFiQA-MR | Faithful ↑ | 55.68 | 57.93 | 59.10 | 56.40 | 54.68 | **61.08** $\Delta$+5.40 |
| | Parametric ↓ | 8.11 | 7.03 | 7.66 | 8.92 | 9.19 | **5.50** $\Delta$-2.61 |
| | $F/(F+P)$ ↑ | 87.29 | 89.18 | 88.53 | 86.34 | 85.61 | **91.75** $\Delta$+4.46 |
| ConFiQA-MC | Faithful ↑ | 51.02 | 54.69 | 52.56 | 49.74 | 53.58 | **57.08** $\Delta$+6.06 |
| | Parametric ↓ | 4.01 | 3.67 | 3.33 | 4.52 | 4.35 | **3.24** $\Delta$-0.77 |
| | $F/(F+P)$ ↑ | 92.71 | 93.71 | 94.05 | 91.67 | 92.49 | **94.63** $\Delta$+1.92 |
| SQuAD2 | Faithful ↑ | 57.50 | 59.17 | 58.17 | 58.67 | 59.83 | **60.92** $\Delta$+3.42 |
| | Parametric ↓ | 9.17 | 7.67 | 6.67 | 8.50 | 8.75 | **5.25** $\Delta$-3.92 |
| | $F/(F+P)$ ↑ | 86.25 | 88.53 | 89.72 | 87.34 | 87.24 | **92.07** $\Delta$+5.82 |
| *Llama-2-13B-Chat* | | | | | | | |
| ConFiQA-QA | Faithful ↑ | 77.89 | 80.41 | 79.15 | 76.86 | 81.10 | **82.47** $\Delta$+4.58 |
| | Parametric ↓ | 21.65 | 17.75 | 19.36 | 22.22 | 17.53 | **14.78** $\Delta$-6.87 |
| | $F/(F+P)$ ↑ | 78.25 | 81.91 | 80.35 | 77.57 | 82.23 | **84.81** $\Delta$+6.56 |
| ConFiQA-MR | Faithful ↑ | 42.88 | 46.13 | 47.57 | 44.68 | 45.14 | **50.00** $\Delta$+7.12 |
| | Parametric ↓ | 13.15 | 10.99 | 12.25 | 12.52 | 10.09 | **8.56** $\Delta$-4.59 |
| | $F/(F+P)$ ↑ | 76.53 | 80.76 | 79.52 | 78.11 | 81.73 | **85.38** $\Delta$+8.85 |
| ConFiQA-MC | Faithful ↑ | 41.13 | 44.28 | 43.17 | 41.89 | 40.19 | **47.87** $\Delta$+6.74 |
| | Parametric ↓ | 8.53 | 7.42 | 6.66 | 9.30 | 7.85 | **6.14** $\Delta$-2.39 |
| | $F/(F+P)$ ↑ | 82.82 | 85.64 | 86.64 | 81.83 | 83.66 | **88.63** $\Delta$+5.81 |
| SQuAD2 | Faithful ↑ | 58.08 | 59.25 | 58.67 | 58.75 | 59.92 | **60.67** $\Delta$+2.59 |
| | Parametric ↓ | 3.83 | 3.42 | **2.83** | 3.67 | 3.75 | 2.92 $\Delta$-0.91 |
| | $F/(F+P)$ ↑ | 93.81 | 94.55 | 95.39 | 94.13 | 94.11 | **95.41** $\Delta$+1.60 |
| *Mistral-7B-Instruct* | | | | | | | |
| ConFiQA-QA | Faithful ↑ | 43.87 | 47.87 | 46.62 | 42.54 | 49.21 | **51.88** $\Delta$+8.01 |
| | Parametric ↓ | 25.94 | 22.60 | 20.68 | 27.19 | 24.35 | **18.02** $\Delta$-7.92 |
| | $F/(F+P)$ ↑ | 62.84 | 67.93 | 69.27 | 61.00 | 66.89 | **74.22** $\Delta$+11.38 |
| ConFiQA-MR | Faithful ↑ | 35.41 | 38.83 | 37.66 | 36.49 | 40.09 | **42.70** $\Delta$+7.29 |
| | Parametric ↓ | 21.08 | 19.55 | **17.93** | 19.91 | 21.44 | 18.20 $\Delta$-2.88 |
| | $F/(F+P)$ ↑ | 62.68 | 66.51 | 67.75 | 64.70 | 65.15 | **70.12** $\Delta$+7.44 |
| ConFiQA-MC | Faithful ↑ | 23.21 | 28.24 | 26.28 | 21.67 | 29.78 | **33.28** $\Delta$+10.07 |
| | Parametric ↓ | 16.04 | 15.10 | 13.48 | 17.15 | 16.21 | **13.31** $\Delta$-2.73 |
| | $F/(F+P)$ ↑ | 59.13 | 65.16 | 66.09 | 55.82 | 64.75 | **71.43** $\Delta$+12.30 |
| SQuAD2 | Faithful ↑ | 66.00 | 68.42 | 67.08 | 67.17 | 69.83 | **71.75** $\Delta$+5.75 |
| | Parametric ↓ | 1.17 | 1.00 | 0.67 | 1.08 | 1.33 | **0.58** $\Delta$-0.59 |
| | $F/(F+P)$ ↑ | 98.26 | 98.56 | 99.02 | 98.41 | 98.13 | **99.19** $\Delta$+0.93 |

### 5.2 Experimental Setup

**Datasets, models, and baselines.** We evaluate mitigation on ConFiQA-QA/MR/MC and SQuAD2 using Llama-2-7B-Chat, Llama-2-13B-Chat, and Mistral-7B-Instruct. For each model–dataset pair, the NADM detector and intervention threshold are fit using only the corresponding training and calibration data and are frozen before held-out evaluation. We compare against Vanilla (greedy) decoding and four inference-time baselines that do not update the base-model weights: CAD (Shi et al., 2024), COIECD (Yuan et al., 2024), ITI (Li et al., 2023), and AARF (Sun et al., 2025). Implementation and calibration details are given in Appendix C.

**Evaluation.** For ConFiQA and SQuAD2, using the two-reference labeling procedure described in Appendix C.3, we report Faithful rate $F = N_F/N$, Parametric rate $P = N_P/N$, and conditional faithfulness $F/(F + P)$, where $N_F$ and $N_P$ count responses assigned exclusively to the context-faithful and parametric states. Complete response-labeling rules and calibration details are provided in Appendix C.

### 5.3 Results and Analysis

Table 3 compares Vanilla decoding, CAD, COIECD, ITI, AARF, and the NADM reranker using the three faithfulness metrics defined above. For NADM, we additionally show the percentage-point change from Vanilla decoding.

Across the 36 model–dataset–metric comparisons, NADM ranks first in 34 and second in the remaining two. It achieves the highest Faithful rate and $F/(F + P)$ in all 12 model–dataset settings, and the lowest Parametric rate in 10 of 12. The two exceptions are Llama-2-13B-Chat on SQuAD2 and Mistral-7B-Instruct on ConFiQA-MR, where COIECD attains slightly lower Parametric rates. Relative to Vanilla decoding, NADM increases the Faithful rate by $2.59$–$10.07$ percentage points, decreases the Parametric rate by $0.59$–$8.73$ points, and improves $F/(F + P)$ by $0.93$–$12.30$ points.

Reporting faithful and parametric rates together with $F/(F + P)$ distinguishes changes in contextual preference from shifts into unresolved response categories.

More broadly, these experiments test whether the NADM detector from Section 3 can serve as a generation-time scorer rather than only a post-hoc diagnostic. Candidate responses are produced by the model's unmodified computation, and attribution is used only to discriminate among those model-generated alternatives.

## 6 Conclusion

We presented NADM (Neuron Attribution for Detection and Mitigation), a unified framework for detecting and mitigating contextual hallucination in retrieval-augmented language models using neuron attribution as a shared internal signal. The NADM detector represents generated responses through response-conditioned attribution patterns and uses them to distinguish context-faithful from contextually hallucinated generation. Causal intervention analyses of the key neurons identified by the detector show that they participate in context–parametric competition, but transferring their activations alone reproduces only a small fraction of the full context-versus-parametric preference difference. This distinction motivates treating the attribution signal as a predictive readout rather than directly overwriting the selected neurons.

Building on this observation, the NADM reranker samples alternative responses from the frozen base model and reuses the same NADM detector to score and select candidates. NADM therefore connects detection and inference-time mitigation without updating the base-model parameters or assuming that a small set of neurons forms a complete causal bottleneck. More broadly, our results highlight a useful separation between *where an internal signal is predictive* and *what can be safely controlled through direct intervention*, and suggest that attribution-based readouts can provide a practical bridge between diagnosing and mitigating contextual hallucination.

### AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT, to assist with manuscript drafting and language editing, organization and restructuring of the paper, LaTeX formatting, literature and reference checks, preparation and revision of a scientific figure, and feedback on the presentation and interpretation of existing experimental analyses. The numerical results reported in the paper were taken from the authors' experimental outputs rather than generated by the AI tools. All AI-assisted text, citations, figures, and analytical suggestions were reviewed and revised by the authors against the underlying experiments and source materials. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work studies the detection and mitigation of contextual hallucination in language models and does not involve human subjects or the collection of private or sensitive personal data. Our experiments use publicly released benchmark datasets and pretrained language models.

### REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the methodology, experimental setup, and evaluation protocols needed to reproduce our results in the main paper and appendices. These include dataset construction and response labeling, attribution computation, detector training and cross-validation, neuron selection, causal intervention protocols and statistical analyses, mitigation calibration, candidate generation, and baseline implementations. We also report the model configurations, evaluation procedures, and relevant hyperparameters used throughout our experiments. Upon acceptance, we will release the source code for NADM together with the scripts needed to reproduce the reported experiments.

## REFERENCES

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-DPO: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10280–10300. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.536.

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 *International Conference on Learning Representations*, 2024.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.581.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-demo.16.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Gabriel Franco, Lucas M. Tassis, Azalea Rohr, and Mark Crovella. Finding highly interpretable prompt-specific circuits in language models. *arXiv preprint arXiv:2602.13483*, 2026.

Linfeng Gao, Qinggang Zhang, Baolong Bi, Bo Zeng, Zheng Yuan, Zerui Chen, Zhimin Wei, Shenghua Liu, Linlong Xu, Longyue Wang, Weihua Luo, and Jinsong Su. Beyond black-box

interventions: Latent probing for faithful retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29981–30000. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1499.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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 *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215. Association for Computational Linguistics, 2024. 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.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34074–34104. PMLR, 2025.

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

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.565.

Potsawee Manakul, Adian Liusie, and Mark J. F. 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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-2124.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-short.69.

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 *International Conference on Learning Representations*, 2025.

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

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.486.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3903–3922. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.234.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21863–21882. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1062.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.968.

# A Additional Detection Details

## A.1 Datasets and models

ConFiQA-QA, ConFiQA-MR, and ConFiQA-MC (Bi et al., 2025) provide increasingly complex context–parametric conflict settings: QA is single-hop, whereas MR and MC introduce multi-hop reasoning. We additionally sample 6,000 examples from the answerable subset of SQuAD 2.0 (Rajpurkar et al., 2018). Unlike ConFiQA, SQuAD2 was not designed to provide context–parametric conflicts and supplies only an answer supported by the passage. We therefore adapt it into an auxiliary detection benchmark on a model-by-model basis.

For each SQuAD2 question, let $A_C$ denote the dataset-provided answer supported by the passage. We first query the same base model with the question alone, without the passage, and record its closed-book answer as $A_P$. We then provide the original passage and query the model normally, obtaining a context-conditioned response $A_G$. Using the same normalized short-answer matching rule as in the rest of the short-form evaluation, $A_G$ is labeled FAITHFUL when it matches $A_C$ but not $A_P$. It is labeled PARAMETRIC when it matches $A_P$ while $A_P$ does not match $A_C$. Cases that match both references, match neither reference, or otherwise do not provide an unambiguous distinction are excluded from the binary detection set. Thus, SQuAD2 serves as a naturally grounded auxiliary setting in which the context-supported reference comes from the dataset and the competing parametric reference is induced from the evaluated model's own closed-book response; we do not assume that SQuAD2 itself natively contains context–parametric conflicts.

Experiments use Llama-2-7B-Chat and Llama-2-13B-Chat (Touvron et al., 2023), together with Mistral-7B-Instruct (Jiang et al., 2023) to test a different model family. All base-model parameters are frozen; only the linear readout used by the NADM detector is fitted.

### A.2 Attribution Coordinates and Short-Form Labels

The analyzed internal units are scalar FFN intermediate coordinates immediately before each transformer's down-projection. Their total number therefore depends on the model architecture. Attribution is computed for all such coordinates before any classifier-induced sparsification, using the realized token's pre-softmax logit as the target. Positive attribution is normalized across all analyzed coordinates at each prediction step as in Section 3.1, and EOS is not included in the response content average.

For ConFiQA, the supplied counterfactual answer is the context-faithful reference and the corresponding original answer is the parametric reference. For SQuAD2, the context-faithful reference $A_C$ is the dataset-provided answer and the model-specific parametric reference $A_P$ is obtained by closed-book querying as described in Appendix A.1. The subsequent context-conditioned response is matched against these two references; only unambiguous FAITHFUL and PARAMETRIC cases are retained for binary detection.

### A.3 Detection baselines

For short-form evaluation, P(True) (Kadavath et al., 2022) uses the model's self-assessed probability that its response is correct; RefChecker (Hu et al., 2024) extracts claims and verifies them against supplied evidence; SelfCheckGPT (Manakul et al., 2023) estimates hallucination risk from consistency across stochastic generations; EigenScore (Chen et al., 2024) characterizes uncertainty through the eigenspectrum of hidden representations; SEP (Kossen et al., 2024) learns a hidden-state probe associated with semantic uncertainty; and ReDeEP (Sun et al., 2025) combines internal signals associated with retrieved-context utilization and parametric-knowledge contribution.

### A.4 Cross-validation and metrics

For each short-form model–dataset configuration, we use repeated stratified five-fold cross-validation with five independent repeats. PARAMETRIC is the positive class throughout. Each example is scored only when held out, yielding one out-of-fold parametric probability per repeat; we average these probabilities across the five repeats before computing the reported metrics. AUROC is computed directly from the averaged parametric probabilities. For Macro-F1, an averaged probability of at least $0.5$ is classified as PARAMETRIC and a lower probability as FAITHFUL; the two class-wise F1 scores are then averaged with equal weight. Feature selection, regularization selection, and classifier fitting for NADM are performed entirely within the training portion of each fold. Trainable baselines use the same data partitions, and all methods are evaluated on the same base-model responses.

## B Additional Causal Analysis

This appendix details the causal analyses in Section 4: neuron selection, first-divergence activation transfer, temporal activation transfer, and the low-dimensional analysis of the selected neurons.

### B.1 Selection of Neurons for Causal Analysis

Throughout the causal analysis, a *neuron* denotes a scalar FFN intermediate coordinate at the input to a transformer's down-projection, the same coordinates used by the NADM detector. Llama-2-7B-Chat has $32$ layers with $11{,}008$ such coordinates per layer, giving $352{,}256$ candidate coordinates.

Neuron identities are selected using only the $3{,}806$ examples in the official ConFiQA-QA training split. We repeat stratified five-fold cross-validation for five repeats, yielding $25$ detector fits. Within each fold, the fold-training data are class balanced, and an inner $20\%$ validation split selects the $\ell_1$ regularization strength by AUROC. We then record which attribution coordinates receive nonzero classifier coefficients. A neuron is retained if its coefficient is nonzero in at least $20$ of the $25$ fits and its sign agrees in at least $90\%$ of the fits in which it is selected. This procedure yields $134$ neurons, all frozen before held-out causal evaluation.

Of these $134$ neurons, $76$ have positive mean coefficients across the fits in which they are selected and $58$ have negative mean coefficients. Because the detector treats parametric responses as the positive

class, positive coefficients indicate predictive association with parametric generation and negative coefficients indicate the opposite association. Coefficient sign is used only to describe the detector association and is not treated as a fixed causal functional label.

### B.2 First-Divergence Activation Transfer

The main causal experiment evaluates 228 held-out ConFiQA-QA questions. For each conflict example, the context-supported candidate is the answer supported by the supplied counterfactual context, while the parametric candidate is the corresponding original benchmark answer associated with prior knowledge. The protocol evaluates two alternative context-supported answer instantiations per question: the benchmark counterfactual answer and a second counterfactual answer drawn from another example with the same relation type, preferring a similar token length while excluding the current context-supported and parametric answers, their aliases, and answers already present in the context. Results are first averaged over these two answer instantiations within each question.

For each answer instantiation, we construct two runs with the same context, question, and candidate answers but different source instructions. One explicitly instructs the model to answer from the retrieved context, while the other instructs it to ignore the context and rely on prior knowledge. The standard condition uses the original baseline prompt without either source-specific instruction. These source instructions are used only to create controlled donor and receiver runs for the causal intervention analysis.

We append the common token prefix shared by the context-supported and parametric candidate answers and evaluate the first position at which the two answers diverge. If one answer is a strict prefix of the other, EOS is used as the divergent token for the shorter answer. Let $v_C$ and $v_P$ denote the context-supported and parametric divergent tokens. The behavioral endpoint is

$$M_{\text{div}} = z(v_C) - z(v_P), \tag{11}$$

so larger values indicate stronger preference for the context-supported candidate.

For activation transfer, the *donor* is the run from which activations are copied and the *receiver* is the run whose activations are replaced. At the divergence position, we replace all 134 selected receiver down-projection-input coordinates with the corresponding donor values. The receiver prompt, model weights, all non-selected coordinates, and all other positions remain unchanged. We test both directions between the context-instructed and prior-knowledge-instructed runs, and separately transfer each source-instructed run into the standard prompt. Let $s_d = +1$ for a context-instructed donor and $s_d = -1$ for a prior-knowledge-instructed donor. We report the aligned patch effect

$$E_{\text{aligned}} = s_d \left( M_{\text{patch}} - M_{\text{receiver}} \right), \tag{12}$$

and define the corresponding unpatched donor–receiver difference with the same sign convention. Positive values therefore always indicate movement of the receiver toward the donor's source preference.

The main text averages the two transfer directions between the context and prior-knowledge instructions, and separately averages the two transfers into the standard prompt.

We use two layer-matched control neuron sets. The random control contains 134 non-selected neurons sampled to match the selected set's per-layer counts. The high-attribution control contains 134 non-selected neurons, again matched by layer, chosen by their mean positive normalized attribution share on the actual training responses. The random control tests arbitrary matched neuron identity, while the high-attribution control tests whether large attribution alone explains the selected-neuron effect.

To quantify how much of the full donor–receiver difference is reproduced by the intervention, we use a ratio of means,

$$R_{\text{rec}} = \frac{\mathbb{E}[\Delta M_{\text{patch}}]}{\mathbb{E}[\Delta M_{\text{natural}}]}, \tag{13}$$

where $\Delta M_{\text{patch}}$ is the sign-aligned change caused by activation transfer and $\Delta M_{\text{natural}}$ is the corresponding sign-aligned unpatched donor–receiver difference. We use a ratio of means rather than averaging per-example ratios, which would be unstable when an individual unpatched difference is close to zero.

Each held-out question is the statistical unit after averaging its two answer instantiations and, for the summary columns, the paired transfer directions. Confidence intervals use $5{,}000$ question-level bootstrap resamples and two-sided $p$-values use $5{,}000$ question-level sign flips. For the first-divergence and low-dimensional validation analyses, Benjamini–Hochberg correction is applied jointly to the $330$ condition/direction/family tests in the condition summary; the $16$ planned paired control contrasts are corrected as a separate family.

### B.3 Activation Transfer During Answer Generation and at the Final Prompt Position

The temporal experiment uses $276$ held-out questions, balanced between $138$ context-faithful and $138$ parametric baseline responses. In the *conflict* condition, the retrieved context supports a contextual answer $A_C$ that conflicts with the model's parametric answer $A_P$. The matched *consistent* condition replaces the conflicting contextual answer with $A_P$, keeping the question and prompt template otherwise unchanged.

For trajectory-wide transfer, $A_C$ and $A_P$ are replayed separately under teacher forcing. Within each transfer, donor and receiver runs replay the same candidate-token sequence, and the selected receiver activations are replaced with donor activations at every answer-prediction position. EOS is excluded from the answer average. The behavioral endpoint is

$$M_{\mathrm{traj}} = \frac{1}{|A_C|}\sum_t \log P(A_{C,t}) - \frac{1}{|A_P|}\sum_t \log P(A_{P,t}), \tag{14}$$

so larger values indicate greater preference for the contextual candidate.

We separately patch the $76$ neurons with positive detector coefficients and the $58$ neurons with negative coefficients. Table 4 reports both donor-to-receiver directions. The positive-coefficient neurons show significant effects in both directions after FDR correction, whereas the negative-coefficient neurons do not show the same bidirectional pattern on this endpoint.

Table 4: Answer-trajectory natural activation transfer on 276 held-out questions. "Transfer" denotes donor → receiver. Effects are changes in the contextual-minus-parametric mean token log-probability margin.

| Neuron set | Transfer | Effect | 95% CI | FDR $q$ |
|---|---|---|---|---|
| Positive-coefficient neurons (76) | Conflict → Consistent | $+0.0510$ | $[+0.0275, +0.0767]$ | $0.0010$ |
| Positive-coefficient neurons (76) | Consistent → Conflict | $-0.0289$ | $[-0.0506, -0.0082]$ | $0.0240$ |
| Negative-coefficient neurons (58) | Conflict → Consistent | $+0.0279$ | $[+0.0019, +0.0555]$ | $0.0873$ |
| Negative-coefficient neurons (58) | Consistent → Conflict | $-0.0100$ | $[-0.0345, +0.0145]$ | $0.5796$ |

After direction alignment, the positive-coefficient neurons produce an average donor-directed shift of approximately $0.0400$. The corresponding unpatched conflict-versus-consistent margin difference is $5.0244$, so trajectory-wide transfer recovers roughly $0.8\%$ of that behavioral difference.

We also intervene on the same neuron identities only at the final prompt position, immediately before the first answer token. Let $v_C^{(1)}$ and $v_P^{(1)}$ denote the first tokens of the contextual and parametric candidate answers. The final-prompt-position endpoint is

$$M_{\mathrm{first}} = \log P(v_C^{(1)}) - \log P(v_P^{(1)}). \tag{15}$$

Activations are replaced only at this final prefill position; subsequent autoregressive generation proceeds without intervention. For the positive-coefficient neurons, the conflict → consistent effect on $M_{\mathrm{first}}$ is $+0.00136$ with 95% CI $[0, +0.00362]$, while the reverse effect is $-0.00181$ with CI $[-0.00498, 0]$; both have FDR $q = 1$. Effects for the negative-coefficient neurons are likewise negligible.

The trajectory-wide and final-prompt-position experiments therefore provide complementary temporal evidence: the measurable donor-directed effect appears when the selected activations are transferred throughout answer prediction, but not when they are transferred only at the final prompt position. The two protocols use different behavioral endpoints, so their effect magnitudes are interpreted within protocol.

### B.4 Low-Dimensional Structure of the Selected Neurons

We next ask whether the causal effect expressed by the 134 selected neurons is broadly distributed across coordinates or concentrated in a lower-dimensional direction. The basis is learned from 256 training questions that are disjoint from the 228 held-out causal-evaluation questions. For each basis-learning question, we use the same two alternative context-supported answer instantiations described above and evaluate each under the context instruction and the prior-knowledge instruction, yielding $256 \times 2 = 512$ paired activation differences.

Let $a_i^C, a_i^P \in \mathbb{R}^{134}$ denote the selected-neuron activation vectors for the context-instructed and prior-knowledge-instructed runs of basis-learning pair $i$. Using the pooled basis-learning activations, we compute a per-coordinate RMS scale vector $s$ and form standardized source contrasts

$$d_i = (a_i^C - a_i^P) \oslash s, \tag{16}$$

where $\oslash$ denotes elementwise division. We apply uncentered SVD to the $512 \times 134$ matrix with rows $d_i$. The uncentered decomposition preserves the mean context-versus-prior-knowledge difference rather than removing it by centering. The learned basis is frozen before held-out evaluation.

Let $V_k \in \mathbb{R}^{134 \times k}$ contain the first $k$ right-singular vectors. For a held-out donor activation $a_d$ and receiver activation $a_r$, we standardize their natural difference,

$$\delta_{d \to r} = (a_d - a_r) \oslash s, \tag{17}$$

project it onto the learned subspace, and intervene with

$$a_r' = a_r + s \odot V_k V_k^\top \delta_{d \to r}, \tag{18}$$

where $\odot$ denotes elementwise multiplication. Thus, the held-out intervention adds only the Rank-$k$ component of the naturally occurring donor–receiver activation difference.

We summarize descriptive concentration using the fraction of squared singular values captured by the first $k$ directions,

$$E_k = \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_i \sigma_i^2}. \tag{19}$$

The first singular direction explains $55.15\%$ of this activation-difference energy; the first four and eight directions explain $65.04\%$ and $72.02\%$, respectively.

Table 5 reports held-out projected interventions. The Rank-1 intervention on the NADM-selected neurons produces $0.1589$ for transfer between the context and prior-knowledge instructions and $0.0796$ for transfer from either source-instructed run into the standard prompt, corresponding to

$$\frac{0.1589}{0.2203} = 72.1\%, \qquad \frac{0.0796}{0.1088} = 73.2\% \tag{20}$$

of the effect obtained by replacing all 134 selected neurons.

The table includes several controls. "Complement of selected-neuron Rank-1" removes the learned Rank-1 component from the natural selected-neuron difference and patches only the remaining component. "Random direction orthogonal to the learned top-8 subspace" samples a Gaussian direction in the same 134-neuron space, projects out the first eight learned directions, and orthogonalizes the remainder; ten Rank-1 random directions are evaluated. "Rank-1 from random matched neurons" learns the first SVD direction from an exactly layer- and size-matched random 134-neuron set that excludes the NADM-selected neurons. "Rank-1 from high-attribution matched neurons" learns the corresponding direction from the 134 layer-matched non-selected neurons with the highest mean positive normalized attribution on the training responses.

The orthogonal component outside the learned Rank-1 direction retains a smaller but significant effect, so the selected-neuron effect is not strictly one-dimensional. At the same time, Rank-1 transfer accounts for only about $5\%$ of the full unpatched donor–receiver margin difference in either transfer setting. Thus, the 72–73% figures describe compression of the causal effect available within the selected neurons, not recovery of 72–73% of the full unpatched context-versus-parametric margin.

Table 5: Held-out low-dimensional activation transfer for the NADM-selected neurons and matched controls. Effects are donor-directed first-divergence logit-margin changes; brackets denote 95% confidence intervals.

| | Context instruction ↔ Prior-knowledge instruction | | | Either source instruction → Standard prompt | | |
|---|---|---|---|---|---|---|
| Condition | Effect | 95% CI | FDR $q$ | Effect | 95% CI | FDR $q$ |
| All 134 NADM-selected neurons | 0.2203 | $[0.1812, 0.2623]$ | 0.0013 | 0.1088 | $[0.0889, 0.1294]$ | 0.0013 |
| Selected-neuron Rank-1 | 0.1589 | $[0.1376, 0.1806]$ | 0.0013 | 0.0796 | $[0.0691, 0.0905]$ | 0.0013 |
| Selected-neuron Rank-4 | 0.1685 | $[0.1468, 0.1911]$ | 0.0013 | 0.0834 | $[0.0723, 0.0945]$ | 0.0013 |
| Selected-neuron Rank-8 | 0.1873 | $[0.1616, 0.2130]$ | 0.0013 | 0.0908 | $[0.0778, 0.1040]$ | 0.0013 |
| Complement of selected-neuron Rank-1 | 0.0595 | $[0.0294, 0.0898]$ | 0.0024 | 0.0343 | $[0.0199, 0.0495]$ | 0.0013 |
| Random direction orthogonal to learned top-8 subspace | $-0.0017$ | $[-0.0057, 0.0022]$ | 0.6718 | $-0.0002$ | $[-0.0013, 0.0009]$ | 0.8861 |
| Rank-1 from random matched neurons | $-0.0018$ | $[-0.0079, 0.0043]$ | 0.7996 | $-0.0003$ | $[-0.0039, 0.0034]$ | 0.9711 |
| Rank-1 from layer-matched high-attribution neurons | 0.0045 | $[-0.0068, 0.0159]$ | 0.7191 | 0.0055 | $[-0.0010, 0.0119]$ | 0.2454 |

# C Additional Mitigation Details

## C.1 Detector fitting and calibration

For each model–dataset pair, the NADM detector is fitted on the corresponding training data. A disjoint calibration subset from the same dataset is used to select the candidate-acceptance margin $\tau$. The detector, selected attribution coordinates, and acceptance margin are then frozen before evaluation on the held-out test set.

## C.2 Candidate generation

For each input, the Vanilla (greedy) baseline candidate is retained in the candidate set. Candidate branching uses the top-$M$ first-token alternatives, with $N$ sampled continuations generated from each forced first-token branch. All reported mitigation experiments use $M = 5$ and $N = 3$. Sampling uses temperature $0.7$, top-$p$ $0.9$, top-$k = 0$, and a maximum of 96 generated tokens. Duplicate token sequences are removed before attribution scoring. Each complete candidate response is scored with the NADM detector. Candidate construction does not use contextual reference answers, parametric reference answers, or test labels.

## C.3 Response labeling and metrics

For ConFiQA, the context-supported and parametric references are the benchmark counterfactual and original answers, respectively. For SQuAD2, the dataset-provided context-supported answer is the faithful reference, while the same base model's closed-book answer to the question is used as the parametric reference. Generated responses are categorized as FAITHFUL, PARAMETRIC, BOTH, or OTHER by matching against these two references with the benchmark-specific normalization and alias rules. $N_F$ and $N_P$ count responses assigned exclusively to the first two categories. We report $N_F/N$, $N_P/N$, and $N_F/(N_F + N_P)$. The last metric conditions on examples that can be assigned unambiguously to one of the two competing answer sources, while BOTH and OTHER remain represented in the absolute rates through the total test-set denominator.

## C.4 Intervention baselines

We focus on inference-time baselines that do not update the base-model parameters. **CAD** contrasts the next-token distributions obtained with and without the supplied context, amplifying context-dependent evidence during decoding (Shi et al., 2024). **COIECD** uses contextual information-entropy signals to adapt decoding when contextual and parametric knowledge conflict (Yuan et al., 2024). **ITI** intervenes on a selected set of attention heads by shifting their activations along learned directions at inference time (Li et al., 2023). **AARF**, introduced with ReDeEP, increases the contribution of Copying Heads while reducing the contribution of Knowledge FFNs (Sun et al., 2025). All methods are evaluated on the same held-out examples and base-model checkpoints, and method-specific hyperparameters are selected using calibration data without access to test labels.