RENA: DETECTING CONTEXTUAL HALLUCINATIONS IN RAG VIA REFERENCE-CONTRAST NEURON ATLAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) reduces reliance on parametric knowledge by grounding language models in retrieved evidence. However, models may still override accurate context and generate responses from conflicting parametric memory. Existing detectors either evaluate outputs externally or analyze internal signals through representations or predefined component functions, leaving open the question of which specific components consistently support different knowledge sources. We instead investigate which specific neurons consistently participate in supporting contextual versus parametric generation. We introduce **ReNA**, a Reference-Contrast Neuron Atlas for detecting contextual hallucinations in RAG. Given the same context and question, ReNA contrasts gradient-based neuron attribution between a context-supported reference and a conflicting parametric reference. It then constructs a sparse atlas of attention and FFN neurons exhibiting strong and consistent source preferences. An actual response is classified from its attribution pattern over the discovered atlas. Experiments across multiple models and datasets show that ReNA consistently outperforms strong output-based and internal-state baselines. Ablation studies further demonstrate the effectiveness of reference contrast and identity-specific neuron attribution. Causal interventions identify the Atlas as a distributed, branch-conditioned readout and amplifier, suggesting that faithful/parametric information propagates as a broad network state expressed across many components rather than through a small, independently sufficient circuit.

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) Shuster et al. (2021); Gao et al. (2023) grounds large language model outputs in external evidence, offering a practical means of reducing errors caused by incomplete or outdated parametric knowledge. Retrieval alone, however, does not guarantee that the model will use the retrieved evidence. When accurate context conflicts with parametric memory, the generated response may either follow the context or revert to the parametric alternative Xu et al. (2024); Niu et al. (2024). We study post-hoc response-state detection in this context–memory conflict setting: given a context–question prompt and a model-generated answer, the goal is to determine whether the answer follows the provided context or the model's parametric knowledge.

Existing hallucination detectors can be broadly categorized by whether they access the generator's internal computation. Model-external methods evaluate evidence alignment and factual consistency, or estimate hallucination risk from sampling disagreement and semantic uncertainty Zha et al. (2023); Min et al. (2023); Manakul et al. (2023); Farquhar et al. (2024). Although broadly applicable, they characterize hallucination through observable behavior and do not reveal how the model internally resolves conflicts between contextual evidence and parametric knowledge.

Model-internal methods instead exploit hidden representations, attention patterns, or FFN computations, building on evidence that knowledge-related computation can be localized to individual FFN and attention neurons Geva et al. (2021); Dai et al. (2022); Yu & Ananiadou (2024). Representation-based approaches use hidden-state probes, response-representation geometry, or attention statistics as predictive features Azaria & Mitchell (2023); Chen et al. (2024); Chuang et al. (2024). Mechanism-guided approaches derive signals from hypothesized component functions, identifying

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.

context and memory heads, context-aware FFN neurons, or components associated with context utilization and parametric-knowledge injection Jin et al. (2024); Shi et al. (2024); Sun et al. (2025).

Despite their differences, existing internal approaches primarily characterize either what information internal representations contain or what functions particular components are hypothesized to perform. Representation-based methods establish that hidden states or attention statistics are predictive of hallucination, whereas mechanism-guided methods associate selected heads or FFNs with predefined roles such as contextual retrieval, copying, or parametric recall. However, functional characterization alone does not reveal whether knowledge-source selection is associated with a stable pattern of which specific native components participate. In particular, comparatively little attention has been paid to whether fixed component identities exhibit source-selective participation patterns across examples, without first assigning those components a semantic or mechanistic role. This motivates our central question: *Can we identify specific component identities whose participation patterns consistently distinguish contextual from parametric generation, without prespecifying what functions those components perform?*

We address this question with a **Reference-Contrast Neuron Atlas (ReNA)**, an identity-centric framework that separates source-selective neuron discovery from supervised response-state classification. For each training prompt, ReNA contrasts neuron attribution between a context-supported reference and a conflicting parametric reference under the same context–question input, thereby identifying neurons that consistently favor one knowledge source over the other. These neurons form a sparse, fixed atlas discovered without using actual generated responses or response-state labels. Once the atlas is fixed, attribution patterns of actual responses over these neuron identities are used by a lightweight classifier to distinguish context-faithful from parametric generation.

ReNA thus shifts the focus from assigning predefined functions to model components toward identifying persistent component identities whose participation is source selective. Across two model sizes and four evaluation settings, ReNA consistently improves contextual-hallucination detection over strong output- and model-internal baselines. Ablation and transfer analyses show that its gains depend on same-prompt reference contrast and the preservation of neuron identity, while causal interventions further indicate that the selected neurons can modulate knowledge-source preference. Together, these results support identity-specific neuron participation as a complementary signal for understanding and detecting context–memory conflict.

2 Related Work

Hallucinations can persist in retrieval-augmented generation even when relevant evidence is available, particularly when contextual and parametric knowledge conflict Niu et al. (2024); Xu et al. (2024). Output-based detectors assess evidence alignment, factual consistency, or uncertainty, including AlignScore Zha et al. (2023), FActScore Min et al. (2023), SelfCheckGPT Manakul et al. (2023), and semantic entropy Farquhar et al. (2024). Model-internal approaches instead exploit hidden representations or attention statistics, such as truthfulness probes Azaria & Mitchell (2023), INSIDE Chen et al. (2024), and Lookback Lens Chuang et al. (2024). These methods establish useful behavioral and representation-level signals, but do not directly identify persistent native components that selectively support competing knowledge sources.

Mechanistic work more directly studies context–memory competition. PH3 identifies opposing context and memory heads through path patching Jin et al. (2024); IRCAN locates and strengthens context-aware FFN neurons using integrated-gradient attribution Shi et al. (2024); and ReDeEP combines Copying Heads and Knowledge FFNs to detect RAG hallucinations Sun et al. (2025). These methods provide interpretable, source-oriented accounts of model behavior, but generally begin from hypotheses about what particular components do.

ReNA differs in both its discovery signal and objective. Rather than assigning components predefined functions, we contrast neuron attribution to contextual and parametric target continuations under the same context–question prompt and identify fixed component identities whose source preferences are strong and consistent across examples. The resulting atlas therefore characterizes *which neurons participate selectively* in competing answer trajectories and uses these identity-specific attribution patterns for response-state detection.

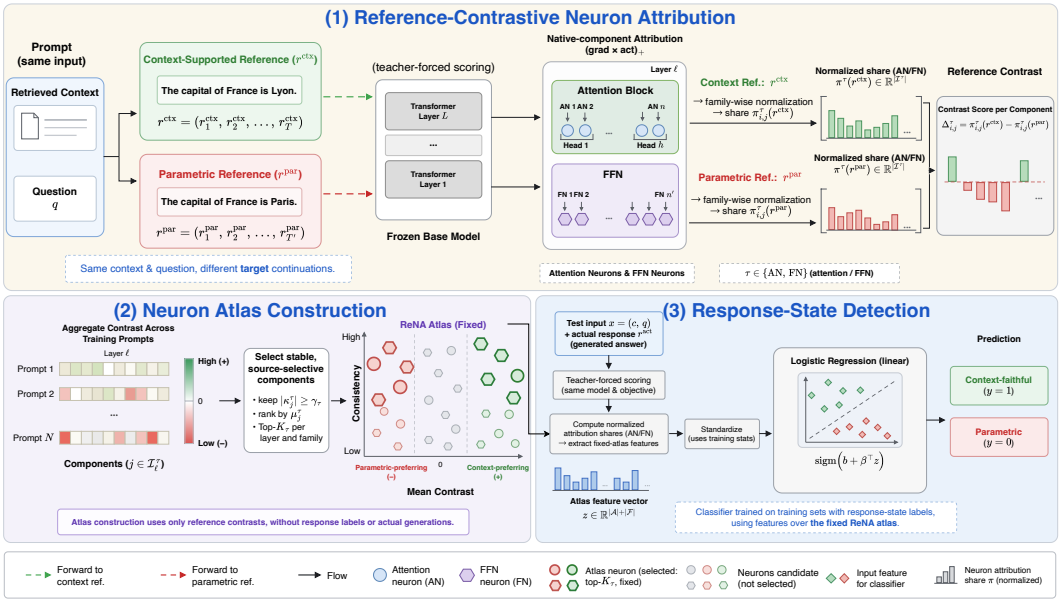


Figure 1: Overview of ReNA. Reference contrasts identify source-selective attention and FFN neurons to form a fixed atlas; attribution over this atlas then classifies actual responses as context-faithful or parametric.

3 Methodology

3.1 Problem Setup and Method Overview

We study post-hoc response-state detection under context–memory conflict. Given a prompt containing retrieved context and a question, together with an answer generated by a frozen base model, the detector predicts whether the answer follows the provided context or a conflicting answer supported by the model's parametric knowledge.

Let $x_i = (c_i, q_i)$ denote the prompt formed from context c_i and question q_i. For N training examples, the training set is

$$\mathcal{D}_{\text{tr}} = \left\{ \left(x_i, r_i^{\text{ctx}}, r_i^{\text{par}}, r_i^{\text{act}}, y_i \right) \right\}_{i=1}^{N}, \tag{1}$$

where r_i^{ctx} is supported by the provided context, r_i^{par} is a conflicting answer supported by parametric knowledge, and r_i^{act} is the answer generated by the base model. The label $y_i \in \{0, 1\}$ equals 1 for a context-faithful response and 0 for a parametric (context-conflicting) response.

Our central hypothesis is that knowledge-source selection is reflected in *which fixed neuron identities receive target-supporting attribution*. Under the same prompt, the contextual and parametric references define two conflicting answer trajectories. A neuron that repeatedly receives a larger share of positive attribution on one trajectory exhibits a stable directional preference for that reference type. If this preference is related to knowledge source, context-faithful and parametric actual responses should use the selected neurons differently. Here, *participation* refers to normalized positive gradient–activation attribution rather than raw activation magnitude.

With the attention and FFN neurons defined as the units of analysis, the method proceeds in three stages (shown in Figure 1). First, we compute reference-contrastive neuron attribution from the two teacher-forced reference continuations, aggregating attribution over their target prediction positions. Second, we construct a fixed atlas by retaining neurons whose reference preferences are directionally consistent and whose attribution contrasts are large. Third, we classify actual responses from their attribution patterns over the fixed atlas. Atlas construction uses only $(x_i, r_i^{\text{ctx}}, r_i^{\text{par}})$, whereas classifier fitting uses $(x_i, r_i^{\text{act}}, y_i)$ after the atlas has been fixed.

3.2 Attention and FFN Neuron Decomposition

Building on the FFN value-decomposition view of Geva et al. (Geva et al., 2021), neuron-level factual-knowledge attribution (Dai et al., 2022), and the extension to attention value–output coordinates by Yu and Ananiadou (Yu & Ananiadou, 2024), we define a neuron as a scalar activation multiplying a fixed output vector. This is a formal decomposition of the trained model's computation; it does not require each neuron to form an independent causal mechanism. We write $\bar{\boldsymbol{h}}_u^{\ell,\mathrm{attn}}$ and $\bar{\boldsymbol{h}}_u^{\ell,\mathrm{ffn}}$ for the actual normalized representations supplied to the attention and FFN sublayers at layer ℓ and token position u, respectively.

FFN neurons. For a two-layer FFN with residual width d, intermediate width d_{ff}, and elementwise nonlinearity ϕ, let $W_1^\ell \in \mathbb{R}^{d_{\mathrm{ff}}\times d}$ and $W_2^\ell \in \mathbb{R}^{d\times d_{\mathrm{ff}}}$. Its output is

$$\boldsymbol{F}_u^\ell = W_2^\ell \phi\big(W_1^\ell \bar{\boldsymbol{h}}_u^{\ell,\mathrm{ffn}}\big) = \sum_{k=1}^{d_{\mathrm{ff}}} a_{\ell,k}^{\mathrm{FN}}(u)\, \boldsymbol{o}_{\ell,k}^{\mathrm{FN}}, \tag{2}$$

where

$$a_{\ell,k}^{\mathrm{FN}}(u) = \phi\big((W_1^\ell)_{k,:}\bar{\boldsymbol{h}}_u^{\ell,\mathrm{ffn}}\big), \qquad \boldsymbol{o}_{\ell,k}^{\mathrm{FN}} = (W_2^\ell)_{:,k}. \tag{3}$$

Thus, the kth row of W_1^ℓ and the kth column of W_2^ℓ define FFN neuron (ℓ, k): the row determines its input-dependent activation, and the column is its fixed output vector. For a gated FFN, $a_{\ell,k}^{\mathrm{FN}}(u)$ denotes the corresponding gated intermediate activation before the output projection.

Attention neurons. We extend the same row–column decomposition to multi-head attention. Let L be the number of Transformer layers, H the number of attention heads, and d_h the per-head width. For head h, let $W_{Q,h}^\ell, W_{K,h}^\ell, W_{V,h}^\ell \in \mathbb{R}^{d_h\times d}$ and $W_{O,h}^\ell \in \mathbb{R}^{d\times d_h}$. The attention output at query position u is

$$\boldsymbol{A}_u^\ell = \sum_{h=1}^{H}\sum_{p\le u} \alpha_{h,u,p}^\ell\, W_{O,h}^\ell \left(W_{V,h}^\ell \bar{\boldsymbol{h}}_p^{\ell,\mathrm{attn}}\right). \tag{4}$$

The causal attention logit and resulting weight from query position u to source position $p \le u$ are

$$\begin{aligned} \xi_{h,u,p}^\ell &= \frac{\left(W_{Q,h}^\ell \bar{\boldsymbol{h}}_u^{\ell,\mathrm{attn}}\right)^\top \left(W_{K,h}^\ell \bar{\boldsymbol{h}}_p^{\ell,\mathrm{attn}}\right)}{\sqrt{d_h}}, \\ \alpha_{h,u,p}^\ell &= \left[\mathrm{softmax}_{p'\le u}\big(\xi_{h,u,p'}^\ell\big)\right]_p. \end{aligned} \tag{5}$$

where the softmax is taken over the source-position index p' across all causally visible positions, and $[\cdot]_p$ selects its pth component. Equation equation 5 shows the standard dot-product form. For brevity, biases and model-specific normalization, residual placement, and positional details (e.g., rotary positional embeddings) are omitted from the notation; all activations and attention weights, including those under grouped- or multi-query attention, are taken from the model's native computation.

As in the FFN case, the mth row of $W_{V,h}^\ell$ is paired with the mth column of $W_{O,h}^\ell$. Expanding Eq. equation 4 over the per-head dimension gives

$$\boldsymbol{A}_u^\ell = \sum_{h=1}^{H}\sum_{m=1}^{d_h} a_{\ell,h,m}^{\mathrm{AN}}(u)\, \boldsymbol{o}_{\ell,h,m}^{\mathrm{AN}}, \tag{6}$$

where

$$a_{\ell,h,m}^{\mathrm{AN}}(u) = \sum_{p\le u} \alpha_{h,u,p}^\ell (W_{V,h}^\ell)_{m,:}\bar{\boldsymbol{h}}_p^{\ell,\mathrm{attn}}, \qquad \boldsymbol{o}_{\ell,h,m}^{\mathrm{AN}} = (W_{O,h}^\ell)_{:,m}. \tag{7}$$

We call (ℓ, h, m) an attention neuron. Its activation aggregates the mth value-coordinate response over source positions according to the attention distribution, while its output vector is fixed by the corresponding column of $W_{O,h}^\ell$. The source position affects the activation but is not part of the neuron identity. For grouped- or multi-query attention, $W_{K,h}^\ell$ and $W_{V,h}^\ell$ denote the key and value projections assigned to query head h by the model's native query-to-KV-head mapping; multiple query heads may therefore share these matrices.

For later use, define the layer-specific identity sets

$$\begin{aligned}\mathcal{I}_\ell^{\mathrm{AN}} &= \{(\ell, h, m) : 1 \le h \le H,\ 1 \le m \le d_h\},\\ \mathcal{I}_\ell^{\mathrm{FN}} &= \{(\ell, k) : 1 \le k \le d_{\mathrm{ff}}\},\end{aligned} \tag{8}$$

and let $\mathcal{I}^\tau = \bigcup_{\ell=1}^{L} \mathcal{I}_\ell^\tau$ for $\tau \in \{\mathrm{AN}, \mathrm{FN}\}$.

3.3 Reference-Contrastive Neuron Attribution

Teacher-forced target objective. Let f denote the frozen base model and p_f its next-token distribution. For a continuation $r = (r_1, \ldots, r_T)$ of length T, we evaluate every token under teacher forcing: when scoring r_t, the model conditions on the prompt x_i and the gold prefix $r_{<t}$. The resulting sequence objective is

$$J_i(r) = \frac{1}{T} \sum_{t=1}^{T} \log p_f(r_t \mid x_i, r_{<t}) \,. \tag{9}$$

The contextual and parametric references are evaluated in separate teacher-forced passes under the same prompt x_i; the same sequence-attribution procedure is later applied to actual responses. Let $u_{i,t}$ denote the position in the concatenated sequence $x_i \| r$ whose output logits predict r_t, where $\|$ denotes sequence concatenation. Unless otherwise stated, all activations and gradients below are computed from the teacher-forced pass on (x_i, r), and this dependence is suppressed in the notation.

Neuron-level attribution. For an observed scalar neuron activation a, we use the gradient-times-activation score, a first-order Taylor attribution that approximates the decrease in the target objective caused by ablating the activation (Molchanov et al., 2017; Ancona et al., 2018). Specifically, the decrease is approximated by $a\,\partial J/\partial a$. We retain only its positive part, $[z]_+ = \max(z, 0)$, because the atlas is designed to capture neurons that locally support the target continuation rather than neurons that suppress it. These scores are local predictive attributions around the observed forward pass, not exact causal intervention effects.

For FFN neuron $j = (\ell, k)$, define

$$s_{i,j}^{\mathrm{FN}}(r) = \sum_{t=1}^{T} \left[a_{\ell,k}^{\mathrm{FN}}(u_{i,t}) \frac{\partial J_i(r)}{\partial a_{\ell,k}^{\mathrm{FN}}(u_{i,t})} \right]_+ . \tag{10}$$

For attention neuron $j = (\ell, h, m)$, we use the same neuron-level definition:

$$s_{i,j}^{\mathrm{AN}}(r) = \sum_{t=1}^{T} \left[a_{\ell,h,m}^{\mathrm{AN}}(u_{i,t}) \frac{\partial J_i(r)}{\partial a_{\ell,h,m}^{\mathrm{AN}}(u_{i,t})} \right]_+ . \tag{11}$$

Thus, both neuron families are attributed at the same granularity: one scalar activation per fixed neuron identity and target position. For attention neurons, source routing is already incorporated into $a_{\ell,h,m}^{\mathrm{AN}}(u)$ through Eq. equation 7; we do not further decompose the score by source position. Because $J_i(r)$ is sequence-level, the gradient at $u_{i,t}$ captures the activation's downstream effect on the continuation, including later target tokens. The attribution is applied at the neuron activation after attention aggregation and does not separately assign credit to the query, key, or softmax computations that produced the realized routing.

Normalized reference contrast. Raw attribution magnitudes vary across examples, continuations, and neuron families. We therefore convert each score into its share of the total positive support within the same family:

$$\pi_{i,j}^{\tau}(r) = \frac{s_{i,j}^{\tau}(r)}{\max\left(\sum_{j' \in \mathcal{I}^\tau} s_{i,j'}^{\tau}(r), \varepsilon_\pi\right)}, \qquad \tau \in \{\mathrm{AN}, \mathrm{FN}\}. \tag{12}$$

where $\varepsilon_\pi > 0$ is a numerical-stability constant. This normalization controls for example- and family-level scale variation, making attribution shares comparable across examples and families while preserving how positive attribution is distributed across neuron identities.

For each training example, the same-prompt reference contrast is

$$\Delta_{i,j}^{\tau} = \pi_{i,j}^{\tau}(r_i^{\mathrm{ctx}}) - \pi_{i,j}^{\tau}(r_i^{\mathrm{par}}). \tag{13}$$

Thus, $|\Delta_{i,j}^{\tau}|$ measures how differently neuron j supports the two reference trajectories on example i, while its sign identifies the preferred trajectory. A positive value indicates a larger normalized positive-support share on the contextual reference, and a negative value indicates a larger share on the parametric reference. Because the two references are teacher-forced under the same context and question, the contrast controls for input-side prompt variation. It remains a target-conditioned diagnostic contrast, rather than a causal isolation of knowledge provenance, because the target tokens and gold prefixes differ.

3.4 Direction-Consistent Neuron Atlas Construction

Directional consistency and contrast magnitude. The atlas is designed to identify neurons that satisfy two distinct criteria: the direction of their contextual–parametric preference is consistent across training examples, and the attribution contrast between the two reference trajectories is large. We quantify these properties separately. For neuron j in family τ, define its mean contrast magnitude

$$\mu_j^{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left| \Delta_{i,j}^{\tau} \right|, \tag{14}$$

and its signed directional consistency

$$\kappa_j^{\tau} = \frac{1}{N} \sum_{i=1}^{N} \operatorname{sgn}\left(\Delta_{i,j}^{\tau}\right), \qquad \operatorname{sgn}(0) = 0. \tag{15}$$

The magnitude $|\kappa_j^{\tau}| \in [0, 1]$ measures how consistently the neuron prefers the same reference type across examples, while its sign indicates the dominant direction: positive for contextual references and negative for parametric references. In contrast, μ_j^{τ} measures the average size of the normalized attribution difference between the two trajectories. It does not measure raw attribution on either trajectory alone.

Atlas construction. We first retain neurons whose directional consistency is at least the family-specific threshold $\gamma_{\tau} \in [0, 1]$; the endpoint $\gamma_{\tau} = 0$ disables this filter:

$$\mathcal{C}_{\ell}^{\tau} = \left\{ j \in \mathcal{I}_{\ell}^{\tau} : |\kappa_j^{\tau}| \geq \gamma_{\tau} \right\}. \tag{16}$$

Among these direction-consistent candidates, we select in each layer the K_{τ} neurons with the largest mean contrast magnitude:

$$\mathcal{G}_{\ell}^{\tau} = \mathrm{TopK}_{j \in \mathcal{C}_{\ell}^{\tau}} \left(\mu_j^{\tau}; K_{\tau} \right), \qquad \mathcal{G}^{\tau} = \bigcup_{\ell=1}^{L} \mathcal{G}_{\ell}^{\tau}. \tag{17}$$

If fewer than K_{τ} neurons in a layer satisfy the consistency threshold, all eligible neurons are retained. This two-stage rule treats consistency as a necessary condition and uses contrast magnitude only to rank neurons that pass that condition. Both context-preferring and parametric-preferring neurons may enter the atlas. We denote the resulting attention and FFN atlases by $\mathcal{A} = \mathcal{G}^{\mathrm{AN}}$ and $\mathcal{F} = \mathcal{G}^{\mathrm{FN}}$. Atlas discovery uses only the two teacher-forced reference continuations. It does not use actual responses or response labels.

3.5 Response-State Detection with the Fixed Atlas

After fixing a canonical ordering within $\mathcal{A}$ and $\mathcal{F}$, the feature vector for actual response r_i^{act} is

$$\boldsymbol{z}_i = \left(\pi_{i,j}^{\mathrm{AN}}(r_i^{\mathrm{act}}) \right)_{j \in \mathcal{A}} \oplus \left(\pi_{i,j}^{\mathrm{FN}}(r_i^{\mathrm{act}}) \right)_{j \in \mathcal{F}}, \tag{18}$$

where $\oplus$ denotes concatenation. The actual response is evaluated with the same teacher-forced sequence-attribution procedure used for the references. Before extracting the atlas coordinates, we compute each $\pi_{i,j}^{\tau}(r_i^{\mathrm{act}})$ over the full family $\mathcal{I}^{\tau}$ using the denominator in Eq. equation 12; we then retain only the coordinates indexed by $\mathcal{A}$ or $\mathcal{F}$. Thus, $\boldsymbol{z}_i \in \mathbb{R}^{|\mathcal{A}|+|\mathcal{F}|}$.

Within each training fold, after the atlas has been fixed, we fit a feature scaler, a class-weighted ℓ_2-regularized logistic-regression classifier, and a decision threshold using only the actual-response feature vectors and labels from that fold. The threshold is selected from training predictions to maximize balanced accuracy. For a held-out example or at inference time, the detector receives only the prompt and actual response, computes the same full-family shares, extracts the fixed atlas coordinates, and applies the fitted scaler, classifier, and threshold. It never accesses reference continuations, response labels, or a case-specific atlas selection.

4 Experimental Setup

We evaluate on the QA, MR, and MC subsets of ConFiQA Bi et al. (2024), and additionally construct 6,000 model-specific conflict cases from the answerable subset of SQuAD 2.0 Rajpurkar et al. (2018), using its gold answer as the contextual reference and the model's closed-book answer as the parametric reference (Appendix C.2). Experiments use frozen LLaMA-2-7B-Chat and LLaMA-2-13B-Chat models Touvron et al. (2023). We compare ReNA with seven representative detectors: output-based **P(True)** Kadavath et al. (2022), **RefChecker** Hu et al. (2024), and **SelfCheckGPT** Manakul et al. (2023); representation-based **EigenScore** Chen et al. (2024) and **SEP** Kossen et al. (2024); attention-based **FrequencyAware** Qi et al. (2026); and mechanism-guided **ReDeEP** Sun et al. (2025). Their operational details are summarized in Appendix C. For each model–dataset configuration, we use repeated stratified five-fold cross-validation with five repeats. Atlas discovery, feature scaling, classifier fitting, and threshold selection are performed using training folds only; held-out examples are evaluated with the resulting fixed atlas and detector, and their out-of-fold predictions are averaged across repeats. All trainable methods use identical data partitions and saved model responses. We report AUROC as the primary metric together with Macro-F1 (denoted F). Full implementation, hyperparameter, evaluation, and prompt details are provided in Appendix C.

5 Results and Analysis

5.1 Main Results

Table 1 compares ReNA with representative output-level, representation-based, attention-statistic, and mechanism-guided hallucination detectors across four benchmarks. We report AUROC and Macro-F1 for LLaMA-2-7B-Chat and LLaMA-2-13B-Chat, and the unweighted average of each metric across datasets.

As shown in Table 1, ReNA achieves the best average performance on both model sizes. On LLaMA-2-7B-Chat, ReNA reaches 0.9362 AUROC and 0.9180 Macro-F1, outperforming the strongest baselines by 1.24 and 1.80 percentage points, respectively. On LLaMA-2-13B-Chat, it obtains 0.9268 AUROC and 0.8917 Macro-F1, improving over SEP by 0.88 and 1.73 points. ReNA ranks first on 15 of the 16 dataset-level metrics and achieves the highest Macro-F1 on all eight model–dataset pairs. The only exception is AUROC on MC with LLaMA-2-13B-Chat, where SEP is slightly higher. SEP and FrequencyAware are the strongest competing methods, but ReNA provides a more consistent balance between AUROC and Macro-F1 across datasets. ReDeEP performs substantially worse, suggesting that predefined mechanistic proxies may miss other neuron identities that are diagnostic of knowledge-source selection. Output-level methods, including P(True), RefChecker, and SelfCheckGPT, are generally less stable, particularly in AUROC. Overall, the results indicate that identity-specific neuron attribution patterns provide a more robust signal for distinguishing contextual from parametric responses.

5.2 Ablation Study

We conduct ablation experiments to examine the contributions of the atlas selection criteria, neuron families, and attribution normalization. We also compare ReNA with a matched random atlas to determine whether its performance arises from the identities selected by reference contrast, rather than merely from the atlas size or layer distribution. The experimental results are presented in Table 2.

Table 1: Main detection results on LLaMA-2-7B-Chat and LLaMA-2-13B-Chat. F denotes Macro-F1; "Avg." denotes the unweighted average of each metric across the four datasets. The best result in each model block is shown in **bold**, and the second-best result is underlined.

	ConFiQA-QA		ConFiQA-MR		ConFiQA-MC		SQuAD2		Average	
Method	AUROC	F	AUROC	F	AUROC	F	AUROC	F	AUROC	F
LLaMA-2-7B-Chat										
P(True)	0.5221	0.7171	0.5453	0.6949	0.5496	0.8068	0.6629	0.6800	0.5700	0.7247
RefChecker	0.8276	0.7708	0.6517	0.7740	0.6570	0.7729	0.5660	0.6617	0.6756	0.7449
SelfCheckGPT	0.5916	0.6782	0.5397	0.7062	0.4586	0.4986	0.6260	0.6558	0.5540	0.6347
EigenScore	0.5890	0.5254	0.6522	0.6059	0.6340	0.6321	0.6239	0.6942	0.6248	0.6144
SEP	0.9337	0.9169	0.9386	0.9364	0.9408	0.9308	0.8185	0.8158	0.9079	0.9000
FrequencyAware	0.9670	0.8563	0.9384	0.7927	0.9432	0.7581	0.8467	0.7661	0.9238	0.7933
ReDeEP	0.7455	0.6433	0.7313	0.6847	0.6792	0.7528	0.6271	0.5505	0.6958	0.6578
ReNA (Ours)	**0.9732**	**0.9508**	**0.9600**	**0.9407**	**0.9582**	**0.9523**	**0.8534**	**0.8281**	**0.9362**	**0.9180**
LLaMA-2-13B-Chat										
P(True)	0.5463	0.6426	0.5569	0.6194	0.5573	0.7109	0.6488	0.6817	0.5773	0.6637
RefChecker	0.7889	0.7113	0.6955	0.6645	0.4944	0.6018	0.5245	0.6583	0.6258	0.6590
SelfCheckGPT	0.5589	0.5500	0.5195	0.5404	0.4795	0.5058	0.6366	0.6792	0.5486	0.5689
EigenScore	0.5781	0.5658	0.5028	0.4890	0.5022	0.4826	0.5748	0.6358	0.5395	0.5433
SEP	0.9325	0.8855	0.9479	0.8968	**0.9581**	0.9127	0.8337	0.8025	0.9180	0.8744
FrequencyAware	0.9441	0.8841	0.9353	0.8556	0.9443	0.8426	0.8362	0.7649	0.9150	0.8368
ReDeEP	0.6614	0.6677	0.6766	0.6382	0.6955	0.7306	0.6436	0.5938	0.6693	0.6576
ReNA (Ours)	**0.9580**	**0.9186**	**0.9528**	**0.9087**	0.9505	**0.9159**	**0.8460**	**0.8235**	**0.9268**	**0.8917**

Table 2: Ablation results on the ConFiQA macro-average and SQuAD2 (AUROC / Macro-F1). Matched Random Atlas reports mean ± standard deviation over five draws with per-layer attention/FFN counts matched to ReNA.

Variant	**LLaMA-2-7B-Chat**		**LLaMA-2-13B-Chat**	
	ConFiQA Avg. *AUROC / F*	SQuAD2 *AUROC / F*	ConFiQA Avg. *AUROC / F*	SQuAD2 *AUROC / F*
Matched Random Atlas	0.8876/0.8519 ±0.0124/ ± 0.0141	0.8068/0.7846 ±0.0187/ ± 0.0162	0.8943/0.8571 ±0.0136/ ± 0.0155	0.7985/0.7764 ±0.0203/ ± 0.0189
Magnitude Only	0.9507/0.9326	0.8321/0.8067	0.9415/0.8982	0.8267/0.8036
Consistency Only	0.9549/0.9368	0.8392/0.8135	0.9458/0.9029	0.8328/0.8104
Attention Neurons Only	0.9452/0.9258	0.8338/0.8074	0.9399/0.8956	0.8295/0.8068
FFN Neurons Only	0.9386/0.9173	0.8197/0.7961	0.9312/0.8861	0.8164/0.7937
Without Family Normalization	0.9274/0.9046	0.8035/0.7808	0.9186/0.8748	0.7952/0.7719
ReNA	**0.9638/0.9479**	**0.8534/0.8281**	**0.9538/0.9144**	**0.8460/0.8235**

Ablation analysis. Using either contrast magnitude or directional consistency alone degrades performance, indicating that the two criteria are complementary: consistency filters neurons with unstable reference preferences, whereas magnitude prioritizes neurons with stronger contextual–parametric contrasts. Restricting the atlas to only attention or FFN neurons also reduces performance, suggesting that the two neuron families capture complementary aspects of response-state variation. The degradation without family-wise normalization further shows the importance of controlling attribution-scale differences.

The matched random atlas remains clearly predictive despite preserving only randomly selected neurons from the same layers and neuron families. This suggests that response state is not confined to a unique sparse set of specialized "hallucination neurons," but is reflected in a broadly distributed and partially redundant attribution signature. A sufficiently large random subset can therefore retain part of this signal. Nevertheless, ReNA consistently outperforms random atlases of identical size and layer-wise composition. We interpret the learned atlas as a compact, signal-enriched readout of the distributed attribution field: reference contrast selects coordinates on which the response-state signature is expressed more consistently and with higher contrast, thereby improving signal-to-noise ratio over arbitrary neuron subsets.

5.3 Sensitivity to Atlas Size

We study the sensitivity of ReNA to atlas size while fixing the attention-to-FFN neuron ratio at 2:1. Performance generally improves with atlas size and then saturates. For LLaMA-2-7B-Chat, $128/64$ achieves the best ConFiQA Macro-F1, while $256/128$ performs best on SQuAD2. For LLaMA-2-13B-Chat, $512/256$ achieves the best overall performance, and further increasing the atlas size yields no additional gains. Detailed results are provided in Appendix A and Table A.1.

5.4 Cross-Domain Transfer Analysis

We test whether neuron identities discovered on one ConFiQA domain transfer to another by freezing the source-domain atlas and training only the response-state classifier on the target domain; target-domain references are not used for atlas discovery. Across the six cross-domain transfers for each model, performance remains close to the corresponding in-domain setting: for LLaMA-2-7B-Chat, average AUROC and Macro-F1 decrease by 0.48 and 0.44 percentage points, respectively, with a similarly small gap for LLaMA-2-13B-Chat. This indicates that the selected neuron identities capture participation patterns shared across QA, multi-hop reasoning, and multi-conflict settings. Full transfer matrices are reported in Appendix C.1.

5.5 Causal Characterization of the Attribution Atlas

We examine whether Atlas coordinates are causally involved in the model's faithful-versus-parametric answer preference or merely correlate with a decision formed elsewhere. Using held-out examples and layer-, module-, and cardinality-matched random controls, we scale the selected AN/FN coordinates along their sign-stable directions. At the largest evaluated budget, the random-adjusted change in the contextual-versus-parametric target-logit margin ranges from 2.85 to 5.91 for faithful-support coordinates and from 0.22 to 1.50 for parametric-support coordinates across ConFiQA-QA, ConFiQA-MR, ConFiQA-MC, and SQuAD2, with all confidence intervals excluding zero. Within the evaluated budgets, the effect also becomes stronger as more Atlas coordinates are included. These results show that Atlas coordinates are not purely correlational: intervening on them can causally modulate the model's knowledge-source preference.

We next use counterfactual activation patching to test whether the Atlas itself is sufficient to carry this preference. Here, activations are transferred between valid model runs under the original and counterfactual prompts rather than being synthetically perturbed. Although replacing Atlas activations can disrupt the current answer preference in some settings, transplanting the original Atlas activations into an answer-swapped counterfactual run does not restore the original preference. Relative to layer-, module-, and cardinality-matched random restoration, the sufficiency effect is negative in all four datasets, ranging from -0.146 to -0.062.

Together, these results suggest that Atlas neurons are causally relevant, state-dependent supporters of the model's current answer branch. They can modulate and sometimes help maintain an existing knowledge-source preference, but the underlying decision is encoded in a broader network state that cannot be reconstructed from the Atlas alone. We therefore interpret the Atlas as a distributed, branch-conditioned causal readout and amplifier rather than a self-contained source-selection circuit. Full intervention protocols and results are provided in Appendix B; the directional-scaling budget sweep and cross-dataset intervention results are reported in Tables B.1 and B.2, respectively.

6 Conclusion

ReNA is a reference-contrast neuron atlas for detecting contextual hallucinations in RAG. By contrasting contextual and parametric reference trajectories, it identifies attention and FFN neurons with stable source preferences and classifies generated responses from attribution over their fixed identities. Across ConFiQA and SQuAD2 on two LLaMA-2 sizes, ReNA outperforms strong output- and representation-based baselines, while ablation and transfer analyses support the roles of reference contrast, neuron identity, and complementary neuron families. Causal interventions show that Atlas coordinates can modulate knowledge-source preference, whereas counterfactual patching indicates that they are not independently sufficient to reconstruct the decision state. These results support neuron-level participation as a compact signal of knowledge-source selection.

AI USE STATEMENT

In preparing this manuscript, the authors used generative AI tools, including ChatGPT, to assist with manuscript drafting and language editing, organization and restructuring, LaTeX formatting, literature and reference checks, preparation and revision of a scientific figure, and feedback on the presentation and interpretation of analyses based on existing experimental results. All numerical results reported in this paper were obtained from the authors' experiments and were not generated by AI tools. All AI-assisted text, citations, figures, and analytical suggestions were independently checked and revised by the authors against the underlying experimental outputs and source materials. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript and all associated artifacts.

ETHICS STATEMENT

This work studies the analysis and detection of context–memory conflicts in large language models and does not involve human subjects or the collection of private or sensitive personal data. Our experiments use publicly available models and benchmark datasets under their respective terms of use. ReNA is intended as a research tool for analyzing and detecting contextual hallucinations rather than as a guarantee of factual reliability; like other hallucination detectors, it may produce false positives or false negatives and should not replace external verification in high-stakes applications. The causal interventions in this work are used solely as controlled analyses of frozen language models. We are not aware of additional ethical concerns specific to the proposed methodology beyond those generally associated with large language model research.

REPRODUCIBILITY STATEMENT

The methodology required to reproduce ReNA, including the neuron decomposition, reference-contrast attribution, atlas construction, and response-state classifier, is described in the main text. The experimental protocol is specified in Section 4, with additional implementation details provided in Appendix C, including dataset construction, atlas hyperparameters, attribution and inference settings, classifier configuration, evaluation procedure, and prompt format. The protocols, statistical procedures, and implementation details for the causal analyses are documented in Appendix B. To further facilitate reproducibility, we will release the source code for the experimental pipeline upon acceptance.

REFERENCES

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30-May 3, 2018, Conference Track Proceedings*. OpenReview, 2018.

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

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, et al. Context-dpo: Aligning language models for context-faithfulness, 2024.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. In *EMNLP*, pp. 1419–1436, 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, 2022.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2023.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6953–6975, 2024.

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, 2024.

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, 2022.

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

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

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

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *International conference on learning representations*, 2017.

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

Siya Qi, Yudong Chen, Runcong Zhao, Qinglin Zhu, Zhanghao Hu, Wei Liu, Yulan He, Zheng Yuan, and Lin Gui. Detecting contextual hallucinations in large language models with frequency-aware attention. In *Forty-third International Conference on Machine Learning*, 2026.

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, 2018.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *Advances in Neural Information Processing Systems*, 37:4997–5024, 2024.

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

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *International Conference on Learning Representations*, volume 2025, pp. 50250–50279, 2025.

Table A.1: Sensitivity to the per-layer atlas size. "AN/FN" denotes the numbers of attention and FFN neurons selected from each layer. ConFiQA reports the macro-average over QA, MR, and MC. F denotes Macro-F1. Bold values indicate the best result in each model–dataset block, and † marks the configuration used in the main experiments.

Model	AN/FN	ConFiQA Avg.		SQuAD2	
		AUROC	F	AUROC	F
LLaMA-2-7B-Chat	$8/4$	0.9454	0.9282	0.7590	0.7127
	$16/8$	0.9466	0.9303	0.7552	0.7113
	$32/16$	0.9543	0.9385	0.7297	0.7717
	$64/32^{\dagger}$	0.9638	0.9479	0.7562	0.7607
	$128/64$	0.9714	**0.9522**	0.8451	0.8150
	$256/128^{\dagger}$	0.9697	0.9409	**0.8534**	**0.8281**
	$512/256$	**0.9729**	0.9228	0.8518	0.8244
LLaMA-2-13B-Chat	$16/8$	0.8977	0.8548	0.7582	0.7197
	$32/16$	0.9113	0.8718	0.7280	0.7228
	$64/32$	0.9275	0.8838	0.7427	0.7317
	$128/64$	0.9441	0.8939	0.7802	0.7582
	$256/128$	0.9526	0.8982	0.8236	0.7815
	$512/256^{\dagger}$	**0.9538**	**0.9144**	**0.8460**	**0.8235**
	$1024/512$	0.9512	0.9038	0.8421	0.8189

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, 2023.

Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey, 2024.

Zeping Yu and Sophia Ananiadou. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3267–3280, 2024.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, 2023.

A Sensitivity to Atlas Size

We analyze the sensitivity of ReNA to the number of neurons retained in the atlas. We fix the ratio of attention neurons to FFN neurons to 2:1, allocate the same number of neurons to every Transformer layer, and vary the per-layer atlas size. Under our neuron definition, each layer of LLaMA-2-7B-Chat contains $4{,}096$ attention neurons and $11{,}008$ FFN neurons, whereas each layer of LLaMA-2-13B-Chat contains $5{,}120$ attention neurons and $13{,}824$ FFN neurons. Because the same budget is used in every layer, the per-layer retention ratios are also the retention ratios over the whole model. The evaluated configurations therefore retain 0.20%–12.50% of the attention neurons and 0.04%–2.33% of the FFN neurons in the 7B model, and 0.31%–20.00% and 0.06%–3.70%, respectively, in the 13B model. For ConFiQA, we report the macro-average over its QA, MR, and MC subsets.

Analysis. Performance generally improves as the atlas grows (shown in Table A.1), indicating that very small atlases do not capture sufficient response-state information. However, the optimal size differs across metrics. On the 7B model, ConFiQA Macro-F1 peaks at $128/64$, while larger atlases provide the strongest AUROC and SQuAD2 results. On the 13B model, ConFiQA Macro-F1 peaks at $256/128$, whereas $512/256$ achieves the best AUROC and SQuAD2 performance. Increasing the atlas further to $1024/512$ yields no additional improvement, suggesting that excessively large atlases may introduce less informative neurons. Overall, the results favor moderate-to-large atlases, with the larger model requiring approximately twice as many neurons per layer to reach a comparable operating range.

B CAUSAL ANALYSIS OF THE RENA ATLAS

ReNA detects whether an actual response follows the retrieved context from its attribution shares on the fixed attention-neuron and FFN-neuron atlases, $\mathcal{A}$ and $\mathcal{F}$. Detection accuracy alone, however, does not establish what causal role these neurons play. A neuron may be predictive because it changes the answer preference when intervened on, because it transports the knowledge-source decision through the model, or simply because it reflects a decision formed elsewhere. We therefore conduct a separate intervention analysis on LLaMA-2-7B-Chat.

We use two complementary interventions to characterize this causal role. Directional scaling tests whether Atlas-selected neurons can amplify contextual or parametric answer preferences, whereas counterfactual activation patching tests whether their activation states independently carry enough information to reconstruct those preferences. Accordingly, the analysis asks two progressively stronger questions:

1. **Causal modulation:** Does intervening on reference-contrast-selected neurons change the model's preference between the contextual and parametric answers?
2. **Causal carriage:** Are the activations of these neurons sufficient to transfer that preference between an original run and a counterfactual run?

The two experiments are therefore complementary: together, they distinguish whether the selected neurons act as state-dependent causal amplifiers or as self-contained causal carriers.

B.1 ATLAS-DERIVED DIRECTION-CONTROLLED INTERVENTION SETS

Relation to the ReNA Atlas. We retain the notation and neuron definitions from the main text. For $\tau \in \{\mathrm{AN}, \mathrm{FN}\}$, $\pi^{\tau}_{i,j}(r)$ is the normalized positive attribution share of neuron j for continuation r, and

$$\Delta^{\tau}_{i,j} = \pi^{\tau}_{i,j}(r^{\mathrm{ctx}}_i) - \pi^{\tau}_{i,j}(r^{\mathrm{par}}_i) \tag{19}$$

is the same-prompt reference contrast used by ReNA. Positive values indicate stronger support for the contextual reference; negative values indicate stronger support for the parametric reference.

The detector Atlas $\mathcal{A} \cup \mathcal{F}$ is optimized for compact prediction by applying the consistency threshold γ^{τ} and ranking eligible neurons by the mean contrast magnitude μ^{τ}_j. For controlled bidirectional interventions, we derive direction-balanced subsets from the same reference-contrast statistics:

$$\mathcal{S}^{\tau}_{\mathrm{ctx}}(B_\tau) \quad \text{and} \quad \mathcal{S}^{\tau}_{\mathrm{par}}(B_\tau), \tag{20}$$

containing B_τ context-preferring and B_τ parametric-preferring neurons per layer, respectively. The balancing and reliability ranking make it possible to compare the two directions at matched budgets and to sweep the number of intervened neurons. This preserves ReNA's neuron definitions, reference contrast, source-preference directions, and training references; the intervention analysis neither changes the detector nor introduces a new test-time feature representation.

Sign-stability ranking. The intervention sets are discovered using only official-training reference contrasts. Actual responses and response-state labels are never read during selection. For neuron j, let n^{+}_j and n^{-}_j be the numbers of non-zero training contrasts with positive and negative signs. We require $n^{+}_j + n^{-}_j \geq 20$ and define non-zero directional stability as

$$\rho^{\tau}_j = \frac{\max(n^{+}_j, n^{-}_j)}{n^{+}_j + n^{-}_j}. \tag{21}$$

Within every layer, neuron family, and dominant sign, neurons are ranked by the 95% Wilson lower confidence bound of ρ^{τ}_j, followed by ρ^{τ}_j, non-zero support, and mean contrast magnitude. This is a finite-sample reliability refinement of the directional-consistency criterion κ^{τ}_j used in the main method. We use "context-preferring" and "parametric-preferring" throughout this appendix because the two directions are defined by reference contrast, not by the label of the generated response.

Each dataset has its own intervention sets, discovered from its official training split: 3,806 QA, 2,831 MR, 2,482 MC, and 4,800 SQuAD2 examples. The largest sets contain 64 ANs and 128 FNs per layer and direction. Matched-random sets exclude all neurons in the largest directional ranking and preserve the layer, neuron family, number of neurons, and assigned intervention sign.

B.2 WHAT IS MEASURED?

The contextual and parametric reference answers may begin with the same tokens. For each example, we teacher-force their longest shared token prefix and locate the first position at which the two references differ. Let v_i^{ctx} and v_i^{par} be the two candidate tokens at this position. The outcome of every intervention is the knowledge-source preference margin

$$m_i = z_i(v_i^{\mathrm{ctx}}) - z_i(v_i^{\mathrm{par}}), \tag{22}$$

where $z_i(v)$ is the model logit for token v. Increasing m_i favors the contextual continuation, while decreasing it favors the parametric continuation. This controlled next-token margin isolates the point at which the two answer trajectories first make different predictions. It should not be interpreted as a direct measure of full-response correctness.

We first use a held-out QA screening set containing 20 context-faithful and 20 parametric responses to compare intervention budgets. We then freeze the largest tested budget and evaluate it on a disjoint QA cohort of 100+100 examples, MR 80+80, MC 49+49, and SQuAD2 100+100. Two SQuAD2 examples have no distinct next reference token after tokenization and are explicitly excluded, leaving 198 completed examples. No other experiment has skipped examples. Response labels are used only to balance these evaluation cohorts and for post-hoc subgroup analysis.

B.3 EXPERIMENT 1: CAN ATLAS NEURONS MODULATE ANSWER PREFERENCE?

Intervention and effect metric. At the prediction position in every layer, we modify the AN activation before the attention output projection and the FN activation before the FFN output projection. Let $\sigma_j = +1$ for a context-preferring neuron and $\sigma_j = -1$ for a parametric-preferring neuron. With dose $\lambda = 0.5$, a context-directed intervention uses

$$a_j^{\rightarrow \mathrm{ctx}} = a_j(1 + \lambda\sigma_j), \tag{23}$$

whereas a parametric-directed intervention reverses the sign:

$$a_j^{\rightarrow \mathrm{par}} = a_j(1 - \lambda\sigma_j). \tag{24}$$

Thus, the two runs differ only in whether the selected neurons are scaled toward their learned reference-preference direction or away from it.

For intervention set $\mathcal{S}$, the per-example directional scaling effect is

$$e_i^{\mathrm{scale}}(\mathcal{S}) = m_i^{\rightarrow \mathrm{ctx}}(\mathcal{S}) - m_i^{\rightarrow \mathrm{par}}(\mathcal{S}). \tag{25}$$

To control for the possibility that changing any large set of neurons affects the logits, the reported quantity is the paired random-adjusted effect

$$\delta_i^{\mathrm{scale}}(\mathcal{S}) = e_i^{\mathrm{scale}}(\mathcal{S}) - \frac{1}{R}\sum_{r=1}^{R} e_i^{\mathrm{scale}}(\mathcal{S}_r^{\mathrm{rand}}). \tag{26}$$

A positive value means that the reference-contrast-selected neurons control the contextual-versus-parametric margin more strongly in the intended direction than matched random neurons. The QA replication uses $R = 20$ random sets, while the screening and cross-domain runs use $R = 10$.

Budget sweep. Table B.1 reports the QA screening experiment. In the tested range, increasing the number of selected neurons strengthens the random-adjusted effect in both directions. We do not extrapolate this trend beyond the evaluated budgets or claim that no larger-budget saturation exists. Rather, the result shows that the observed effect is not confined to only the smallest selected subset.

AN/FN composition. At a fixed total of 48 neurons per layer, context-preferring effects are similar for AN-only, FN-only, and mixed allocations (1.482–1.670). Parametric-preferring effects are smaller for FN-only (0.240) than for the mixed 24/24 allocation (0.399) or the 16/32 allocation (0.424). The first direction is therefore not attributable to only one neuron family, while the second shows modest complementarity between ANs and FNs.

Table B.1: QA directional-scaling budget sweep. Effects are the mean random-adjusted changes in Eq. 26; brackets denote 95% paired-bootstrap confidence intervals. All rows use the same 40 held-out examples.

AN/layer	FN/layer	Total	Direction	Effect	95% CI
4	8	384	Context-preferring	0.450	[0.215, 0.687]
8	16	768	Context-preferring	0.846	[0.338, 1.372]
16	32	1,536	Context-preferring	1.530	[0.603, 2.548]
32	64	3,072	Context-preferring	2.756	[1.205, 4.363]
64	128	6,144	Context-preferring	5.449	[3.001, 8.013]
4	8	384	Parametric-preferring	0.121	[0.053, 0.195]
8	16	768	Parametric-preferring	0.197	[0.090, 0.312]
16	32	1,536	Parametric-preferring	0.424	[0.206, 0.651]
32	64	3,072	Parametric-preferring	0.697	[0.313, 1.112]
64	128	6,144	Parametric-preferring	1.098	[0.521, 1.742]

Independent replication and cross-domain results. After the QA screen, we freeze the largest evaluated intervention set (64 ANs and 128 FNs per layer and direction). Table B.2 (a) shows the results on the disjoint QA cohort and three additional datasets. The random-adjusted effect is positive in all eight dataset–direction conditions, and all confidence intervals exclude zero. Context-preferring effects range from 2.851 to 5.912, while parametric-preferring effects range from 0.221 to 1.497.

Interpretation. Directional scaling establishes *causal modulation and amplification*: deliberately changing the selected neuron activations provides reliable directional control over the answer-preference margin, with substantially larger effects than matched random neurons in all eight dataset–direction conditions. These results show that the selected neurons do not merely record the response state; their activations can amplify contextual or parametric answer branches. Scaling alone does not identify them as the exclusive natural mediation pathway, so we test independent causal carriage separately.

B.4 Experiment 2: Do Atlas neurons carry the decision state?

Counterfactual runs. We construct two counterfactual model runs and record their native AN/FN activations. The primary *answer-swapped* counterfactual replaces the contextual answer string in the retrieved Context with the conflicting parametric answer, while leaving the system prompt, question, and remaining Context unchanged. This creates a same-question counterfactual in which the answer-bearing evidence changes but the prompt format and most semantics are preserved. As an auxiliary stress test, the *no-Context* counterfactual retains the system prompt and question but removes the complete Context block. For receiver activation a_j^r, donor activation a_j^d, and interpolation coefficient α, we patch

$$\widetilde{a}_j^r = a_j^r + \alpha(a_j^d - a_j^r). \tag{27}$$

The patch is applied to selected AN/FN activations at the prediction position in all layers. QA uses $\alpha \in \{0.25, 0.5, 1.0\}$ to verify dose behavior; cross-domain comparisons use complete replacement ($\alpha = 1$).

Necessity and sufficiency metrics. The necessity test uses the original run as receiver and the counterfactual run as donor:

$$e_i^{\mathrm{nec}}(\mathcal{S}) = m_i^{\mathrm{orig}} - m_i^{\mathrm{orig}\leftarrow\mathrm{cf}}(\mathcal{S}). \tag{28}$$

A positive value means that replacing $\mathcal{S}$ with its counterfactual state removes part of the original contextual margin. The sufficiency test reverses the transfer:

$$e_i^{\mathrm{suf}}(\mathcal{S}) = m_i^{\mathrm{cf}\leftarrow\mathrm{orig}}(\mathcal{S}) - m_i^{\mathrm{cf}}. \tag{29}$$

A positive value means that restoring the original activations recovers part of the contextual margin in the counterfactual run. We test the context-preferring set, the parametric-preferring set, and their

Table B.2: Cross-dataset causal intervention results. "Effect" denotes the selected-minus-matched-random change in the contextual-minus-parametric margin; brackets denote 95% paired-bootstrap confidence intervals.

(a) Directional scaling: 64 ANs and 128 FNs per layer and direction

Dataset	Direction	n	Effect	95% CI
QA	Context-preferring	200	5.912	[4.743, 7.116]
	Parametric-preferring	200	1.497	[1.191, 1.797]
MR	Context-preferring	160	2.851	[2.092, 3.573]
	Parametric-preferring	160	0.233	[0.121, 0.349]
MC	Context-preferring	98	3.298	[2.029, 4.512]
	Parametric-preferring	98	0.221	[0.099, 0.343]
SQuAD2	Context-preferring	198	2.919	[1.923, 3.921]
	Parametric-preferring	198	1.390	[1.034, 1.771]

(b) Answer-swapped patching: union of both directional sets

Dataset	Test	n	Selected	Random	Effect	95% CI
QA	Necessity	200	0.019	0.205	-0.186	$[-0.229, -0.149]$
MR	Necessity	160	0.070	0.123	-0.053	$[-0.093, 0.004]$
MC	Necessity	98	0.024	0.090	-0.065	$[-0.087, -0.046]$
SQuAD2	Necessity	198	0.057	0.098	-0.041	$[-0.085, 0.019]$
QA	Sufficiency	200	0.043	0.189	-0.146	$[-0.183, -0.108]$
MR	Sufficiency	160	0.036	0.114	-0.078	$[-0.101, -0.057]$
MC	Sufficiency	98	0.023	0.086	-0.062	$[-0.078, -0.048]$
SQuAD2	Sufficiency	198	0.016	0.095	-0.079	$[-0.109, -0.049]$

union. As in Experiment 1, the reported effect subtracts the per-example mean of matched random sets. A transferable causal carrier should outperform random controls, particularly on the sufficiency test.

Answer-swapped results. Table B.2 (b) reports the union intervention for the primary answer-swapped counterfactual. Patching produces small positive raw restoration effects, showing that the transferred activations are not inert. However, the selected neurons do not preferentially restore the original preference beyond matched random neurons. Random-adjusted sufficiency is negative in all four datasets, from -0.146 to -0.062, because matched random neurons of the same layerwise composition produce a larger shift. Thus, the selected activations provide limited restoration but do not constitute an Atlas-specific, independently sufficient carrier of the knowledge-source preference.

No-Context stress test. With the no-Context donor, union necessity is positive relative to random on QA (0.171, 95% CI [0.070, 0.271]) and MC (0.089, [0.009, 0.168]), but is unresolved on MR and SQuAD2. This indicates that the selected neurons can help maintain the current preference under a large input-state change in some datasets. Nevertheless, union sufficiency is negative in every dataset: -0.566 on QA, -0.199 on MR, -0.146 on MC, and -0.168 on SQuAD2. Because deleting the Context changes both the answer evidence and the overall task state, we treat this donor as a stress test and base the carrier conclusion primarily on the cleaner answer-swapped counterfactual.

Interpretation. Counterfactual patching localizes the level of the causal claim. The selected activations carry some state-dependent influence, as indicated by their small raw restoration effects, but they do not reconstruct the original preference more effectively than equally sized random subsets. Combined with Experiment 1, this means that the selected neurons are causally coupled to and can amplify the current answer branch, while the complete knowledge-source state remains distributed beyond their activation values.

B.5 Statistical procedure

All selected-versus-random comparisons are paired at the example level. For each example, we subtract the mean effect of its matched-random intervention sets before aggregation. We estimate two-

sided 95% percentile confidence intervals using 5,000 bootstrap resamples of examples. Permutation p-values use 100,000 random sign flips of the paired differences, followed by Benjamini–Hochberg correction within each analysis family. For high-budget directional scaling, all adjusted q-values are below 6×10^{-4}. For answer-swapped union sufficiency, all adjusted q-values are at most 6×10^{-5}. Our conclusions are based primarily on the direction, magnitude, and confidence interval of the paired effects rather than isolated significance thresholds.

B.6 Diagnostic first-order analysis

We additionally verify why the selected neurons can exhibit strong directional amplification but only limited preferential restoration. This is a diagnostic analysis, not an additional causal criterion. Let $g_j = \partial m / \partial a_j$ be the local sensitivity of the margin. The first-order scaling effect is

$$\widehat{\Delta m}_{\text{scale}} = \sum_{j \in \mathcal{S}} \sigma_j g_j a_j, \tag{30}$$

whereas a counterfactual patch is locally approximated by

$$\widehat{\Delta m}_{r \leftarrow d} = \sum_{j \in \mathcal{S}} g_j^r (a_j^r - a_j^d). \tag{31}$$

On the QA screening cohort, the predicted and measured scaling effects have Pearson correlations of 0.900 for context-preferring neurons and 0.998 for parametric-preferring neurons. Across the 12 combinations of donor, patching direction, and intervention set, patching correlations range from 0.958 to 0.993. Thus, scaling and state transfer probe different perturbation directions: scaling follows the observed activation direction used by gradient-times-activation attribution, whereas patching follows the cross-state activation displacement.

B.7 Implementation details and audit

All interventions use the frozen LLaMA-2-7B-Chat model in evaluation mode with the native transformer forward pass. For ANs, we intervene on the post-attention-aggregation scalar activations $a^{\text{AN}}_{\ell,h,m}(u)$ defined in the main text, immediately before the output projection. We do not rewrite the attention softmax, assign neuron-specific attention weights, or intervene on query/key computations. For FNs, we intervene on the gated intermediate activations $a^{\text{FN}}_{\ell,k}(u)$ immediately before the FFN output projection.

Reference-attribution extraction retains the top 512 AN and top 1,024 FN scores per layer and target on ConFiQA, and the top 1,024 AN and top 2,048 FN scores on SQuAD2. Family-normalization denominators are computed from the full positive attribution totals before retention.

B.8 Conclusion and claim boundary

Together, the two intervention experiments support a coherent causal characterization of the ReNA Atlas. First, directional scaling produces substantially larger contextual-versus-parametric margin changes for reference-contrast-selected neurons than for layer-, family-, and cardinality-matched random neurons across all four datasets. The selected neurons are therefore not merely passive correlates of response state: their activations provide reliable directional control over the model's answer preference and can amplify contextual or parametric branches.

Second, counterfactual activation patching shows that this causal influence does not make the selected neurons an independently sufficient carrier of the knowledge-source decision. Transferring their original activation values produces limited positive raw restoration but does not restore the original margin more effectively than matched random neurons. Thus, the Atlas can modulate an existing answer preference, but its activation state alone is insufficient to reconstruct that preference in a counterfactual model state.

These results connect the causal analysis to ReNA's detection mechanism. The actual-response feature $\pi^{\tau}_{i,j}(r^{\text{act}}_i)$ measures how positive target support is distributed across fixed neuron identities. The matched-random-atlas results in the main text show that part of this predictive signal is broadly distributed, while ReNA's advantage over random atlases shows that reference contrast selects a

Table C.1: Cross-domain transfer of atlas identities. Rows denote the atlas discovery domain, and columns denote the target domain. Each entry reports AUROC / Macro-F1.

Atlas source \ Target	QA	MR	MC
LLaMA-2-7B-Chat			
QA	**0.9732 / 0.9508**	0.9498 / 0.9370	0.9525 / 0.9513
MR	0.9707 / 0.9451	**0.9600 / 0.9407**	0.9577 / 0.9513
MC	0.9727 / 0.9455	0.9505 / 0.9313	**0.9582 / 0.9523**
LLaMA-2-13B-Chat			
QA	**0.9580 / 0.9186**	0.9446 / 0.9021	0.9468 / 0.9098
MR	0.9536 / 0.9132	**0.9528 / 0.9087**	0.9491 / 0.9124
MC	0.9544 / 0.9140	0.9457 / 0.8998	**0.9505 / 0.9159**

more signal-enriched readout. Directional scaling further demonstrates that the selected neurons are causally coupled to the answer-preference logits and can amplify them. At the same time, the absence of preferential patching sufficiency indicates that the complete knowledge-source state extends beyond these selected neurons. We therefore interpret the ReNA Atlas as a compact, branch-conditioned causal readout and amplifier embedded in a broader distributed network state, rather than as a small self-contained source-selection circuit.

The scope of this interpretation is bounded by the intervention protocol. The causal endpoint is the contextual-minus-parametric margin at the first divergent reference token rather than full-response accuracy, directional scaling establishes controllability rather than exclusive natural mediation, and the intervention experiments are conducted on LLaMA-2-7B-Chat. Within these boundaries, the causal results directly support the mechanistic interpretation presented in the main text.

C Additional Experimental Details

C.1 Cross-domain transfer details.

Table C.1 reports the full source–target transfer matrices used in Section 5.4. Rows indicate the domain used to discover the fixed atlas, while columns indicate the target domain on which the response-state classifier is trained and evaluated.

C.2 SQuAD2 conflict-case construction.

SQuAD2 is not natively annotated for context–parametric knowledge conflict, so we construct model-specific cases from its answerable subset. For each example, we first query the base model without the SQuAD2 context and record its closed-book answer as A_p, which serves as the parametric reference r_i^{par}. The SQuAD2 gold answer, supported by the original context, serves as the contextual reference r_i^{ctx}.

We then provide the original context and question to the same model and generate the actual response r_i^{act}. An example is retained as *context-faithful* when the context-conditioned response is correct with respect to the SQuAD2 gold answer and differs from A_p. It is retained as *parametric* when the context-conditioned response preserves the closed-book answer, i.e., $r_i^{\mathrm{act}} = A_p$, while A_p is incorrect with respect to the gold answer. Examples satisfying neither condition are excluded. After filtering, 6,000 cases are retained for evaluation. This construction tests whether the model switches to the context-supported answer after observing the evidence or persists with a conflicting parametric answer.

C.3 Baseline details.

We use seven representative hallucination detectors spanning output-, representation-, attention-, and mechanism-based signals. **P(True)** Kadavath et al. (2022) uses the model's self-assessed probability that its response is correct; **RefChecker** Hu et al. (2024) extracts response claims and verifies them against the supplied context; and **SelfCheckGPT** Manakul et al. (2023) estimates hallucination

risk from consistency across stochastic generations. **EigenScore** Chen et al. (2024) measures dispersion through the eigenspectrum of sampled-response hidden representations, while **SEP** Kossen et al. (2024) learns a hidden-state probe approximating semantic-entropy-based uncertainty. **FrequencyAware** Qi et al. (2026) uses high-frequency energy in attention trajectories across generation steps. **ReDeEP** Sun et al. (2025) combines retrieved-context utilization measured through Copying Heads with parametric-knowledge contribution measured through Knowledge FFNs.

C.4 ATLAS HYPERPARAMETERS.

We use $\tau \in \{\mathrm{AN}, \mathrm{FN}\}$ to denote attention and FFN neuron families, respectively. For LLaMA-2-7B on the three ConFiQA subsets, we set the layer-wise Atlas budgets to $K^{\mathrm{AN}} = 64$ and $K^{\mathrm{FN}} = 32$, with directional-consistency thresholds $\gamma_{\mathrm{AN}} = \gamma_{\mathrm{FN}} = 0.70$. For SQuAD2, we used $K^{\mathrm{AN}} = 256$, $K^{\mathrm{FN}} = 128$, and $\gamma_{\mathrm{AN}} = \gamma_{\mathrm{FN}} = 0.55$. For LLaMA-2-13B, we used $K^{\mathrm{AN}} = 512$ and $K^{\mathrm{FN}} = 256$ across all datasets and did not apply an additional directional-consistency filter (equivalently, $\gamma_{\mathrm{AN}} = \gamma_{\mathrm{FN}} = 0$). Atlas construction was performed independently within each layer and outer-fold training partition. Among neurons satisfying $|\kappa_j^\tau| \geq \gamma_\tau$, we selected those with the largest mean contrast magnitude μ_j^τ; both contextual- and parametric-preferring neurons were eligible. We set the normalization constant in Eq. 12 to $\varepsilon_\pi = 10^{-12}$.

C.5 INFERENCE AND ATTRIBUTION.

All language-model parameters were frozen, and the models were evaluated in inference mode on NVIDIA A800 GPUs with 80 GB of memory. Model responses were generated using greedy decoding without sampling. To compute neuron attribution, we applied teacher forcing separately to the contextual reference r_i^{ctx}, the parametric reference r_i^{par}, and the actual response r_i^{act}. Each continuation was scored using the mean token log-likelihood objective in Eq. 9, followed by one backward pass to obtain the positive gradient–activation attribution of attention and FFN neurons. Attention weights and value activations were obtained from the model's native forward computation without modifying the attention implementation. Models were loaded in FP16, while attribution aggregation was performed in FP32.

C.6 CLASSIFIER.

ReNA uses a linear classifier over the actual-response attribution shares at the fixed Atlas coordinates defined in Eq. 18. Features were standardized using statistics estimated from the current training fold and then provided to an ℓ_2-regularized logistic-regression classifier with class-balanced weights. We used the `liblinear` solver and included an intercept. The regularization parameter was set to $C = 1.0$, except for LLaMA-2-7B on SQuAD2, where $C = 0.03$ was used. The maximum number of optimization iterations was set to 5,000 to ensure convergence. The classifier was implemented using scikit-learn 1.6.1. We intentionally used a linear detector so that predictive performance primarily reflects the information encoded by the fixed-Atlas attribution pattern rather than additional nonlinear classifier capacity.

C.7 EVALUATION.

For each model–dataset configuration, we used repeated stratified five-fold cross-validation with five repeats. Atlas discovery, feature standardization, classifier fitting, and decision-threshold calibration were repeated independently within every outer training fold. Each example was therefore evaluated only as held-out data and received five out-of-fold probabilities, which were averaged across repeats. The decision threshold was selected using only the corresponding training partition to maximize balanced accuracy; neither held-out labels nor held-out reference continuations were accessed during model selection. We report AUROC as the primary metric because it is independent of the decision threshold, together with Macro-F1 as the threshold-dependent metric. All methods were evaluated using identical data partitions and the same saved model responses.

C.8 PROMPT FORMAT.

To standardize response form and reduce attribution variation caused by verbose answer scaffolding, we used the same short-answer instruction for both model sizes and all datasets. Each Context-containing input was formatted using the native LLaMA-2 chat template as follows:

```
<s>[INST] <<SYS>>
You are a helpful assistant. Directly answer the question in
no more than five words.
<</SYS>>

Context:
{context}

Question: {question} [/INST]
```

Here, {context} and {question} are replaced by the passage and query of the corresponding example. The same prompt x_i is used to generate the actual response and to teacher-force the contextual reference r_i^{ctx}, parametric reference r_i^{par}, and actual response r_i^{act} in Eq. 9; only the continuation following [/INST] changes. Thus, the reference contrast in Eq. 13 is computed under an identical Context and question. The five-word instruction is a soft generation constraint rather than constrained decoding: responses are not truncated post hoc to five words.