WHEN EVIDENCE FAILS TO PERSUADE: TRACING INTERNAL RESISTANCE TO EVIDENCE-DRIVEN PREFERENCE REVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidence resistance, the tendency of language models to resist external correction despite explicit supporting evidence, poses a fundamental challenge to external knowledge augmentation and evidence-dependent reasoning. Despite its practical importance, the mechanisms underlying this behavior remain poorly understood, with limited conceptual separation between the model's resistant behavior and the contextual–parametric conflict underlying its emergence. To address this gap, we formulate evidence resistance as a quantitative attribution problem over the internal dynamics of evidence-driven preference revision, isolating component-level responses that systematically counteract adaptation toward externally supported alternatives. Building on this formulation, we develop *Contextual Shift Attribution*, a signed attribution framework that decomposes evidence-induced answer shifts into directional component contributions and quantifies their relative influence on the model's response to external evidence. By tracing component responses along a controlled evidence-strength trajectory, our framework distinguishes mechanistically relevant resistance from incidental activation changes across progressively reinforced evidence conditions. Across six language models and three knowledge-conflict benchmarks, targeted intervention on the identified components improves the average evidence-following rate by at least 9.5% over existing baselines while largely preserving evidence-free model behavior. Our analysis reveals that *evidence resistance reflects a stable pattern of internal computation sustained by shared structure across heterogeneous evidence settings, rather than an instance-specific response to knowledge conflict.* These findings move the study of evidence integration beyond output-level conflict toward the internal computations that actively resist preference revision, providing a foundation for more precise diagnosis and control of evidence-integration failures.

1 INTRODUCTION

Modern language models increasingly integrate external evidence with their parametric knowledge to support reliable reasoning in knowledge-intensive tasks (Gao et al., 2023; Lin et al., 2025; Yadav et al., 2026; Zhao et al., 2026c). Yet models may fail to adequately shift their answer preferences toward externally supported alternatives, resulting in persistent reliance on their original predictions (Longpre et al., 2021; Xie et al., 2024; Xu et al., 2024b). We term this tendency *evidence resistance*, a behavioral response to externally induced conflict between contextual and parametric knowledge. As illustrated in Figure 1, evidence-driven preference revision can be viewed as a directional transition toward the evidence-supported alternative, shaped by competing internal responses. Evidence resistance captures the component-level contributions that oppose this transition and sustain the model's established answer preference. Conceptually, evidence resistance differs from knowledge conflict in both scope and object of analysis, shifting attention from informational inconsistency across competing knowledge sources to the internal computation that counteracts evidence-driven preference revision. Such resistance is especially consequential for evidence-dependent applications, as external knowledge augmentation and inference-time knowledge updating rely on effective revision of established model preferences. Controlling this phenomenon thus requires mechanistic characterization beyond phenomenological observation: *How can evidence resistance be quantitatively characterized within the model's internal computation?*

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.

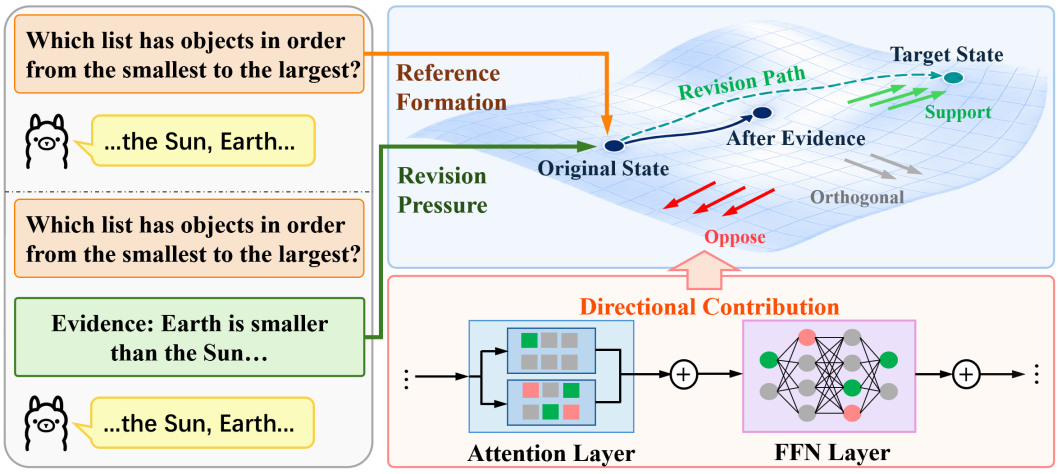

Figure 1: **Conceptual overview of evidence resistance and its component-level structure. Left:** Behavioral manifestation of evidence resistance, with the model retaining its established preference despite explicit evidence favoring an alternative answer. **Right:** A directional view of evidence-driven preference revision and its component-level decomposition, tracing distinct influences to attention heads and feed-forward network (FFN) neurons. Evidence resistance reflects the structured internal computation that counteracts movement toward the evidence-supported alternative and sustains the model's established preference.

Existing research on contextual–parametric conflict has largely studied model responses empirically at the output level, revealing varying degrees of receptivity to contextual evidence and persistence of parametric knowledge (Longpre et al., 2021; Chen et al., 2022; Xie et al., 2024; Xu et al., 2024b). More recent studies have extended this line of inquiry into the model's internal processing by localizing components associated with competing knowledge sources and probing representations correlated with contextual or parametric reliance (Ferrando et al., 2024; Zheng et al., 2024; Zhang et al., 2024; Li et al., 2025a). Such analyses reveal the internal distribution of conflict-related components, but provide limited insight into their contribution to preference revision and thus fail to distinguish resistance-driving responses from merely conflict-associated activity (Tighidet et al., 2024; Zhao et al., 2026a). As a result, the mechanistic understanding of evidence resistance remains quantitatively unresolved, hindering systematic analysis of its internal expression and targeted mitigation (Zhao et al., 2026a; Augenstein, 2025; Zhang et al., 2026).

To address this question, we ground evidence resistance through the internal dynamics of evidence-driven preference revision, focusing on component responses that oppose movement toward externally supported alternatives. To characterize these dynamics under controlled variation in contextual reinforcement, we introduce an evidence-strength trajectory by progressively reinforcing the same external assertion. This construction transforms each conflict instance into a sequence from the evidence-free state to increasingly reinforced evidence conditions, providing a controlled approximation of real-world evidence accumulation through redundant retrieval and repeated exposure. Building on this trajectory, we develop *Contextual Shift Attribution*, a training-free signed attribution framework that decomposes the resulting preference shift into these directional component contributions, as illustrated in the lower panel of Figure 1. The resulting profiles enable systematic characterization of the structure and recurrence of evidence resistance across inputs.

Our analysis reveals a small but influential subset of components whose responses persistently oppose evidence-driven changes in answer preference. Notably, these components recur across heterogeneous evidence settings, revealing *a shared and structurally consistent internal organization underlying the repeated expression of evidence resistance across distinct conflicts*. Despite representing only a small fraction of the model, these components exert disproportionate influence on the model's susceptibility to evidence-driven preference shifts. Guided by this structural regularity, we construct a targeted intervention over consistently resistance-associated components, yielding substantial gains in evidence following while maintaining the model's evidence-free behavior. These findings motivate precise control strategies that selectively target resistance-associated computation rather than broadly altering the model's response to external evidence. Our contributions can be summarized as follows:

• **Problem Formulation.** We identify evidence resistance as a structured model-internal behavior arising from component responses that persistently counteract evidence-driven preference revision, and formulate its mechanistic analysis as a quantitative attribution problem over inference-time revision dynamics. This formulation translates evidence resistance from a behavioral manifestation of knowledge conflict into a measurable computational object for mechanistic analysis.

• **Contextual Shift Attribution.** We introduce a quantitative attribution framework that decomposes evidence-induced preference shifts into directional component contributions, separating resistance-associated responses from incidental activation changes across progressively reinforced evidence conditions. The resulting contribution profiles enable systematic characterization of recurrent resistance patterns across heterogeneous evidence settings.

• **Functional Structure of Evidence Resistance.** We reveal a recurrent subset of components that consistently exert opposing influence on evidence-driven preference revision and demonstrate their significance through targeted intervention. Selectively intervening on these components improves the average evidence-following rate by at least 9.5% over baselines while preserving evidence-free behavior, suggesting a practical path toward targeted correction of evidence integration failures.

2 Related Work

Reliability of Contextual Knowledge Integration. Contextual knowledge integration in real-world applications challenges language models (Chen et al., 2024; Fan et al., 2024; Mei et al., 2025; Cheng et al., 2025), including knowledge conflicts (Xu et al., 2024b; Wang et al., 2024; Liu & Roth, 2025), susceptibility to misinformation (Zhou et al., 2024; Ni et al., 2026; Li et al., 2025b), and contextual unfaithfulness (Ji et al., 2023; Huang et al., 2025). Knowledge conflicts arise from inconsistencies across external sources or between contextual and parametric knowledge (Chen et al., 2022; Xu et al., 2024b), and are addressed through cross-source corroboration, compatibility-aware integration, and source-aware consolidation (Weller et al., 2024; Zhang et al., 2023; Wang et al., 2025). Susceptibility to misinformation has been associated with inadequate robustness to unreliable evidence and increased reliance on user-provided information after instruction tuning (Xu et al., 2024c; Han et al., 2025), with mitigation strategies including information refinement, adversarial fine-tuning, and reinforcement learning for knowledge selection (Xu et al., 2024c; Fang et al., 2024; Lin et al., 2026). Contextual unfaithfulness refers to discrepancies between generated claims and the supplied context (Huang et al., 2025; Liu et al., 2025), and is commonly addressed through context-faithful prompting and evidence-guided response revision (Zhou et al., 2023; Peng et al., 2023). These perspectives primarily characterize antecedent conditions or outcome-level failures (Xu et al., 2024b; Han et al., 2025; Liu et al., 2025), whereas evidence resistance concerns the model's dynamic response that counteracts preference revision as external evidence is reinforced.

Contextual Influence on Parametric Reliance. The practical significance of contextual influence in knowledge-intensive applications (Faridi et al., 2026; Yadav et al., 2026) has motivated approaches that regulate contextual–parametric reliance at both the output and component levels (Tonmoy et al., 2024; Wang et al., 2026; Ferrando et al., 2024; Zheng et al., 2024; Li et al., 2025a; Zhao et al., 2026b). Specifically, at the output level CAD adjusts next-token distributions by contrasting logits obtained with and without external context (Shi et al., 2024), whereas at the component level, PASTA localizes and steers selected attention heads toward designated context spans (Zhang et al., 2024). Our framework instead quantifies the directional contributions of individual components to preference revision, distinguishing resistance-driving responses from conflict-associated activity.

3 Mechanistic Characterization of Evidence Resistance

In this section, we formulate evidence resistance as a component-level attribution problem over the internal dynamics of evidence-driven preference revision and derive a targeted intervention from the resulting attribution structure. We first construct a controlled evidence-response trajectory from the evidence-free state to progressively reinforced evidence conditions. We then introduce a signed attribution framework to quantify the directional contributions of attention heads and FFN neurons, separating supportive from resistance-associated responses. Finally, we identify a sparse set of resistance-associated components and restore their outputs to query-specific evidence-free references, suppressing resistance-associated computation without modifying model parameters.

3.1 Evidence-Driven Preference Revision and Component Representation

We formalize evidence resistance through changes in the model's preference between its original answer and an evidence-supported alternative. Consider a decoder-only Transformer f with fixed parameters and deterministic inference, with an affine unembedding applied to the final normalized hidden state. We assume that the attention-head and FFN contributions defined below enter the residual stream additively without an intervening nonlinear transformation. Let q denote a query-containing prompt, e an external assertion supporting answer a_e, and z a fixed suffix at the prediction position. We define the evidence-free and evidence-conditioned inputs as $x^{(0)} = q\|z$ and $x^{(1)} = q\|e\|z$. Let a_r denote the model's reference answer under $x^{(0)}$, with $a_e \neq a_r$. For prompt x, we define the answer-preference margin

$$\phi(x) = \ell_{a_e}(x) - \ell_{a_r}(x) = \log \frac{p(a_e \mid x)}{p(a_r \mid x)}, \tag{1}$$

where $\ell_a(x)$ is the next-token logit of a. Negative values favor a_r, whereas positive values favor a_e. For $\phi(x^{(0)}) < 0$, the shift $\phi(x^{(1)}) - \phi(x^{(0)})$ measures evidence-driven preference revision, while $\phi(x^{(1)}) \leq 0$ indicates persistent preference for the original answer.

To resolve this shift into internal contributions, we represent attention heads and FFN neurons by their additive writes to the residual stream at the final prompt position:

$$\mathbf{v}^A_{\ell,h} = W^{(h)}_{O,\ell}\mathbf{a}_{\ell,h}, \qquad \mathbf{v}^F_{\ell,r} = \mathbf{w}^{\text{down}}_{\ell,r} m_{\ell,r}, \tag{2}$$

where $\mathbf{a}_{\ell,h}$ is the head output before projection, $m_{\ell,r}$ is the FFN activation before down projection, and the corresponding projection terms map these quantities into their additive residual-stream contributions. We use c to denote either an attention head or an FFN neuron, with $\mathbf{v}_c$ denoting its residual-stream contribution. Interventions are restricted to the final prompt position, while earlier positions remain unmodified.

3.2 Signed Contextual Shift Attribution

To characterize the internal dynamics of evidence-driven preference revision, we vary contextual reinforcement through finite repetition of the same external assertion. This construction yields a response trajectory from the evidence-free baseline to increasingly reinforced evidence conditions. Specifically, we define

$$x^{(n)} = q\|e^n\|z, \qquad n = 0, \ldots, N^\star, \tag{3}$$

where e includes a fixed delimiter and $N^\star \geq 1$ is a finite repetition budget. Let $\mathbf{v}_c(n)$ and $\phi(n)$ denote the component output and answer-preference margin under $x^{(n)}$. Under this setting, the component outputs of a decoder-only Transformer converge to finite limits as $n \to \infty$, with the conditions and proof detailed in Appendix D. This convergence establishes a stable reference for characterizing each component's response to progressively reinforced evidence. Accordingly, we adopt a sufficiently large finite repetition budget $N^\star$ so that $\mathbf{v}_c(N^\star)$ approximates this limiting reference. The resulting displacement $\mathbf{v}_c(N^\star) - \mathbf{v}_c(0)$ captures the component's contextual response relative to its evidence-free state, providing the basis for quantifying its directional influence on evidence-driven preference revision.

For each step, we define the component displacement and downstream sensitivity as

$$\boldsymbol{\delta}_c(n) = \mathbf{v}_c(n) - \mathbf{v}_c(n-1), \qquad \mathbf{G}_c(n) = \nabla_{\mathbf{v}_c}\phi|_{x^{(n)}}. \tag{4}$$

The gradient quantifies the local sensitivity of the answer-preference margin to the component output at $x^{(n)}$, accounting for both direct effects and those mediated by downstream computation. The inner product $\langle G_c(n-1), \delta_c(n)\rangle$ estimates the signed contribution of the component's evidence-induced response to the answer-preference shift. Aggregating these stepwise contributions along the reinforcement trajectory yields our proposed Contextual Shift Attribution score.

Definition 3.1 (Contextual shift attribution, CSA)**.** The cumulative contextual shift attribution score along the finite repetition sequence is

$$\mathcal{A}_c(N^\star) = \sum_{n=1}^{N^\star} \langle \mathbf{G}_c(n-1), \boldsymbol{\delta}_c(n)\rangle, \tag{5}$$

where each summand predicts the margin change from replacing only $\mathbf{v}_c(n-1)$ by $\mathbf{v}_c(n)$ in the computation for $x^{(n-1)}$. The attribution score quantifies the direction and relative strength of each component's contribution to evidence-driven preference revision, enabling quantitative identification of resistance-associated responses within the model.

For computational efficiency, we approximate the cumulative score using gradients evaluated at a set of anchor points $N_1, ..., N_m \in \{0, \ldots, N^\star\}$:

$$\widehat{\mathcal{A}}_c = \frac{1}{m} \sum_{m'=1}^{m} \langle \mathbf{G}_c(N_{m'}), \mathbf{v}_c(N^\star) - \mathbf{v}_c(0) \rangle \,. \tag{6}$$

$\widehat{\mathcal{A}}_c$ approximates the cumulative attribution with an explicit error guarantee, as established by the following proposition.

Proposition 3.2 (Frozen-gradient error)**.** *Let* $V_c = \sum_{n=1}^{N^\star} \|\boldsymbol{\delta}_c(n)\|_2$ *and* $\varepsilon_c = \max_{1 \le n \le N^\star} \|\mathbf{G}_c(n-1) - \frac{1}{m} \sum_{m'=1}^{m} \mathbf{G}_c(N_{m'})\|_2$*. Then*

$$|\mathcal{A}_c(N^\star) - \widehat{\mathcal{A}}_c| \le \varepsilon_c V_c. \tag{7}$$

The two scores have the same sign whenever $|\widehat{\mathcal{A}}_c| > \varepsilon_c V_c$*.*

Proposition 3.2 bounds the discrepancy between $\widehat{\mathcal{A}}_c$ and $\mathcal{A}_c(N^\star)$, establishing the validity of efficient CSA estimation under controlled gradient variation. The resulting score captures the signed influence of each component's contextual response on evidence-driven preference revision, with its local replacement interpretation formalized in Appendix C.2.

3.3 Resistance-Associated Component Identification and Intervention

We next aggregate component-level attribution responses across samples to identify recurrent resistance-associated components and derive a targeted intervention from the resulting attribution structure. Let $\mathcal{D}$ denote a calibration set of query–evidence samples with designated target answers. We summarize the query-dependent attribution of each component by its median over $\mathcal{D}$:

$$\widetilde{\mathcal{A}}_c = \mathrm{median}_{(q,e,a_e) \in \mathcal{D}} \, \widehat{\mathcal{A}}_c(q, e, a_e). \tag{8}$$

Components with strongly negative aggregated attribution consistently oppose evidence-driven preference revision. We therefore identify resistance-associated components by selecting the lower α-quantile of the attribution distribution, with $\alpha \in (0, 1)$:

$$\mathcal{C}_{\mathcal{P}}^{-} = \left\{ c \in \mathcal{P} : \; \widetilde{\mathcal{A}}_c \le Q_\alpha\Big(\{ \widetilde{\mathcal{A}}_{c'} : c' \in \mathcal{P} \} \Big) \right\}, \tag{9}$$

where $\mathcal{P}$ denotes the candidate component pool and Q_α its empirical lower α-quantile. The pool may be defined globally or within individual layers, with the corresponding configurations and selection thresholds specified in Appendix A.2.

We then intervene directly on these localized components. For each query, we obtain the selected components' reference outputs $\mathbf{v}_c(0)$ from the unmodified model on the evidence-free prompt $x^{(0)}$ and substitute these values for the corresponding component outputs at the prediction position during inference on the evidence-conditioned prompt $x^{(1)}$:

$$\widehat{p}(\cdot \mid q, e) = f\Big(x^{(1)} \;\Big|\; \mathrm{do}\big(\mathbf{v}_c = \mathbf{v}_c(0),\; c \in \mathcal{C}_p^{-}\big) \Big). \tag{10}$$

This design integrates component selectivity with query-specific reference substitution to target resistance-associated computation while preserving native predictions on matched evidence-free inputs by construction. To complement our attribution-based formulation with a theoretical characterization of intervention effects, we further relate directional component attribution to preference changes induced by reference patching, providing sufficient conditions for single-component margin improvement and a local analysis of joint interventions detailed in Appendix C.4.

Table 1: **Evidence-guided preference revision under counterfactual knowledge conflict.** Percentage of post-intervention predictions aligned with the evidence-supported answer, evaluated on instances where the unmodified model initially predicts a conflicting answer. Baseline results use the *best-performing configurations* selected through systematic hyperparameter tuning. FE and FA denote FaithEval and FARM, respectively. **Bold** and underlined values indicate the best and second-best results within each model–benchmark setting.

Model	CAD			RCA			PASTA			JuICE			Ours		
	FE	KRE	FA	FE	KRE	FA	FE	KRE	FA	FE	KRE	FA	FE	KRE	FA
Qwen3-1.7B	24.5	38.3	47.4	31.1	51.6	52.5	38.2	66.1	55.8	**82.3**	88.0	76.2	71.4	**88.3**	**82.9**
Qwen3-4B	24.7	70.5	72.4	31.9	71.1	57.2	60.8	89.6	83.9	68.1	92.2	68.8	**75.7**	**93.7**	**93.0**
Qwen3-8B	37.2	72.7	76.9	44.1	74.5	76.5	79.5	94.7	88.6	81.6	92.2	77.9	**87.9**	**96.3**	**95.4**
Llama-3.2-1B	57.1	66.8	72.8	48.2	70.1	64.8	69.0	82.8	**78.7**	62.5	65.2	69.9	**77.0**	**84.1**	75.0
Llama-3.2-3B	63.4	74.1	87.2	70.8	78.9	80.8	83.5	94.0	79.8	77.9	90.9	78.2	**91.0**	**94.4**	**88.1**
Llama-3.1-8B	53.1	66.2	85.2	68.6	71.1	86.2	82.1	85.7	89.8	77.9	78.8	59.6	**90.0**	**95.0**	**93.6**
Mean	43.3	64.8	73.7	49.1	69.6	69.7	68.9	85.5	79.4	75.1	84.6	71.8	**82.2**	**92.0**	**88.0**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks and Models. We evaluate six large language models from two model families across a range of scales: Qwen3-1.7B, Qwen3-4B, Qwen3-8B (Team, 2025), Llama-3.2-1B, Llama-3.2-3B (Meta AI, 2024), and Llama-3.1-8B (Grattafiori et al., 2024). These models span small ($<$3B), medium (3B–6B), and large ($>$6B) parameter regimes. Experiments are conducted on three knowledge-conflict benchmarks: FaithEval (Ming et al., 2025) (977 QA instances), KRE (Ying et al., 2024) (1,120), and FARM (Xu et al., 2024a) (885). Each benchmark provides counterfactual evidence that conflicts with factual knowledge, enabling evaluation across diverse conflict settings. Our framework is training-free, with experiments conducted on NVIDIA A100 and RTX A6000 GPUs. Additional benchmark and model details are provided in Appendix A.1.

Component Selection and Reference Patching. We compute CSA scores for each attention head and MLP neuron, aggregate them across the evaluation set by the median, and select a sparse subset with the strongest negative responses. During evidence-conditioned inference, the selected outputs at the prediction position are replaced with their evidence-free counterparts from the same query. This query-specific reference patching removes resistance-associated contextual responses while preserving the remaining computation. Implementation details are provided in Appendix A.2.

Cross-Benchmark Transfer and Generalization Protocol. To rule out benchmark-specific calibration effects, we evaluate cross-benchmark generalization by localizing resistance-associated components on one benchmark and directly transferring them to the others without target-specific reselection. We assess structural and functional generalization using component-set overlap and evidence-following performance under these frozen source-localized component sets.

4.2 COMPARATIVE EVALUATION OF EVIDENCE-GUIDED MODEL STEERING

We first evaluate the effectiveness of the localized resistance-associated components through evidence-guided preference revision under knowledge conflict. We compare our intervention against four representative inference-time steering baselines spanning decoding- and component-level contextual control including CAD (Shi et al., 2024), RCA (Zhao et al., 2026b), PASTA (Zhang et al., 2024), and JuICE (Li et al., 2025a). All methods are evaluated on instances where the unmodified model's prediction conflicts with the provided evidence, isolating cases that require revision of an established model preference. To ensure a competitive comparison, we systematically tune every baseline and report its *best-performing configuration subject* to preserving linguistically meaningful outputs. Full hyperparameter search results are provided in Appendix A.3. As shown in Table 1, our proposed intervention consistently produces stronger evidence-guided preference revision across heterogeneous models and conflict settings. Notably, our intervention surpasses existing methods

by at least **9.5%** in overall average performance while modifying only **1.74%** of attention heads and **0.29%** of FFN neurons. The effectiveness of such sparse intervention supports the specific role of the localized components in resisting evidence-driven preference revision, beyond a generic increase in contextual reliance. More importantly, these results empirically validate our mechanistic formulation of *evidence resistance as a structured internal behavior supported by a distinct, manipulable computational substrate rather than an incidental consequence of knowledge conflict*.

4.3 Preservation of Parametric Knowledge under Intervention

We next evaluate the preservation of the model's original parametric knowledge following intervention. We evaluate each method on evidence-free questions using prediction accuracy against ground-truth answers, with Table 2 summarizing knowledge preservation across models. The prompt template is provided in Appendix A.4. Our intervention consistently preserves strong evidence-free performance, with substantially less degradation than several baselines. This behavior is consistent with the prediction invariance induced by reference patching on the corresponding evidence-free input. These results indicate that improved responsiveness to external evidence does not arise from broadly weakening original knowledge or indiscriminately suppressing prior preferences. Instead, the intervention selectively modulates resistance-associated computation while preserving evidence-free behavior ensured by the reference-patching design. These results further support the localized resistance mechanism's specificity and demonstrate enhanced evidence-guided revision without globally disrupting parametric knowledge.

Table 2: **Parametric knowledge preservation after intervention.** Prediction accuracy on evidence-free questions, with higher values indicating better retention of the model's original knowledge.

Model	CAD	RCA	PASTA	JuICE	Ours
Qwen3-1.7B	58.6	59.1	**60.5**	22.8	60.2
Qwen3-4B	74.5	69.1	73.9	70.9	**74.9**
Qwen3-8B	79.5	79.2	76.2	76.1	**79.8**
Llama-3.2-1B	40.2	41.3	29.1	27.0	**42.5**
Llama-3.2-3B	54.7	55.1	53.7	44.2	**59.6**
Llama-3.1-8B	66.6	66.9	58.5	36.1	**68.9**
Mean	62.4	61.8	58.7	46.2	**64.3**

4.4 Ablation and Structural Analysis of Resistance-Associated Components

We next compare targeted and random patching under an identical intervention budget to isolate the contribution of component selection from the reference-patching operation. The random control uses the same intervention procedure and component budget, replacing resistance-associated components with randomly sampled attention heads and FFN neurons. Results are averaged over three random seeds and reported as mean ± standard deviation in Table 3. As shown, random patching remains close to the unmodified model across most settings, whereas targeted patching consistently yields substantially stronger evidence-guided preference revision across settings. The resulting separation demonstrates that the intervention effect is driven by the identity of the localized components rather than by patching-induced perturbation alone, providing direct validation of the resistance-associated computation identified by our framework.

We further examine the distribution of attribution responses across component types to characterize the internal structure of evidence resistance. For each query, we aggregate attention-head and FFN attribution responses within each layer and pool the resulting layer-wise values across queries to obtain benchmark-level distributions. Figure 2 reveals

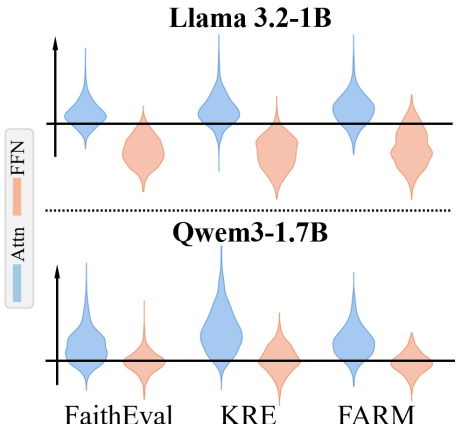

Figure 2: **Component-level attribution distributions.** Violin plots compare attention-head and FFN-neuron scores across settings, with positive and negative values indicating supportive and opposing contributions to preference revision. FFN profiles are consistently more negative than attention-head profiles, revealing a component-type asymmetry in resistance-associated computation.

Table 3: **Ablation study of resistance-associated component selection under matched intervention budgets.** Evidence-guided preference revision for the unmodified model (Dense), random patching (Random), and targeted patching (Ours). Random controls use the same reference-patching procedure with matched numbers of attention heads and FFN neurons. Results for Random are reported as mean ± standard deviation over three random seeds.

Model	Benchmarks								
	FaithEval			KRE			FARM		
	Dense	Random	Ours	Dense	Random	Ours	Dense	Random	Ours
Qwen3-1.7B	17.2	13.7 ± 8.1	**71.4**	31.1	29.1 ± 4.2	**88.3**	39.3	31.5 ± 12.9	**82.9**
Qwen3-4B	16.5	14.2 ± 0.2	**75.7**	52.7	43.7 ± 4.1	**93.7**	61.4	56.3 ± 2.1	**93.0**
Qwen3-8B	21.8	28.4 ± 1.9	**87.9**	56.6	66.7 ± 5.3	**96.3**	65.4	52.4 ± 17.9	**95.4**
Llama-3.2-1B	37.2	36.6 ± 1.2	**77.0**	45.3	42.2 ± 2.3	**84.1**	57.3	46.7 ± 13.4	**75.0**
Llama-3.2-3B	49.4	47.4 ± 9.0	**91.0**	59.8	55.5 ± 17.0	**94.4**	79.5	79.7 ± 3.4	**88.1**
Llama-3.1-8B	38.1	34.6 ± 7.7	**90.0**	51.5	54.0 ± 4.2	**95.0**	77.2	62.5 ± 25.6	**93.6**

a consistent directional asymmetry between attention heads and FFN neurons across settings. Attention-head responses are predominantly concentrated on the positive side of the preference-revision axis, whereas FFN responses exhibit substantially stronger negative mass. This separation indicates that resistance-associated computation is not uniformly distributed across the Transformer architecture, but is preferentially expressed through FFN components that oppose evidence-induced answer changes. The recurrence of this pattern across heterogeneous settings further supports a structured and selective organization of resistance-associated computation within models. Appendix B.2 provides the full block-level analysis.

4.5 Cross-Benchmark Generalization of Resistance-Associated Components

We next examine the consistency of resistance-associated components across heterogeneous evidence sources and conflict settings. For each model, we independently localize components on FaithEval, KRE, and FARM to assess their cross-benchmark structural recurrence. We quantify the overlap among the components using the Szymkiewicz–Simpson coefficient. As shown in Table 4, resistance-associated attention heads exhibit substantial overlap across benchmarks, with average pairwise overlaps of at least **80.0%** and a comparably high three-way overlap. This consistency indicates that localization captures a recurrent subset of attention heads across distinct conflicts without being dominated by benchmark-specific variation, supporting shared resistance-associated computation under heterogeneous evidence conditions and its potential cross-setting transferability.

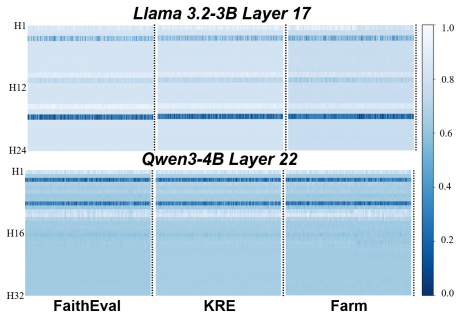

Figure 3: **Cross-benchmark consistency of attention-head attribution response.** Heatmaps show query-level attribution profiles for representative layers across benchmarks, with rows denoting attention heads and columns individual queries. The persistence of head-specific attribution structure across benchmarks supports a shared pattern of resistance-associated attention behavior.

To examine this consistency beyond set-level overlap, we visualize query-level attribution responses for representative layers in Figure 3. The resulting response patterns remain highly consistent across FaithEval, KRE, and FARM, with the same heads exhibiting distinctive attribution profiles across heterogeneous queries. This fine-grained consistency complements the overlap statistics in Table 4, showing that cross-benchmark recurrence extends beyond the selected component sets to the underlying response structure of individual heads. Appendix B.3 extends this analysis to all settings.

We finally evaluate out-of-benchmark generalization under a cross-benchmark calibration–test protocol, which constitutes a more stringent distribution-shift test than a within-benchmark split, to

Table 4: **Cross-benchmark overlap of resistance-associated attention heads.** Szymkiewicz–Simpson coefficients quantify the overlap between attention-head sets independently localized on FaithEval, KRE, and FARM. Pairwise and three-way overlaps are reported for each model, with parenthesized fractions showing the intersection size relative to the smaller selected set.

Model	FE–KRE	FE–FA	KRE–FA	All three
Qwen3-1.7B	0.88 (7/8)	0.88 (7/8)	0.88 (7/8)	0.88 (7/8)
Qwen3-4B	0.75 (27/36)	0.64 (23/36)	0.67 (24/36)	0.58 (21/36)
Qwen3-8B	0.87 (26/30)	0.83 (25/30)	0.87 (26/30)	0.77 (23/30)
Llama-3.2-1B	0.75 (3/4)	0.75 (3/4)	0.75 (3/4)	0.75 (3/4)
Llama-3.2-3B	0.88 (7/8)	0.88 (7/8)	0.75 (6/8)	0.75 (6/8)
Llama-3.1-8B	0.90 (9/10)	0.90 (9/10)	0.90 (9/10)	0.90 (9/10)
Mean	0.84	0.81	0.80	0.77

Table 5: **Cross-benchmark transferability of resistance-associated components.** Values report evidence-following rates (%) obtained by applying component sets identified on a source benchmark to target benchmarks within the same model. Column groups denote source benchmarks, and subcolumns denote target benchmarks. Results with matching source and target benchmarks correspond to benchmark-specific component selection and serve as the reference performance.

Model	**Source**								
	FaithEval			**KRE**			**FARM**		
	FaithEval	KRE	FARM	FaithEval	KRE	FARM	FaithEval	KRE	FARM
Qwen3-1.7B	71.4	87.4	80.4	72.4	88.3	80.2	72.7	88.9	82.9
Qwen3-4B	75.7	93.8	92.9	69.8	93.7	91.9	67.6	93.4	93.0
Qwen3-8B	87.9	96.2	95.8	86.8	96.3	95.9	87.9	94.7	95.4
Llama-3.2-1B	77.0	82.6	77.9	82.2	84.1	77.7	74.8	80.4	75.0
Llama-3.2-3B	91.0	95.8	87.4	88.9	94.4	91.1	90.5	95.8	88.1
Llama-3.1-8B	90.0	95.2	92.5	89.6	95.0	92.8	89.8	96.0	93.6

assess whether the proposed framework identifies a shared and transferable computational mechanism of evidence resistance rather than specific patterns. For each model, components localized on one source benchmark are directly applied to all remaining benchmarks without target-specific reselection. As shown in Table 5, cross-benchmark component transfer largely preserves the original evidence-guided preference revision performance, with an average decrease of only **0.44%** relative to benchmark-specific component selection. This result extends the structural consistency in Table 4 to practical transferability, showing that resistance-associated components identified in one evidence setting remain effective across distinct conflicts. Such cross-setting stability indicates that the observed intervention gains are not confined to the benchmark used for component localization, but persist under out-of-benchmark transfer without target-specific component scoring or reselection.

5 Conclusions

We characterized evidence resistance as a distinct model-internal behavior underlying failures of evidence-driven preference revision and formulated its mechanistic analysis as a component-level attribution problem for identifying internal responses that systematically oppose external correction. We instantiated this formulation through a signed attribution framework that tracks component responses along controlled changes in evidence strength and attributes their directional influence on preference revision. Our results show that resistance is concentrated in a sparse and recurrent subset of components rather than diffusely distributed across the model, with these components exhibiting stable opposing effects across heterogeneous evidence settings. Intervening on this localized computation substantially improves responsiveness to external evidence while preserving native predictions on matched evidence-free inputs by construction, indicating that resistance can be selectively mitigated without disrupting parametric knowledge. These findings establish evidence resistance as a quantitatively characterizable internal behavior, offering a principled pathway toward its mechanistic diagnosis and targeted mitigation.

AI USE STATEMENT

In this work, we used a generative AI coding agent to assist with method implementation and experimental execution, as well as the evaluation of prompt templates described in Appendix A.4.

We have **not** used generative AI tools for generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis.

Interpreting results and assisting with translation are not applicable to this work.

We have reviewed all AI-assisted work. We used a generative AI coding agent to implement the computational procedures derived from our theoretical formulation, including CSA score computation, median aggregation (Equation 8), component selection (Equation 9), and reference patching (Equation 10). The agent was also used to execute the corresponding experiments, conduct hyperparameter searches and evaluations for the baseline methods, and collect numerical results. Additionally, we used the agent to evaluate and select the prompt templates described in Appendix A.4, with selection guided by the unmodified (dense) models' compliance with the specified response format and output-validity criteria. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Isabelle Augenstein. Understanding the interplay between llms' utilisation of parametric and contextual knowledge: A keynote at ecir 2025. In *ACM SIGIR Forum*, volume 59, pp. 1–11. ACM New York, NY, USA, 2025.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.146. URL https://aclanthology.org/2022.emnlp-main.146/.

Yuxuan Chen, Daniel Röder, Justus-Jonas Erker, Leonhard Hennig, Philippe Thomas, Sebastian Möller, and Roland Roller. Retrieval-augmented knowledge integration into language models: A survey. In Sha Li, Manling Li, Michael JQ Zhang, Eunsol Choi, Mor Geva, Peter Hase, and Heng Ji (eds.), *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pp. 45–63, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.knowllm-1.5. URL https://aclanthology.org/2024.knowllm-1.5/.

Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, et al. A survey on knowledge-oriented retrieval-augmented generation. *ACM Transactions on Information Systems*, 2025.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 6491–6501, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671470. URL https://doi.org/10.1145/3637528.3671470.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10028–10039, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.540. URL https://aclanthology.org/2024.acl-long.540/.

Arman Rasool Faridi, Faraz Masood, Keshvi, and Tabassum Yunus. Retrieval-augmented generation for large language models: Evolution, architectures, applications, and challenges (2020–2025). *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 16(3):e70122, 2026.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-Jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kyubeen Han, Junseo Jang, Hongjin Kim, Geunyeong Jeong, and Harksoo Kim. Exploring the impact of instruction-tuning on LLM's susceptibility to misinformation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26711–26731, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1295. URL https://aclanthology.org/2025.acl-long.1295/.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. *arXiv preprint arXiv:2503.10996*, 2025a.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. Knowledge boundary of large language models: A survey. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5131–5157, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.256. URL https://aclanthology.org/2025.acl-long.256/.

Chenyu Lin, Yilin Wen, Du Su, Hexiang Tan, Fei Sun, Muhan Chen, Chenfu Bao, and Zhonghou Lv. Resisting contextual interference in rag via parametric-knowledge reinforcement. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 46721–46742, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4dbcd4ba82866fdefc7ea0d15e9a72f2-Paper-Conference.pdf.

Yu-Hsuan Lin, Qian-Hui Chen, Yi-Jie Cheng, Jia-Ren Zhang, Yi-Hung Liu, Liang-Yu Hsia, and Yun-Nung Chen. Llm inference enhanced by external knowledge: A survey. *arXiv preprint arXiv:2505.24377*, 2025.

Siyi Liu and Dan Roth. Conflicts in texts: Data, implications and challenges. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10073–10091, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.534. URL https://aclanthology.org/2025.findings-emnlp.534/.

Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. Towards long context hallucination detection. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7842–7850, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.436. URL https://aclanthology.org/2025.findings-naacl.436/.

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, 2021.

Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*, 2025.

Meta AI. Llama 3.2 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md, 2024. Accessed 2026.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". In *International Conference on Learning Representations*, volume 2025, pp. 29430–29456, 2025.

Bo Ni, Zheyuan Liu, Yongjia Lei, Leyao Wang, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Xin Dong, Yinglong Xia, Krishnaram Kenthapadi, et al. Towards trustworthy retrieval augmented generation for large language models: A survey. *ACM Computing Surveys*, 58(15):1–36, 2026.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *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, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zineddine Tighidet, Jiali Mei, Benjamin Piwowarski, and Patrick Gallinari. Probing language models on their knowledge source. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 604–614, 2024.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1476. URL https://aclanthology.org/2025.acl-long.1476/.

Haoran Wang, Xiongxiao Xu, Philip S Yu, and Kai Shu. Beyond tokens: A survey on decoding methods for large language and vision-language models. *ACM SIGKDD Explorations Newsletter*, 28(1):1–20, 2026.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. Knowledge mechanisms in large language models: A survey and perspective. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7097–7135, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.416. URL https://aclanthology.org/2024.findings-emnlp.416/.

Orion Weller, Aleem Khan, Nathaniel Weir, Dawn Lawrie, and Benjamin Van Durme. Defending against disinformation attacks in open-domain question answering. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 402–417, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-short.35. URL https://aclanthology.org/2024.eacl-short.35/.

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*, volume 2024, pp. 35623–35646, 2024.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16259–16303, 2024a.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. Unsupervised information refinement training of large language models for retrieval-augmented generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 133–145, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.9. URL https://aclanthology.org/2024.acl-long.9/.

Itisha Yadav, Sirko Schindler, Diana Peters, and Roman Klinger. External knowledge integration in large language models: A survey on methods, challenges, and future directions. *Semantic Web*, 17(4):22104968261453132, 2026.

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. Intuitive or dependent? investigating llms' behavior style to conflicting prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4221–4246, 2024.

Hengyuan Zhang, Zhihao Zhang, Mingyang Wang, Zunhai Su, Yiwei Wang, Qianli Wang, Shuzhou Yuan, Ercong Nie, Xufeng Duan, Qibo Xue, et al. Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models. *Computer Science Review*, 62:101011, 2026.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for llms. In *International Conference on Learning Representations*, volume 2024, pp. 42411–42430, 2024.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. Merging generated and retrieved knowledge for open-domain QA. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4710–4728, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.286. URL https://aclanthology.org/2023.emnlp-main.286/.

Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Understanding parametric and contextual knowledge reconciliation within large language models. *Advances in Neural Information Processing Systems*, 38:102978–103012, 2026a.

Mingkuan Zhao, Yide Gao, Wentao Hu, Suquan Chen, Tianchen Huang, Zhenhua An, Zetao Chang, Xiayu Sun, and Yuheng Min. Resonant context anchoring: Decoupling attention routing and signal gain at inference time. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 36600–36611, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1824. URL https://aclanthology.org/2026.findings-acl.1824/.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *Data Science and Engineering*, 11(1):1–29, 2026c.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*, 2024.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968/.

Yujia Zhou, Wenbo Zhang, Jingying Shao, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Jason Chen Zhang, et al. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*, 2024.

A Implementation Details

A.1 Benchmarks Pre-processing

FaithEval. Experiments are performed on the counterfactual subset of FaithEval, which contains 1,000 multiple-choice instances. Each instance pairs a question with a context passage written to support a counterfactual option, which is designated as the answer key, and a justification that identifies the fabricated evidence in the passage. We retain only instances with exactly four options labeled A–D, yielding 977 instances. Since the answer key marks the counterfactual option, we recover the factual answer to each question from its original ARC annotation via the shared question ID. For each instance, the counterfactual option is used as the distractor. Rather than the full context passage, we use as the external evidence the fabricated sentence quoted in the justification, keeping the shortest one when several are quoted; when the justification quotes no sentence, we take the context sentence with the highest word overlap with the justification.

KRE. Experiments are performed on the COSE subset of KRE, which contains 1,221 multiple-choice instances. Each instance provides a question with five options, a gold answer, a designated counterfactual option, and a short negative context supporting the counterfactual option. We first remove 4 instances in which the gold and counterfactual options have identical text due to duplicated options. We then reduce each instance to four options by keeping the gold and counterfactual options and randomly sampling two of the remaining three, and shuffle the option order. Finally, we manually remove 97 instances whose gold answer is questionable or whose counterfactual option is not a meaningful alternative answer (e.g., off-topic, ungrammatical, or arguably also correct), yielding 1,120 instances. For each instance, the counterfactual option is used as the distractor, and the negative context is used verbatim as the external evidence.

FARM. Experiments are performed on FARM, which consists of four subsets: NQ1 (488 instances), NQ2 (489), TruthfulQA (484), and BoolQ (491). Each instance provides a question, a misinformation target, and persuasive passages in four styles: a plain assertion of the target, and three passages each of logical, credibility, and emotional appeal. All subsets except BoolQ additionally provide an option list that marks the correct answer, the misinformation target, and the remaining incorrect options. We exclude BoolQ, which has binary answers and no options, and NQ1, whose misinformation targets are negations of the correct answer (e.g., "not Yale") rather than alternative options. All 489 NQ2 instances already contain four options and are retained, with the option marked as correct taken as the gold answer. TruthfulQA instances contain 2 to 24 options, and we reduce each to four by keeping the best answer and the misinformation target and randomly sampling two incorrect options. We discard 87 TruthfulQA instances with fewer than two incorrect options, rather than filling the remaining slots with additional correct answers, and one instance with an empty misinformation target, leaving 396. After shuffling the option order, this yields 885 instances in total. For each instance, the misinformation target is used as the distractor, and the first logical-appeal passage is used as the external evidence.

A.2 Model Patching

Component selection. For each component, we take the median of its CSA over all questions of a benchmark. A negative median CSA means that the component's response to the evidence pulls the model back toward its initial answer, so we patch the components with the most negative median CSA, ranking attention heads and MLP neurons separately. For five of the six models we use a global top-k rule: all heads are ranked jointly across layers and the k_A most negative ones are selected, and k_M neurons are selected in the same way. This rule imposes no per-layer quota, so the selected heads concentrate in a few layers. For Qwen3-4B we use a per-layer top-quantile rule: in every layer, we select the components whose median CSA is at or below that layer's 0.005 quantile, i.e., the most negative 0.5%. Because the threshold is inclusive, this selects the single most negative head in each of the 36 layers and 49–50 of the 9,728 neurons in each layer. In all 18 model–benchmark pairs, every selected component has a negative median CSA. The rule and its size are fixed per model and shared by the three benchmarks. They are chosen among global budgets (k_A, k_M) on a grid and the per-layer 0.5% rule, as the configuration that maximizes the mean evidence-following rate over the three benchmarks while keeping the mean probability mass on the option letters at or above 0.5 on every benchmark, the same protocol used to tune the baselines (Table 6).

Table 6: Selection rule and patched components per model, shown as patched / total (share). Under the global top-k rule, the patched counts are k_A and k_M. Mean is the average share over the six models.

Model	Selection rule	Attention heads	MLP neurons
Llama-3.2-1B	global top-k	4 / 512 (0.78%)	250 / 131,072 (0.19%)
Llama-3.2-3B	global top-k	8 / 672 (1.19%)	500 / 229,376 (0.22%)
Llama-3.1-8B	global top-k	10 / 1,024 (0.98%)	250 / 458,752 (0.05%)
Qwen3-1.7B	global top-k	8 / 448 (1.79%)	500 / 172,032 (0.29%)
Qwen3-4B	per-layer 0.5% quantile	36 / 1,152 (3.13%)	1,764–1,766 / 350,208 (0.50%)
Qwen3-8B	global top-k	30 / 1,152 (2.60%)	2,000 / 442,368 (0.45%)
Mean		1.74%	0.29%

Patching. We patch attention heads at their output before the output projection W_O and MLP neurons at their post-activation value entering the down-projection. For each question, we record the activations of the selected components on the evidence-free prompt, i.e., the same question and options without the evidence section. When the model reads the evidence-bearing prompt, we overwrite these components at the final prompt position with the recorded values and leave all other components and positions unchanged. The intervention updates no parameters and requires one extra forward pass. The patched components amount to 0.78–3.13% (mean 1.74%) of a model's attention heads and 0.05–0.50% (mean 0.28%) of its MLP neurons.

A.3 Comparison Methods Implementation

We compare against four training-free methods that steer the model toward the evidence: CAD, RCA, PASTA, and JUICE. All four are run on the same prompts, the same samples, and the same evidence-following criterion as our method. Each method has a strength parameter, which we search on every benchmark; for each model we then fix one setting that is shared by the three benchmarks. Tables 7–11 report, for every searched setting, the evidence-following rate (the share of usable samples with $P(d) > P(c_0)$ after the intervention) and the mean probability mass that the intervened model places on the option letters, on each benchmark. Let ℓ_{ctx} and ℓ_{prior} denote the next-token logits at the answer position with and without the evidence section.

Selection rule. A setting is admissible for a model if it has been evaluated on all three benchmarks and its mean option-letter mass is at least 0.5 on every benchmark; among admissible settings we select the one with the highest mean evidence-following rate over the three benchmarks. The mass guard excludes settings that raise the rate by pushing the model off the option letters altogether, in which case the comparison $P(d) > P(c_0)$ is won trivially. The same rule selects the component budget of our method.

CAD. CAD decodes from $(1+\alpha)\,\ell_{\mathrm{ctx}} - \alpha\,\ell_{\mathrm{prior}}$. We search $\alpha \in \{0, 0.25, 0.5, 1, 1.5, 2, 3\}$, where $\alpha = 0$ is the unmodified model. The selected value is $\alpha = 1$ for Llama-3.2-1B, Llama-3.2-3B, and Qwen3-4B, and $\alpha = 0.5$ for Llama-3.1-8B, Qwen3-1.7B, and Qwen3-8B (Table 7). The guard is binding for every model: the next larger α still raises the mean rate but drops the letter mass below 0.5 on at least one benchmark.

RCA. RCA rescales, at the decoding step and in every layer and head, the value vector of each attended position i by $\lambda_i = 1 + \gamma\,\mathrm{softplus}(s_i)$, where s_i is the pre-softmax attention score. The released implementation applies the gain to every prompt position; we additionally evaluate a variant that restricts it to the evidence span. We search $\gamma \in \{0, 0.02, 0.04, 0.05, 0.08, 0.12, 0.2, 0.5, 1.5\}$ for both variants (Tables 8 and 9). The evidence-span variant is selected for every model, with $\gamma = 0.2$ for Llama-3.2-1B and Qwen3-4B, $\gamma = 0.5$ for Llama-3.2-3B and Llama-3.1-8B, and $\gamma = 0.12$ for Qwen3-1.7B and Qwen3-8B. The full-prompt gain reaches a lower admissible mean rate on every Llama model and, on the Qwen models, stays close to or below the unmodified model until the letter mass collapses. Since RCA acts at a decoding step after the prompt is prefilled, its $\gamma = 0$ rows can differ from the dense values by a few tenths of a point.

PASTA. PASTA adds $\log \alpha$ to the pre-softmax attention scores of all key positions outside the evidence span for a selected set of heads, so that these heads attend more to the evidence. Following the original procedure, we steer one head at a time with $\alpha = 0.01$ on the first 48 questions of each benchmark and keep the 20 heads with the highest evidence-following rate. We search $\alpha \in \{1, 0.3, 0.1, 0.03, 0.01, 0.003\}$, where $\alpha = 1$ is the unmodified model. The selected value is $\alpha = 0.1$ for the three Llama models, $\alpha = 0.3$ for Qwen3-1.7B, and $\alpha = 0.003$ for Qwen3-4B and Qwen3-8B (Table 10). On the Llama models, $\alpha = 0.1$ is also the setting with the highest mean rate, so the guard is not binding. On Qwen3-1.7B, every $\alpha \leq 0.1$ reaches a higher mean rate but lowers the letter mass on FaithEval to 0.31 or less, so the selection stops at $\alpha = 0.3$. On Qwen3-4B and Qwen3-8B, all six settings pass the guard and the rate saturates from $\alpha = 0.03$ downward.

JUICE. JUICE runs the model twice: the first pass records the output $o_{l,h}$ of each selected head, and the second pass adds $\alpha_h \, o_{l,h}$ to the attention output at every position, with $\alpha_h = -\alpha_s$ for a set of heads to suppress and $\alpha_h = +\alpha_l$ for a set of heads to enlarge. Following the official procedure, the two sets (10 heads each) are found on 4 questions by scoring every head's effect on the probability of the evidence-supported option under three prompt settings. We evaluate $(\alpha_s, \alpha_l) \in \{0, \ldots, 5\}^2$ on the first 300 questions of each benchmark, take the three best settings per benchmark among those that keep at least 0.55 of the probability mass on the option letters, and run the four settings with the highest summed evidence-following rate on the full set of all three benchmarks. The selected setting is $(\alpha_s, \alpha_l) = (3, 4)$ for Llama-3.2-1B, $(1, 2)$ for Llama-3.2-3B, $(2, 0)$ for Llama-3.1-8B, $(3, 1)$ for Qwen3-1.7B, $(1, 4)$ for Qwen3-4B, and $(1, 3)$ for Qwen3-8B (Table 11). All candidates pass the guard, so the selection is by mean rate alone. For Llama-3.1-8B the selected setting has $\alpha_l = 0$, i.e., only the suppression set is active.

A.4 Prompt Template

All prompts share a single sectioned template and are fed to the model as plain text, without a chat template. An instruction line asks for a single letter, followed by a `Question:` section, a `Choices:` section listing one option per line, an optional `Evidence:` section, and a final `Answer:` marker. The model's answer is read from its next-token distribution at the last prompt position, restricted to the option letters. For a given question, all prompts share the same instruction, question, and options, and differ only in the Evidence section. We use four variants: the clean prompt and the repeated-hint prompt for locating components, the natural-evidence prompt for evaluating evidence following, and the placeholder prompt for measuring accuracy without evidence. The examples below use the same KRE question with Qwen3-8B.

A.4.1 Clean Prompt

The clean prompt $x^{(0)}$ contains only the question and its options: the Evidence section, including its header, is removed entirely. It has three uses. The option letter with the highest probability on $x^{(0)}$ defines the model's initial answer c_0. Its activations serve as the evidence-free reference when computing CSA. Its activations at the final position are also the values written into the selected components during patching.

Clean Template

```
Reply with only one letter:
Question:
<question>
Choices:
A. <option A>
B. <option B>
C. <option C>
D. <option D>
Answer:
```

Example (the model answers D):

Table 7: CAD: evidence-following rate (%) and mean option-letter probability mass for every searched α. Dense is the unmodified model. Guard passes when the mass is at least 0.50 on every benchmark; the selected setting (bold) is the guard-passing setting with the highest mean rate.

Setting	FaithEval		KRE		FARM		Mean	Guard
	rate	mass	rate	mass	rate	mass	rate	
Llama-3.2-1B (dense: 37.2 / 45.3 / 57.3)								
$\alpha = 0$	37.2	0.98	45.3	0.98	57.3	0.96	46.6	✓
$\alpha = 0.25$	46.9	0.97	54.0	0.97	65.3	0.92	55.4	✓
$\alpha = 0.5$	51.6	0.95	60.2	0.96	69.5	0.85	60.4	✓
$\alpha = 1$	**57.1**	**0.85**	**66.8**	**0.89**	**72.8**	**0.60**	**65.6**	✓
$\alpha = 1.5$	62.9	0.67	73.0	0.75	76.4	0.35	70.8	–
$\alpha = 2$	67.1	0.44	74.7	0.54	78.1	0.17	73.3	–
$\alpha = 3$	70.7	0.12	78.6	0.20	81.0	0.03	76.8	–
Llama-3.2-3B (dense: 49.4 / 59.8 / 79.5)								
$\alpha = 0$	49.4	0.97	59.8	0.97	79.5	0.94	62.9	✓
$\alpha = 0.25$	54.7	0.97	65.9	0.95	83.5	0.91	68.1	✓
$\alpha = 0.5$	58.4	0.95	68.8	0.90	85.2	0.86	70.8	✓
$\alpha = 1$	**63.4**	**0.81**	**74.1**	**0.68**	**87.2**	**0.63**	**74.9**	✓
$\alpha = 1.5$	66.9	0.54	76.3	0.40	88.8	0.36	77.3	–
$\alpha = 2$	68.9	0.30	77.8	0.21	89.7	0.19	78.8	–
$\alpha = 3$	73.0	0.09	80.2	0.06	90.9	0.05	81.4	–
Llama-3.1-8B (dense: 38.1 / 51.5 / 77.2)								
$\alpha = 0$	38.1	0.85	51.5	0.87	77.2	0.72	55.6	✓
$\alpha = 0.25$	47.5	0.81	60.7	0.83	83.0	0.64	63.8	✓
$\alpha = 0.5$	**53.1**	**0.75**	**66.2**	**0.75**	**85.2**	**0.53**	**68.2**	✓
$\alpha = 1$	63.0	0.52	74.1	0.45	88.2	0.29	75.1	–
$\alpha = 1.5$	70.7	0.26	79.3	0.19	90.0	0.14	80.0	–
$\alpha = 2$	75.6	0.12	83.4	0.07	91.1	0.06	83.4	–
$\alpha = 3$	83.0	0.03	86.4	0.01	92.5	0.01	87.3	–
Qwen3-1.7B (dense: 17.2 / 31.1 / 39.3)								
$\alpha = 0$	17.2	0.95	31.1	0.98	39.3	0.92	29.2	✓
$\alpha = 0.25$	21.2	0.93	35.1	0.97	43.9	0.89	33.4	✓
$\alpha = 0.5$	**24.5**	**0.85**	**38.3**	**0.92**	**47.4**	**0.79**	**36.7**	✓
$\alpha = 1$	31.8	0.43	43.5	0.58	56.0	0.42	43.7	–
$\alpha = 1.5$	39.1	0.13	48.6	0.27	61.0	0.15	49.6	–
$\alpha = 2$	44.2	0.04	52.0	0.13	65.6	0.06	53.9	–
$\alpha = 3$	50.8	0.01	56.2	0.06	70.8	0.02	59.3	–
Qwen3-4B (dense: 16.5 / 52.7 / 61.4)								
$\alpha = 0$	16.5	0.97	52.7	0.96	61.4	0.96	43.5	✓
$\alpha = 0.25$	19.2	0.96	60.0	0.96	66.4	0.95	48.6	✓
$\alpha = 0.5$	21.5	0.93	65.4	0.93	69.5	0.93	52.1	✓
$\alpha = 1$	**24.7**	**0.67**	**70.5**	**0.76**	**72.4**	**0.83**	**55.9**	✓
$\alpha = 1.5$	28.0	0.34	73.8	0.56	75.2	0.68	59.0	–
$\alpha = 2$	31.4	0.18	76.0	0.43	76.5	0.57	61.3	–
$\alpha = 3$	36.3	0.08	79.4	0.29	79.3	0.44	65.0	–
Qwen3-8B (dense: 21.8 / 56.6 / 65.4)								
$\alpha = 0$	21.8	0.94	56.6	0.95	65.4	0.95	47.9	✓
$\alpha = 0.25$	29.8	0.90	66.0	0.94	72.7	0.96	56.2	✓
$\alpha = 0.5$	**37.2**	**0.72**	**72.7**	**0.84**	**76.9**	**0.95**	**62.3**	✓
$\alpha = 1$	46.3	0.29	78.5	0.49	82.1	0.83	69.0	–
$\alpha = 1.5$	53.0	0.15	82.9	0.29	85.1	0.71	73.7	–
$\alpha = 2$	59.9	0.10	85.1	0.19	87.6	0.62	77.5	–
$\alpha = 3$	68.3	0.05	88.8	0.10	88.8	0.51	82.0	–

Table 8: RCA with the released full-prompt gain: evidence-following rate (%) and mean option-letter probability mass for every searched γ. Dense is the unmodified model. Guard passes when the mass is at least 0.50 on every benchmark; the selected setting (bold) is the guard-passing setting with the highest mean rate. No full-prompt setting is selected for any model (see Table 9).

Setting	FaithEval		KRE		FARM		Mean	Guard
	rate	mass	rate	mass	rate	mass	rate	
Llama-3.2-1B (dense: 37.2 / 45.3 / 57.3)								
$\gamma = 0$	36.4	0.98	45.0	0.98	57.6	0.96	46.4	✓
$\gamma = 0.02$	38.8	0.98	50.2	0.98	59.8	0.96	49.6	✓
$\gamma = 0.04$	40.3	0.98	53.7	0.98	62.4	0.96	52.1	✓
$\gamma = 0.05$	41.1	0.98	55.9	0.98	63.5	0.96	53.5	✓
$\gamma = 0.08$	44.2	0.97	59.8	0.97	65.1	0.95	56.4	✓
$\gamma = 0.12$	47.4	0.95	66.3	0.96	66.3	0.93	60.0	✓
$\gamma = 0.2$	45.6	0.89	71.1	0.94	65.8	0.89	60.8	✓
$\gamma = 0.5$	44.1	0.46	43.6	0.70	39.8	0.29	42.5	–
$\gamma = 1.5$	54.3	0.12	42.3	0.08	40.7	0.06	45.8	–
Llama-3.2-3B (dense: 49.4 / 59.8 / 79.5)								
$\gamma = 0$	49.0	0.97	60.2	0.97	80.1	0.94	63.1	✓
$\gamma = 0.02$	48.8	0.97	60.1	0.97	80.1	0.94	63.0	✓
$\gamma = 0.04$	48.8	0.97	61.5	0.97	80.3	0.94	63.5	✓
$\gamma = 0.05$	49.1	0.97	61.9	0.97	80.5	0.94	63.8	✓
$\gamma = 0.08$	50.1	0.96	62.1	0.96	80.3	0.94	64.1	✓
$\gamma = 0.12$	52.5	0.95	63.1	0.96	79.9	0.93	65.2	✓
$\gamma = 0.2$	55.3	0.94	65.0	0.95	80.2	0.91	66.9	✓
$\gamma = 0.5$	51.6	0.86	68.8	0.90	73.8	0.84	64.7	✓
$\gamma = 1.5$	41.4	0.49	35.1	0.57	46.9	0.48	41.1	–
Llama-3.1-8B (dense: 38.1 / 51.5 / 77.2)								
$\gamma = 0$	37.8	0.85	51.6	0.87	76.8	0.72	55.4	✓
$\gamma = 0.02$	41.5	0.86	53.9	0.89	77.9	0.73	57.8	✓
$\gamma = 0.04$	43.0	0.87	55.3	0.90	78.4	0.75	58.9	✓
$\gamma = 0.05$	43.7	0.88	56.0	0.90	79.4	0.75	59.7	✓
$\gamma = 0.08$	46.0	0.89	56.6	0.91	79.8	0.76	60.8	✓
$\gamma = 0.12$	47.4	0.91	55.0	0.92	79.0	0.76	60.5	✓
$\gamma = 0.2$	42.6	0.94	49.7	0.94	75.8	0.75	56.0	✓
$\gamma = 0.5$	33.7	0.94	35.9	0.96	66.0	0.73	45.2	✓
$\gamma = 1.5$	31.2	0.68	33.5	0.66	46.2	0.30	37.0	–
Qwen3-1.7B (dense: 17.2 / 31.1 / 39.3)								
$\gamma = 0$	16.8	0.95	30.8	0.98	38.9	0.93	28.8	✓
$\gamma = 0.02$	22.0	0.97	38.8	0.99	41.2	0.95	34.0	✓
$\gamma = 0.04$	22.4	0.96	38.3	0.99	43.3	0.94	34.7	✓
$\gamma = 0.05$	17.7	0.90	34.1	0.98	40.4	0.88	30.7	✓
$\gamma = 0.08$	10.1	0.24	15.5	0.43	22.7	0.09	16.1	–
$\gamma = 0.12$	27.0	0.00	25.2	0.00	34.5	0.00	28.9	–
$\gamma = 0.2$	34.9	0.00	37.6	0.00	38.9	0.00	37.1	–
$\gamma = 0.5$	45.4	0.00	45.1	0.00	45.3	0.00	45.3	–
$\gamma = 1.5$	44.4	0.00	42.1	0.00	45.3	0.00	43.9	–
Qwen3-4B (dense: 16.5 / 52.7 / 61.4)								
$\gamma = 0$	16.7	0.97	52.5	0.96	60.9	0.96	43.4	✓
$\gamma = 0.02$	11.3	0.98	36.5	0.98	53.2	0.98	33.7	✓
$\gamma = 0.04$	6.3	0.99	17.6	0.99	41.2	0.99	21.7	✓
$\gamma = 0.05$	5.7	0.99	11.4	0.99	35.1	0.99	17.4	✓
$\gamma = 0.08$	5.1	0.99	5.6	0.99	21.3	0.98	10.6	✓
$\gamma = 0.12$	5.5	0.93	5.7	0.96	15.7	0.93	9.0	✓
$\gamma = 0.2$	19.9	0.01	15.2	0.02	18.0	0.01	17.7	–
$\gamma = 0.5$	41.7	0.00	33.5	0.00	33.7	0.01	36.3	–
$\gamma = 1.5$	50.8	0.00	42.7	0.00	37.5	0.00	43.7	–
Qwen3-8B (dense: 21.8 / 56.6 / 65.4)								
$\gamma = 0$	21.7	0.94	56.8	0.95	65.4	0.95	48.0	✓
$\gamma = 0.02$	17.9	0.94	53.3	0.95	63.5	0.95	44.9	✓
$\gamma = 0.04$	13.0	0.94	44.4	0.94	57.1	0.94	38.2	✓
$\gamma = 0.05$	11.1	0.92	38.8	0.93	50.7	0.92	33.5	✓
$\gamma = 0.08$	5.7	0.76	22.1	0.79	34.6	0.76	20.8	✓
$\gamma = 0.12$	4.6	0.17	11.9	0.27	18.5	0.23	11.7	–
$\gamma = 0.2$	13.1	0.00	16.7	0.00	26.8	0.00	18.9	–
$\gamma = 0.5$	23.1	0.00	22.1	0.00	29.9	0.00	25.0	–
$\gamma = 1.5$	25.9	0.00	25.4	0.00	30.2	0.01	27.2	–

Table 9: RCA with the gain restricted to the evidence span: evidence-following rate (%) and mean option-letter probability mass for every searched γ. Dense is the unmodified model. Guard passes when the mass is at least 0.50 on every benchmark; the selected setting (bold) is the guard-passing setting with the highest mean rate.

Setting	FaithEval		KRE		FARM		Mean	Guard
	rate	mass	rate	mass	rate	mass	rate	
Llama-3.2-1B (dense: 37.2 / 45.3 / 57.3)								
$\gamma = 0$	36.4	0.98	45.0	0.98	57.6	0.96	46.4	✓
$\gamma = 0.02$	39.0	0.98	47.2	0.98	58.5	0.95	48.2	✓
$\gamma = 0.04$	40.3	0.98	50.3	0.98	59.5	0.95	50.0	✓
$\gamma = 0.05$	40.2	0.98	51.7	0.98	59.5	0.95	50.4	✓
$\gamma = 0.08$	42.2	0.98	55.5	0.98	60.5	0.95	52.7	✓
$\gamma = 0.12$	44.6	0.97	60.5	0.98	61.8	0.94	55.6	✓
$\gamma = 0.2$	**48.2**	**0.97**	**70.1**	**0.97**	**64.8**	**0.92**	**61.0**	✓
$\gamma = 0.5$	56.3	0.92	72.0	0.93	50.8	0.74	59.7	✓
$\gamma = 1.5$	54.0	0.32	40.4	0.21	40.5	0.10	45.0	–
Llama-3.2-3B (dense: 49.4 / 59.8 / 79.5)								
$\gamma = 0$	49.0	0.97	60.2	0.97	80.1	0.94	63.1	✓
$\gamma = 0.02$	51.0	0.97	63.2	0.97	80.2	0.94	64.8	✓
$\gamma = 0.04$	52.3	0.97	65.8	0.97	80.8	0.94	66.3	✓
$\gamma = 0.05$	53.5	0.97	66.7	0.97	80.8	0.94	67.0	✓
$\gamma = 0.08$	56.6	0.97	70.3	0.97	82.6	0.94	69.8	✓
$\gamma = 0.12$	59.4	0.97	73.2	0.96	83.0	0.93	71.9	✓
$\gamma = 0.2$	63.2	0.97	77.3	0.96	82.6	0.92	74.3	✓
$\gamma = 0.5$	**70.8**	**0.93**	**78.9**	**0.91**	**80.8**	**0.89**	**76.8**	✓
$\gamma = 1.5$	41.9	0.59	31.8	0.53	38.9	0.67	37.5	✓
Llama-3.1-8B (dense: 38.1 / 51.5 / 77.2)								
$\gamma = 0$	37.8	0.85	51.6	0.87	76.8	0.72	55.4	✓
$\gamma = 0.02$	42.9	0.85	54.2	0.87	79.0	0.72	58.7	✓
$\gamma = 0.04$	44.8	0.85	57.0	0.87	80.4	0.72	60.7	✓
$\gamma = 0.05$	46.6	0.85	58.6	0.87	81.2	0.72	62.1	✓
$\gamma = 0.08$	49.9	0.85	61.8	0.87	82.0	0.71	64.6	✓
$\gamma = 0.12$	54.3	0.85	67.0	0.87	83.4	0.71	68.2	✓
$\gamma = 0.2$	61.1	0.85	71.6	0.87	85.6	0.68	72.8	✓
$\gamma = 0.5$	**68.6**	**0.80**	**71.1**	**0.82**	**86.2**	**0.54**	**75.3**	✓
$\gamma = 1.5$	55.5	0.70	62.7	0.59	48.0	0.36	55.4	–
Qwen3-1.7B (dense: 17.2 / 31.1 / 39.3)								
$\gamma = 0$	16.8	0.95	30.8	0.98	38.9	0.93	28.8	✓
$\gamma = 0.02$	20.0	0.95	35.6	0.98	43.0	0.94	32.9	✓
$\gamma = 0.04$	23.4	0.95	40.6	0.97	45.8	0.94	36.6	✓
$\gamma = 0.05$	24.9	0.94	42.3	0.97	46.5	0.95	37.9	✓
$\gamma = 0.08$	28.5	0.94	49.1	0.95	51.0	0.95	42.9	✓
$\gamma = 0.12$	**31.1**	**0.92**	**51.6**	**0.94**	**52.5**	**0.93**	**45.0**	✓
$\gamma = 0.2$	29.0	0.76	43.6	0.88	38.3	0.43	37.0	–
$\gamma = 0.5$	37.5	0.02	30.1	0.06	41.6	0.00	36.4	–
$\gamma = 1.5$	42.5	0.00	41.8	0.00	44.0	0.00	42.8	–
Qwen3-4B (dense: 16.5 / 52.7 / 61.4)								
$\gamma = 0$	16.7	0.97	52.5	0.96	60.9	0.96	43.4	✓
$\gamma = 0.02$	19.4	0.97	60.1	0.97	63.9	0.96	47.8	✓
$\gamma = 0.04$	21.6	0.96	64.2	0.97	65.3	0.95	50.4	✓
$\gamma = 0.05$	22.7	0.96	65.5	0.97	65.7	0.95	51.3	✓
$\gamma = 0.08$	25.2	0.94	67.4	0.97	64.3	0.94	52.3	✓
$\gamma = 0.12$	27.4	0.92	69.3	0.96	62.1	0.92	52.9	✓
$\gamma = 0.2$	**31.9**	**0.86**	**71.1**	**0.95**	**57.2**	**0.83**	**53.4**	✓
$\gamma = 0.5$	35.7	0.13	37.4	0.07	32.2	0.03	35.1	–
$\gamma = 1.5$	50.9	0.00	45.2	0.00	47.2	0.00	47.7	–
Qwen3-8B (dense: 21.8 / 56.6 / 65.4)								
$\gamma = 0$	21.7	0.94	56.8	0.95	65.4	0.95	48.0	✓
$\gamma = 0.02$	27.1	0.93	62.8	0.95	69.0	0.95	53.0	✓
$\gamma = 0.04$	31.0	0.93	68.0	0.95	72.9	0.95	57.3	✓
$\gamma = 0.05$	34.4	0.92	70.2	0.95	73.9	0.95	59.5	✓
$\gamma = 0.08$	39.2	0.91	73.1	0.95	74.7	0.94	62.3	✓
$\gamma = 0.12$	**44.1**	**0.87**	**74.5**	**0.94**	**76.5**	**0.90**	**65.0**	✓
$\gamma = 0.2$	45.3	0.71	74.4	0.85	70.5	0.65	63.4	✓
$\gamma = 0.5$	27.3	0.11	42.4	0.17	36.3	0.07	35.3	–
$\gamma = 1.5$	42.3	0.00	43.6	0.00	42.5	0.00	42.8	–

Table 10: PASTA: evidence-following rate (%) and mean option-letter probability mass for every searched α (20 selected heads per benchmark). Dense is the unmodified model. Guard passes when the mass is at least 0.50 on every benchmark; the selected setting (bold) is the guard-passing setting with the highest mean rate.

Setting	FaithEval		KRE		FARM		Mean	Guard
	rate	mass	rate	mass	rate	mass	rate	
Llama-3.2-1B	(dense: 37.2 / 45.3 / 57.3)							
$\alpha = 1$	37.2	0.98	45.3	0.98	57.3	0.96	46.6	✓
$\alpha = 0.3$	63.2	0.97	76.9	0.97	76.0	0.93	72.0	✓
$\alpha = 0.1$	**69.0**	**0.94**	**82.8**	**0.89**	**78.7**	**0.88**	**76.8**	✓
$\alpha = 0.03$	61.9	0.83	74.5	0.71	65.3	0.76	67.2	✓
$\alpha = 0.01$	55.0	0.39	61.9	0.60	50.2	0.57	55.7	–
$\alpha = 0.003$	55.6	0.19	53.9	0.49	45.2	0.43	51.5	–
Llama-3.2-3B	(dense: 49.4 / 59.8 / 79.5)							
$\alpha = 1$	49.4	0.97	59.8	0.97	79.5	0.94	62.9	✓
$\alpha = 0.3$	77.3	0.96	89.9	0.96	79.5	0.94	82.2	✓
$\alpha = 0.1$	**83.5**	**0.95**	**94.0**	**0.94**	**79.8**	**0.94**	**85.8**	✓
$\alpha = 0.03$	75.1	0.91	80.1	0.90	80.6	0.93	78.6	✓
$\alpha = 0.01$	58.8	0.85	49.9	0.85	80.2	0.90	63.0	✓
$\alpha = 0.003$	49.6	0.62	37.5	0.81	79.4	0.92	55.5	✓
Llama-3.1-8B	(dense: 38.1 / 51.5 / 77.2)							
$\alpha = 1$	38.1	0.85	51.5	0.87	77.2	0.72	55.6	✓
$\alpha = 0.3$	73.5	0.83	87.0	0.89	88.2	0.73	82.9	✓
$\alpha = 0.1$	**82.1**	**0.72**	**85.7**	**0.87**	**89.8**	**0.69**	**85.8**	✓
$\alpha = 0.03$	78.5	0.39	77.2	0.77	88.5	0.54	81.4	–
$\alpha = 0.01$	70.9	0.19	67.7	0.70	80.9	0.39	73.2	–
$\alpha = 0.003$	64.0	0.15	61.6	0.73	63.8	0.32	63.1	–
Qwen3-1.7B	(dense: 17.2 / 31.1 / 39.3)							
$\alpha = 1$	17.2	0.95	31.1	0.98	39.3	0.92	29.2	✓
$\alpha = 0.3$	**38.2**	**0.75**	**66.1**	**0.94**	**55.8**	**0.97**	**53.4**	✓
$\alpha = 0.1$	46.2	0.31	77.2	0.90	66.9	0.98	63.4	–
$\alpha = 0.03$	47.2	0.15	82.6	0.90	72.7	0.98	67.5	–
$\alpha = 0.01$	47.9	0.14	83.4	0.88	73.5	0.98	68.3	–
$\alpha = 0.003$	46.8	0.15	83.5	0.86	71.4	0.95	67.2	–
Qwen3-4B	(dense: 16.5 / 52.7 / 61.4)							
$\alpha = 1$	16.5	0.97	52.7	0.96	61.4	0.96	43.5	✓
$\alpha = 0.3$	30.4	0.96	79.3	0.98	75.4	0.97	61.7	✓
$\alpha = 0.1$	48.1	0.93	86.7	0.99	81.8	0.98	72.2	✓
$\alpha = 0.03$	58.7	0.85	89.4	0.99	83.6	0.99	77.2	✓
$\alpha = 0.01$	60.4	0.75	89.5	0.99	83.9	0.99	77.9	✓
$\alpha = 0.003$	**60.8**	**0.68**	**89.6**	**0.99**	**83.9**	**0.99**	**78.1**	✓
Qwen3-8B	(dense: 21.8 / 56.6 / 65.4)							
$\alpha = 1$	21.8	0.94	56.6	0.95	65.4	0.95	47.9	✓
$\alpha = 0.3$	46.1	0.93	85.2	0.94	80.0	0.95	70.4	✓
$\alpha = 0.1$	67.4	0.94	92.5	0.95	87.3	0.95	82.4	✓
$\alpha = 0.03$	76.5	0.94	94.0	0.95	88.9	0.95	86.4	✓
$\alpha = 0.01$	78.6	0.93	94.4	0.96	89.6	0.95	87.5	✓
$\alpha = 0.003$	**79.5**	**0.92**	**94.7**	**0.96**	**88.6**	**0.95**	**87.6**	✓

Table 11: JuICE: evidence-following rate (%) and mean option-letter probability mass for the four (α_s, α_l) candidates per model that were run on all three benchmarks; the per-benchmark dev-grid winners that were not run on the other benchmarks are omitted. Dense is the unmodified model. Guard passes when the mass is at least 0.50 on every benchmark; the selected setting (bold) is the guard-passing setting with the highest mean rate.

Setting	FaithEval		KRE		FARM		Mean	Guard
	rate	mass	rate	mass	rate	mass	rate	
Llama-3.2-1B (dense: 37.2 / 45.3 / 57.3)								
$(\alpha_s, \alpha_l) = (2, 0)$	57.8	0.97	68.5	0.98	64.5	0.98	63.6	✓
$(\alpha_s, \alpha_l) = (2, 4)$	55.7	0.96	70.0	0.98	65.4	0.99	63.7	✓
$(\alpha_s, \alpha_l) = (3, 4)$	**62.5**	**0.82**	**65.2**	**0.96**	**69.9**	**0.98**	**65.9**	✓
$(\alpha_s, \alpha_l) = (3, 5)$	61.8	0.83	65.5	0.96	69.0	0.98	65.4	✓
Llama-3.2-3B (dense: 49.4 / 59.8 / 79.5)								
$(\alpha_s, \alpha_l) = (0, 0)$	49.4	0.97	59.8	0.97	79.5	0.94	62.9	✓
$(\alpha_s, \alpha_l) = (0, 1)$	59.9	0.97	75.6	0.98	78.9	0.95	71.5	✓
$(\alpha_s, \alpha_l) = (0, 2)$	66.7	0.97	84.3	0.98	77.8	0.96	76.3	✓
$(\alpha_s, \alpha_l) = (1, 2)$	**77.9**	**0.98**	**90.9**	**0.97**	**78.2**	**0.97**	**82.4**	✓
Llama-3.1-8B (dense: 38.1 / 51.5 / 77.2)								
$(\alpha_s, \alpha_l) = (0, 2)$	40.4	0.86	55.8	0.90	83.5	0.90	59.9	✓
$(\alpha_s, \alpha_l) = (0, 3)$	40.9	0.86	55.9	0.91	82.3	0.92	59.7	✓
$(\alpha_s, \alpha_l) = (0, 4)$	41.7	0.85	53.1	0.92	79.5	0.92	58.1	✓
$(\alpha_s, \alpha_l) = (2, 0)$	**77.9**	**0.95**	**78.8**	**0.95**	**59.6**	**0.97**	**72.1**	✓
Qwen3-1.7B (dense: 17.2 / 31.1 / 39.3)								
$(\alpha_s, \alpha_l) = (1, 0)$	85.1	0.98	93.4	1.00	61.6	0.97	80.0	✓
$(\alpha_s, \alpha_l) = (1, 1)$	83.3	0.99	92.0	1.00	62.8	0.99	79.4	✓
$(\alpha_s, \alpha_l) = (2, 1)$	82.6	0.98	90.5	0.96	71.3	0.99	81.5	✓
$(\alpha_s, \alpha_l) = (3, 1)$	**82.3**	**0.82**	**88.0**	**0.83**	**76.2**	**0.99**	**82.1**	✓
Qwen3-4B (dense: 16.5 / 52.7 / 61.4)								
$(\alpha_s, \alpha_l) = (1, 0)$	29.9	0.98	86.3	0.99	70.7	0.99	62.3	✓
$(\alpha_s, \alpha_l) = (1, 3)$	64.2	0.99	92.8	1.00	69.0	1.00	75.3	✓
$(\alpha_s, \alpha_l) = (1, 4)$	**68.1**	**0.99**	**92.2**	**1.00**	**68.8**	**1.00**	**76.4**	✓
$(\alpha_s, \alpha_l) = (2, 1)$	65.1	0.98	91.7	1.00	69.1	1.00	75.3	✓
Qwen3-8B (dense: 21.8 / 56.6 / 65.4)								
$(\alpha_s, \alpha_l) = (1, 0)$	48.5	0.94	80.2	0.96	76.5	0.99	68.4	✓
$(\alpha_s, \alpha_l) = (1, 3)$	**81.6**	**0.96**	**92.2**	**0.98**	**77.9**	**1.00**	**83.9**	✓
$(\alpha_s, \alpha_l) = (2, 0)$	66.8	0.96	81.7	0.97	76.3	1.00	75.0	✓
$(\alpha_s, \alpha_l) = (2, 1)$	79.4	0.97	89.9	0.99	76.2	1.00	81.8	✓

```
Reply with only one letter:
Question:
A revolving door is convenient for two direction travel, but it
also serves as a security measure at a what?
Choices:
A. department store
B. new york
C. library
D. bank
Answer:
```

A.4.2 REPEATED-HINT PROMPT

To locate components, we construct a prompt $x^{(N)}$ that asserts a distractor N times. For each model and question, the distractor d is drawn uniformly from the options other than the model's initial answer c_0, so d can be the gold answer when c_0 is wrong. The Evidence section repeats the assertion "d. ¡option d¿ is true." N times, separated by single spaces. The evidence-bearing activations are taken at $N = 1000$, and CSA gradients are averaged over $N \in \{200, 400, 600, 800, 1000\}$. Because only the Evidence section grows with N, all tokens before it are identical across every N and match the clean prompt.

Repeated-Hint Template

```
Reply with only one letter:
Question:
<question>
Choices:
A. <option A>
B. <option B>
C. <option C>
D. <option D>
Evidence:
<d>.  <option d> is true.  <d>.  <option d> is true.  ...  <d>.
<option d> is true.  (N times)
Answer:
```

Example ($d =$ A, shown with $N = 3$):

```
Reply with only one letter:
Question:
A revolving door is convenient for two direction travel, but it
also serves as a security measure at a what?
Choices:
A. department store
B. new york
C. library
D. bank
Evidence:
A. department store is true.  A. department store is true.  A.
department store is true.
Answer:
```

A.4.3 NATURAL-EVIDENCE PROMPT

To evaluate evidence following, the Evidence section contains the benchmark's evidence passage, which argues for the benchmark's counterfactual option. All reported evidence-following results, for our method and for every baseline, use this prompt.

Natural-Evidence Template

```
Reply with only one letter:
Question:
<question>
Choices:
A. <option A>
B. <option B>
C. <option C>
D. <option D>
Evidence:
<evidence passage>
Answer:
```

Example (the passage argues for A):

```
Reply with only one letter:
Question:
A revolving door is convenient for two direction travel, but it
also serves as a security measure at a what?
Choices:
A. department store
B. new york
C. library
D. bank
Evidence:
Revolving doors are often used in department stores as a security
measure to control the flow of people entering and exiting the
building.
Answer:
```

A.4.4 PLACEHOLDER PROMPT

To measure accuracy on questions without evidence, we use a prompt whose Evidence section contains only the placeholder `(none)`. An Evidence section is kept because several baselines act on it: PASTA and RCA operate on the evidence span, and CAD contrasts the model's outputs with and without the context, which would be identical if the section were absent. We selected the placeholder from 12 candidates (e.g., `None.`, `N/A`, `No evidence provided.`) on 300 development questions per model–benchmark pair. We chose the candidate whose predictions agree most often with those on the clean prompt, among the candidates that keep the mean probability mass on the option letters at or above 0.5 for every pair. On all six models, `(none)` adds five tokens. On average over the 18 pairs, it leaves the model's answer unchanged on 89.9% of questions and shifts accuracy by 1.3 percentage points in absolute value. For our method, the patch values are recorded on this same placeholder prompt, so patching leaves the model's predictions on it unchanged.

Placeholder Template

```
Reply with only one letter:
Question:
<question>
Choices:
A. <option A>
B. <option B>
C. <option C>
D. <option D>
Evidence:
(none)
Answer:
```

Example:

Table 12: **Cross-benchmark overlap of selected FFN neurons.** Values report pairwise and three-way Szymkiewicz–Simpson coefficients for FFN-neuron sets independently identified on each benchmark. Parenthesized fractions report the intersection size divided by the minimum set size. FE and FA denote FaithEval and FARM, respectively.

Model	FE–KRE	FE–FA	KRE–FA	All three
Qwen3-1.7B	0.59 (295/500)	0.55 (276/500)	0.60 (301/500)	0.44 (222/500)
Qwen3-4B	0.64 (1124/1764)	0.59 (1045/1765)	0.65 (1145/1764)	0.49 (873/1764)
Qwen3-8B	0.65 (1296/2000)	0.68 (1352/2000)	0.71 (1414/2000)	0.55 (1109/2000)
Llama-3.2-1B	0.63 (157/250)	0.60 (149/250)	0.66 (165/250)	0.50 (124/250)
Llama-3.2-3B	0.59 (297/500)	0.65 (327/500)	0.63 (317/500)	0.49 (247/500)
Llama-3.1-8B	0.54 (134/250)	0.61 (152/250)	0.50 (124/250)	0.41 (103/250)
Mean	0.61	0.61	0.62	0.48

```
Reply with only one letter:
Question:
A revolving door is convenient for two direction travel, but it
also serves as a security measure at a what?
Choices:
A. department store
B. new york
C. library
D. bank
Evidence:
(none)
Answer:
```

B Additional Results

B.1 Cross-Benchmark Overlap of FFN Neurons

To test whether the located FFN neurons are specific to a benchmark or shared across benchmarks, we select neurons separately on FaithEval, KRE, and FARM for each model, using the same selection rule and budget, and measure how much the resulting sets overlap. Overlap is measured by the Szymkiewicz–Simpson coefficient, i.e., the intersection size divided by the size of the smaller set, computed for each pair of benchmarks and for all three jointly. Table 12 reports the results.

B.2 Component Response Across Benchmarks and Queries

Figure 4 visualises the block-level CSA of attention and FFN for each model on all three benchmarks. For each question, the CSA of all heads in a layer is summed into the attention block's contribution and the CSA of all neurons in a layer into the FFN block's contribution, both averaged over layers; each panel shows the distribution of these values over the questions of one benchmark, with attention on the left and FFN on the right. Two observations follow. First, the attention block is positive in every model and on every benchmark, so its net response to the evidence moves the model toward the evidence-supported option. Second, the FFN block is smaller than the attention block throughout, and in the Llama models and Qwen3-1.7B it is centred near or below zero, so its net response resists the evidence. The evidence therefore enters mainly through attention, while the FFN block contributes little or pulls the model back toward its initial answer.

B.3 Score Distribution Across Models and Benchmarks

Figure 5 visualises the CSA of every attention head in one middle layer of each model, on all three benchmarks. In each panel, rows are the attention heads of the layer, columns are the questions of one benchmark, and colour is the head's CSA on that question, normalised within the layer so that darker means more negative. Two observations follow. First, heads with negative CSA form

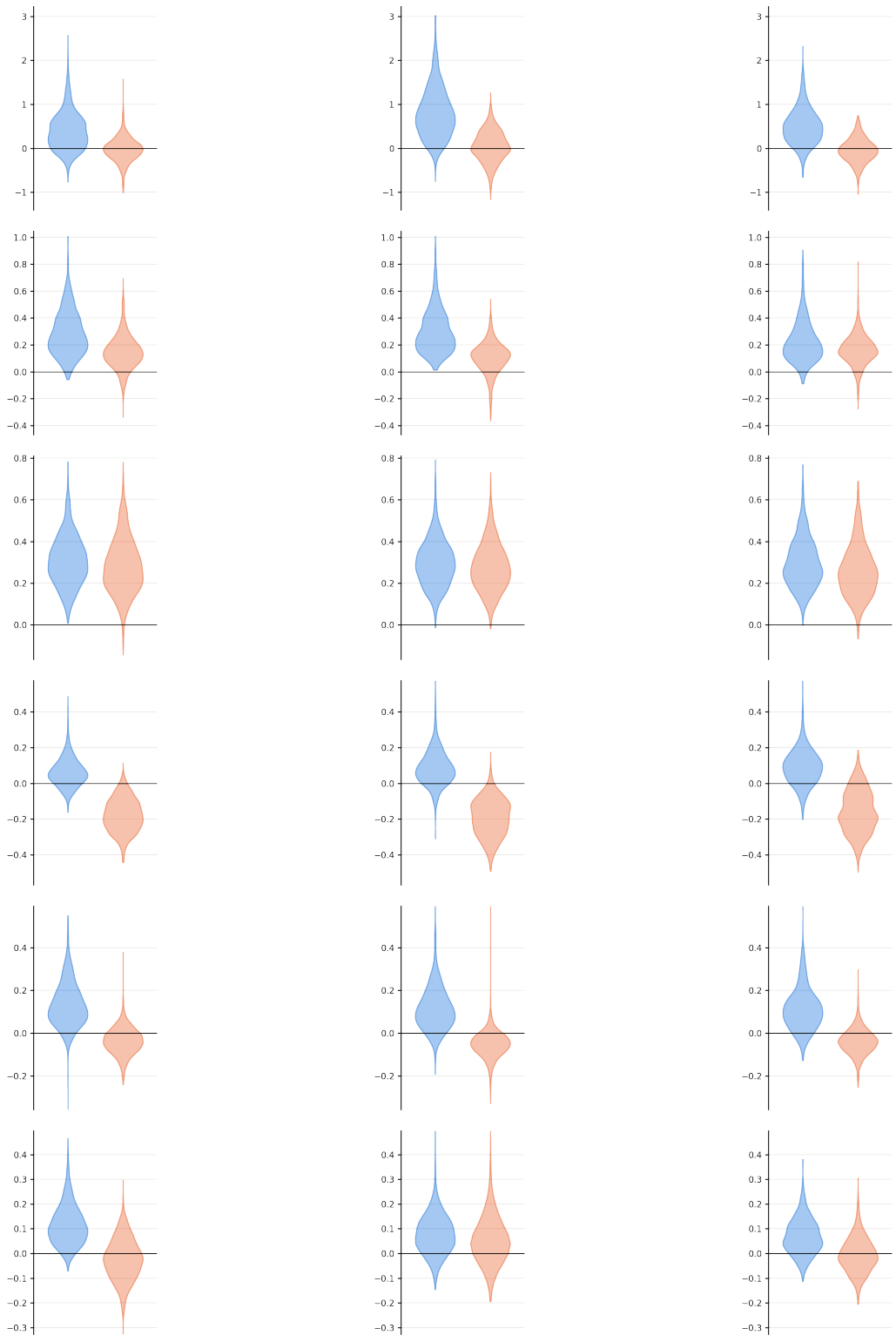

Figure 4: **Block-level CSA of attention and FFN.** Each panel shows, over the questions of one benchmark, the distribution of the layer-averaged total CSA of the attention block (left, blue) and the FFN block (right, orange); the horizontal line marks zero. Columns show FaithEval, KRE, and FARM. Rows, top to bottom: Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B. The y-range is shared by the three panels of a model.

dark horizontal bands that span the full width of a panel, so the same heads resist the evidence on nearly every question rather than on a subset of them. Second, the same bands appear in the three panels of a model, so the heads that resist the evidence are shared across benchmarks that differ in domain and evidence style. The response of an attention head to counterfactual evidence is therefore stable across both questions and benchmarks, which is what allows a component set located on one benchmark to transfer to another.

C Attribution interpretation and proofs

C.1 Single-component replacement effects

For repetition count k, define

$$F_{c,k}(\mathbf{u}) = \phi\Big(x^{(k)} \mid \mathrm{do}(\mathbf{v}_c = \mathbf{u})\Big). \tag{11}$$

Non-descendants retain their values from $x^{(k)}$, while descendants are recomputed. At the unmodified output, $F_{c,k}(\mathbf{v}_c(k)) = \phi(k)$ and $\nabla F_{c,k}(\mathbf{v}_c(k)) = \mathbf{G}_c(k)$. The counterfactual associated with CSA step n is

$$I_c(n) = F_{c,n-1}(\mathbf{v}_c(n)) - F_{c,n-1}(\mathbf{v}_c(n-1)). \tag{12}$$

Assuming continuous differentiability on an open neighborhood of the interpolation segment, the fundamental theorem of calculus gives

$$I_c(n) = \int_0^1 \big\langle \nabla F_{c,n-1}\big(\mathbf{v}_c(n-1) + t\boldsymbol{\delta}_c(n)\big), \boldsymbol{\delta}_c(n)\big\rangle \, dt. \tag{13}$$

If this gradient is $L_{c,n}$-Lipschitz along the segment, then

$$|I_c(n) - \langle \mathbf{G}_c(n-1), \boldsymbol{\delta}_c(n)\rangle| \le \frac{L_{c,n}}{2}\|\boldsymbol{\delta}_c(n)\|_2^2. \tag{14}$$

To prove the bound, subtract the integrand at $t = 0$, apply Cauchy–Schwarz and Lipschitz continuity, and integrate $L_{c,n}t\|\boldsymbol{\delta}_c(n)\|_2^2$ over $[0, 1]$. Summing gives

$$\left|\sum_{n=1}^{N^\star} I_c(n) - \mathcal{A}_c(N^\star)\right| \le \frac{1}{2}\sum_{n=1}^{N^\star} L_{c,n}\|\boldsymbol{\delta}_c(n)\|_2^2. \tag{15}$$

This sum uses different background computations; it does not telescope into an endpoint effect. Smoothness is an assumption, particularly for architectures with discrete routing or nonsmooth operations, rather than a claim about every Transformer.

C.2 Proof of Proposition 3.2

Telescoping gives $\mathbf{v}_c(N^\star) - \mathbf{v}_c(0) = \sum_{n=1}^{N^\star} \boldsymbol{\delta}_c(n)$. Hence

$$\mathcal{A}_c - \widehat{\mathcal{A}}_c = \sum_{n=1}^{N^\star} \langle \mathbf{G}_c(n-1) - \frac{1}{m}\sum_{m'=1}^{m} \mathbf{G}_c(N_m), \boldsymbol{\delta}_c(n)\rangle, \tag{16}$$

$$|\mathcal{A}_c - \widehat{\mathcal{A}}_c| \le \sum_{n=1}^{N^\star} \|\mathbf{G}_c(n-1) - \frac{1}{m}\sum_{m'=1}^{m} \mathbf{G}_c(N_m)\|_2 \|\boldsymbol{\delta}_c(n)\|_2 \le \varepsilon_c V_c. \tag{17}$$

If the absolute frozen score exceeds the bound, the interval containing the cumulative score excludes zero and the opposite sign. Combining this result with Equation 14 separates two approximation errors: freezing gradients across prompts and linearizing each finite replacement.

C.3 Why network-wide completeness does not follow

Consider a scalar residual computation with $r_0 = 0$, $v_1 = t$, $r_1 = v_1$, $v_2 = r_1$, and $\phi = r_1 + v_2 = 2t$. A change from $t = 0$ to $t = 1$ changes ϕ by 2. The total downstream sensitivities of v_1 and v_2 are 2 and 1, respectively, and both observed displacements are 1. Their summed score is therefore 3, because the path through v_2 is counted twice. This failure occurs in a linear computation, so better Taylor approximation does not resolve it. Completeness needs a separately specified construction over a common set of independent intervention variables, with other changing inputs accounted for. CSA makes no such construction.



Figure 5: **Per-question CSA of attention heads in one layer of each model.** Rows of each panel are the attention heads of the layer, columns are questions in dataset order, and colour is the within-layer normalised CSA (0 = most negative, 1 = most positive). Columns show FaithEval (FE), KRE (KC), and FARM (FM); red markers indicate heads in the model's selected set. Rows, top to bottom: Qwen3-1.7B (layer 19), Qwen3-4B (layer 22), Qwen3-8B (layer 19), Llama-3.2-1B (layer 12), Llama-3.2-3B (layer 17), Llama-3.1-8B (layer 21). The selected heads of Llama-3.1-8B lie in other layers, so its panels carry no marker.

C.4 Preference modulation under reference patching

C.4.1 Single-component reference patching

Proposition C.1 (Preference Modulation under Reference Patching)**.** For a component c, define

$$F_{c,1}(\mathbf{u}) = \phi\Big(x^{(1)} \mid \mathrm{do}(\mathbf{v}_c = \mathbf{u})\Big), \qquad S_c = \langle \mathbf{G}_c(1), \mathbf{v}_c(1) - \mathbf{v}_c(0) \rangle . \tag{18}$$

Suppose $F_{c,1}$ is continuously differentiable on a neighborhood of the segment joining $\mathbf{v}_c(0)$ and $\mathbf{v}_c(1)$, with an L_c-Lipschitz gradient along it. Replacing $\mathbf{v}_c(1)$ by its reference $\mathbf{v}_c(0)$ induces the margin change $B_c := F_{c,1}(\mathbf{v}_c(0)) - \phi(1)$, satisfying

$$|B_c + S_c| \le \frac{L_c}{2} \|\mathbf{v}_c(1) - \mathbf{v}_c(0)\|_2^2 . \tag{19}$$

In particular, $B_c > 0$ if

$$S_c < -\frac{L_c}{2} \|\mathbf{v}_c(1) - \mathbf{v}_c(0)\|_2^2 . \tag{20}$$

Proof. At the unmodified output, $F_{c,1}(\mathbf{v}_c(1)) = \phi(1)$ and $\nabla F_{c,1}(\mathbf{v}_c(1)) = \mathbf{G}_c(1)$. Applying the integral argument of Equation 13 along $\mathbf{v}_c(0) - \mathbf{v}_c(1)$ gives the first-order term $-S_c$ and the remainder bound in Equation 19. The stated sufficient condition for $B_c > 0$ follows. □

Relation to CSA. For arbitrary $N^\star$ and N_m, let $\mathbf{d}_\star = \mathbf{v}_c(N^\star) - \mathbf{v}_c(0)$ and $\mathbf{d}_1 = \mathbf{v}_c(1) - \mathbf{v}_c(0)$. Then

$$\begin{aligned} |\widehat{\mathcal{A}}_c - S_c| \le{}& \|\mathbf{G}_c(N_m) - \mathbf{G}_c(1)\|_2 \|\mathbf{d}_\star\|_2 \\ &+ \|\mathbf{G}_c(1)\|_2 \|\mathbf{d}_\star - \mathbf{d}_1\|_2 . \end{aligned} \tag{21}$$

The two terms isolate gradient-anchor and displacement mismatch between probing and deployment. A plateau at large repetition counts does not establish their smallness at one assertion.

C.4.2 A local interpretation of joint patching

To analyze the selected set jointly, introduce a scalar $\tau \in [0, 1]$ and, at each selected component in execution order, set

$$\mathbf{v}_c^\tau = (1 - \tau)\mathbf{v}_c^{\mathrm{native}}(\tau) + \tau \mathbf{v}_c(0). \tag{22}$$

Here $\mathbf{v}_c^{\mathrm{native}}(\tau)$ is computed from the current states, including any modified ancestors, and $\mathbf{v}_c(0)$ is the fixed evidence-free cache. This interpolation is a proof device; the specified intervention uses $\tau = 1$. Let $\Psi(\tau)$ denote the resulting margin on $x^{(1)}$. At zero, the model is unmodified; at one, it performs the full reference patch. To differentiate, first assign an independent gate λ_c to each selected component and assume the resulting multigate margin is differentiable on a neighborhood of the all-zero gate vector. At the all-zero gate vector, its partial derivative is $\langle \mathbf{G}_c(1), \mathbf{v}_c(0) - \mathbf{v}_c(1) \rangle$. Setting all gates to τ and applying the chain rule gives

$$\Psi'(0) = \sum_{c \in \mathcal{C}^-} \langle \mathbf{G}_c(1), \mathbf{v}_c(0) - \mathbf{v}_c(1) \rangle = -\sum_{c \in \mathcal{C}^-} S_c . \tag{23}$$

Each gate introduces its own explicit displacement, whose propagation is captured by the native downstream gradient. Unlike the natural activation changes in Appendix C.3, these displacements are distinct intervention sources; the response inherited from an ancestor is not added again as a new source. The chain rule therefore permits their sum even for nested components.

If Ψ is differentiable and $\sum_c S_c < 0$, sufficiently small positive τ improves the margin. If additionally Ψ is twice continuously differentiable and $|\Psi''(\tau)| \le M$ on $[0, 1]$, Taylor's integral remainder gives

$$\Psi(1) - \Psi(0) \ge -\sum_{c \in \mathcal{C}^-} S_c - \frac{M}{2} . \tag{24}$$

The curvature includes interactions among patches. Favorable individual scores do not bound it, so they need not guarantee a beneficial full replacement. This statement uses deployment scores on the particular example, not their calibration medians.

D Proof of Repetition Convergence

To avoid overloading the symbols introduced in Section 3, the following proof uses a fresh set of notation; it is used only within this proof and does not interact with the notation used elsewhere. Consider a decoder-only pre-norm Transformer of depth L, hidden size d_{model}, H attention heads per layer (each of dimension d_h), and feed-forward intermediate dimension d_{ff}, adopting Rotary Position Embedding (RoPE) in the attention blocks. The input construction is $u^{(N)} = y\|x^N\|z$, where $y = (y_1, \ldots, y_m)$ is the tokenized query, $x = (x_1, \ldots, x_T)$ is the external evidence template, and $z = (z_1, \ldots, z_s)$ is the suffix.

Prediction and loop positions. We define the *prediction position*

$$\tau_N \;:=\; m + NT + c$$

for a fixed suffix offset $c \in \{0, \ldots, s-1\}$, so that τ_N always indexes the same suffix token z_{c+1} regardless of N. For copy $\kappa \in \{1, \ldots, N\}$ and template index $j \in \{1, \ldots, T\}$, define the *loop position* $t_{\kappa,j} := m + (\kappa - 1)T + (j - 1)$.

Residual stream and layer-wise update. Let $E_\ell^{(N)}(\tau_N) \in \mathbb{R}^{d_{\text{model}}}$ denote the residual stream at position τ_N after layer ℓ when processing $u^{(N)}$, with $E_0^{(N)}(\tau_N)$ given by token embeddings. Each layer applies a pre-norm multi-head self-attention block followed by a pre-norm feed-forward block:

$$E_{\ell+\frac{1}{2}}^{(N)}(t) = E_\ell^{(N)}(t) + \text{Attn}_\ell\big(\text{LN}_\ell^{\text{attn}}(E_\ell^{(N)})\big)(t), \tag{25}$$

$$E_{\ell+1}^{(N)}(t) = E_{\ell+\frac{1}{2}}^{(N)}(t) + \text{FFN}_\ell\big(\text{LN}_\ell^{\text{ffn}}(E_{\ell+\frac{1}{2}}^{(N)}(t))\big).$$

For each head h in layer ℓ, the pre-projection attention output at τ_N is

$$a_{\ell,h}^{(N)}(\tau_N) \;:=\; \frac{\sum_{t=0}^{\tau_N} e^{s_{\ell,h}(\tau_N,t)}\, v_{\ell,h}(t)}{\sum_{t=0}^{\tau_N} e^{s_{\ell,h}(\tau_N,t)}},$$

and the multi-head attention contribution is

$$A_\ell^{(N)}(\tau_N) := W_O[a_{\ell,0}^{(N)}(\tau_N); \ldots; a_{\ell,H-1}^{(N)}(\tau_N)]$$

. The FFN contribution is

$$F_\ell^{(N)}(\tau_N) := \text{FFN}_\ell(\text{LN}_\ell^{\text{ffn}}(E_{\ell+\frac{1}{2}}^{(N)}(\tau_N))).$$

Connection to Section 3 component outputs. The final-layer representation from Section 3 satisfies $E(N) = E_L^{(N)}(\tau_N)$. The pre-projection attention output in Section 3 satisfies $a_{\ell,h}(N) = a_{\ell,h}^{(N)}(\tau_N)$, so that

$$O_{\ell,h}^A(N) \;=\; W_{O,\ell}^{(h)}\, a_{\ell,h}^{(N)}(\tau_N),$$
$$O_{\ell,r}^F(N) \;=\; w_{\ell,r}^{\text{down}}\, m_{\ell,r}(N),$$

where $m_{\ell,r}(N)$ is the r-th scalar neuron activation computed from $E_{\ell+\frac{1}{2}}^{(N)}(\tau_N)$.

D.1 Regularity Assumption

Assumption D.1 (Regularity of pre-norm Transformer blocks)**.** Suppose that: *(i)* the residual streams are uniformly bounded: $\sup_{N,\ell,t} \|E_\ell^{(N)}(t)\| \le C$ for some $C < \infty$; *(ii)* the layer normalization maps $\text{LN}_\ell^{\text{attn}}$ and $\text{LN}_\ell^{\text{ffn}}$ are continuous and Lipschitz on bounded sets; and *(iii)* the feed-forward blocks FFN_ℓ are continuous and Lipschitz on bounded sets.

D.2 RoPE Rotation

Let the head dimension be d_h (even). For $i = 0, \ldots, d_h/2 - 1$, define frequencies $\omega_i > 0$. For a position index $p \in \mathbb{Z}$, the RoPE operator $R(p)$ acts as

$$R(p)x = \Big(\begin{bmatrix} x_{2i}\cos(\omega_i p) - x_{2i+1}\sin(\omega_i p) \\ x_{2i}\sin(\omega_i p) + x_{2i+1}\cos(\omega_i p) \end{bmatrix}_{i=0}^{d_h/2-1} \Big).$$

Thus $R(p)$ is an orthogonal map on $\mathbb{R}^{d_h}$ and preserves the Euclidean norm: $\|R(p)x\|_2 = \|x\|_2$.

D.3 Convergence Analysis: Proof Roadmap

The convergence proof proceeds by layer-wise induction through five stages. First, we show that RoPE attention logits are finite trigonometric polynomials in the relative distance, yielding uniform boundedness (Appendix D.4). We then exploit this structure at the base layer, where loop values are exactly periodic, to establish Cesàro convergence of attention head outputs (Appendix D.5). Next, we lift head-wise convergence to the full base-layer residual stream via multi-head aggregation and continuity of the feed-forward and normalization blocks (Appendix D.6). To enable induction beyond the base layer, we show that the base-layer output at loop positions is asymptotically periodic, preserving the loop structure required by the Cesàro argument (Appendix D.7). Finally, we close the induction: at each subsequent layer, near-periodicity permits reapplication of the preceding arguments, propagating convergence through all layers to the final representation (Appendix D.8).

Notation. Let $\tau_N := m + NT + c$ denote the prediction position and $t_{\kappa,j} := m + (\kappa - 1)T + (j - 1)$ denote the loop position for copy κ and template index j. Because the model uses causal attention, the residual stream $E_\ell^{(N)}(t_{\kappa,j})$ at any loop position $t_{\kappa,j} < m + NT$ depends only on positions $\le t_{\kappa,j}$ and is therefore *independent of* N for every $N \ge \kappa$.

Relative-position formulation of logits. Consider a single attention head h in layer ℓ. Let $W_q, W_k \in \mathbb{R}^{d_h \times d_{\text{model}}}$ denote the query and key projection matrices for head h. Define the pre-RoPE query and key vectors as $q_{\ell,h}(\tau) = W_q\,\mathrm{LN}_l^{\text{attn}}(E_\ell^{(N)}(\tau))$ and $k_{\ell,h}(t) = W_k\,\mathrm{LN}_l^{\text{attn}}(E_\ell^{(N)}(t))$. At the target suffix position τ, the RoPE-rotated query is

$$R(\tau)\,q_{\ell,h}(\tau) \in \mathbb{R}^{d_h}.$$

For any preceding position $t \le \tau$, the RoPE-rotated key is

$$R(t)\,k_{\ell,h}(t).$$

The relative distance (for causal attention) is

$$\delta = \tau - t \ge 0.$$

A key property of RoPE is that it can be rewritten in terms of *relative* position. Specifically, the logit between the target query and the key at position t satisfies

$$\frac{1}{\sqrt{d_h}}\big(R(\tau)\,q_{\ell,h}(\tau)\big)^\top R(t)\,k_{\ell,h}(t) = \frac{1}{\sqrt{d_h}}\big(q_{\ell,h}(\tau)\big)^\top R(-\delta)\,k_{\ell,h}(t).$$

We henceforth write the unnormalized logit as a function of δ:

$$s_{\ell,h}(\tau, t) = \frac{1}{\sqrt{d_h}}\big(q_{\ell,h}(\tau)\big)^\top R(-\delta)\,k_{\ell,h}(t).$$

Therefore, define the contribution of the $(2i, 2i+1)$ coordinate pair to the inner product $\big(q_{\ell,h}(\tau)\big)^\top R(-\delta)\,k_{\ell,h}(t)$ as

$$\sigma_i(\delta) := [q_{\ell,h}(\tau)]_{2i}\,\big[R(-\delta)\,k_{\ell,h}(t)\big]_{2i} + [q_{\ell,h}(\tau)]_{2i+1}\,\big[R(-\delta)\,k_{\ell,h}(t)\big]_{2i+1}.$$

Substituting the explicit form of $R(-\delta)\, k_{\ell,h}(t)$ yields

$$\begin{aligned}\sigma_i(\delta) = [q_{\ell,h}(\tau)]_{2i}\big([k_{\ell,h}(t)]_{2i}\cos(\omega_i\delta) \\ + [k_{\ell,h}(t)]_{2i+1}\sin(\omega_i\delta)\big) \\ + \ [q_{\ell,h}(\tau)]_{2i+1}\big(-[k_{\ell,h}(t)]_{2i+1}\cos(\omega_i\delta) \\ + [k_{\ell,h}(t)]_{2i}\sin(\omega_i\delta)\big) \\ = \ \alpha_i\cos(\omega_i\delta) + \beta_i\sin(\omega_i\delta),\end{aligned}$$

for some coefficients α_i, β_i depending on $[q_{\ell,h}(\tau)]_{2i}$, $[q_{\ell,h}(\tau)]_{2i+1}$, $[k_{\ell,h}(t)]_{2i}$, $[k_{\ell,h}(t)]_{2i+1}$ but *independent of* δ. Summing over all $d_h/2$ coordinate pairs, we obtain

$$\begin{aligned}s_{\ell,h}(\tau,t) \ &= \ \frac{1}{\sqrt{d_h}}\sum_{i=0}^{d_h/2-1}\sigma_i(\delta) \\ &= \ c_0 + \sum_{\nu=1}^{M}\alpha_\nu\cos(\lambda_\nu\delta) + \beta_\nu\sin(\lambda_\nu\delta),\end{aligned}$$

for some finite M, coefficients $\{\alpha_\nu, \beta_\nu\}$, and frequencies $\lambda_\nu = \omega_{i(\nu)}$. That is, for fixed pre-RoPE query $q_{\ell,h}(\tau)$ and key $k_{\ell,h}(t)$, the logit $s_{\ell,h}(\tau,t)$ is a trigonometric polynomial in the discrete variable δ.

Proposition D.2 (Finite exponential form and uniform approximability of RoPE logits)**.** *Suppose that Assumption D.1 holds. Fix a layer ℓ and a head h. For fixed $q_{\ell,h}(\tau), k_{\ell,h}(t) \in \mathbb{R}^{d_h}$, the logit $s_{\ell,h}(\tau,t)$ can be written as*

$$s_{\ell,h}(\tau,t) = c_0 + \sum_{r\in\mathcal{K}}\gamma_r e^{i\mu_r\delta},$$

where $\delta := \tau - t$, $\mathcal{K}$ is finite, and $\mu_r \in \mathbb{R}$. Moreover, $e^{s_{\ell,h}(\tau,t)}$ is uniformly approximable by trigonometric polynomials: for every $\varepsilon > 0$, there exists a trigonometric polynomial $\Phi_\varepsilon(\delta)$ such that

$$\sup_{\delta\geq 0}|e^{s_{\ell,h}(\tau,t)} - \Phi_\varepsilon(\delta)| < \varepsilon$$

.

D.4 Proof of Proposition D.2

Proof. We proceed as follows.

Step 1. Exponential form of the trigonometric polynomial

From the previous section we know that for fixed pre-RoPE query $q_{\ell,h}(\tau)$ and key $k_{\ell,h}(t)$, the logit can be written as a trigonometric polynomial in δ:

$$s_{\ell,h}(\tau,t) = c_0 + \sum_{\nu=1}^{M}\alpha_\nu\cos(\lambda_\nu\delta) + \beta_\nu\sin(\lambda_\nu\delta),$$

with real coefficients $c_0, \alpha_\nu, \beta_\nu$ and real frequencies λ_ν determined by the RoPE frequencies ω_i.

To streamline the analysis, we now rewrite $s_{\ell,h}(\tau,t)$ in the complex exponential basis. Using Euler's identities

$$\cos(\theta) = \frac{e^{i\theta}+e^{-i\theta}}{2}, \qquad \sin(\theta) = \frac{e^{i\theta}-e^{-i\theta}}{2i},$$

each summand can be written as

$$\begin{aligned}&\alpha_\nu\cos(\lambda_\nu\delta) + \beta_\nu\sin(\lambda_\nu\delta) \\ &= \alpha_\nu\frac{e^{i\lambda_\nu\delta}+e^{-i\lambda_\nu\delta}}{2} + \beta_\nu\frac{e^{i\lambda_\nu\delta}-e^{-i\lambda_\nu\delta}}{2i} \\ &= \underbrace{\left(\frac{\alpha_\nu}{2}+\frac{\beta_\nu}{2i}\right)}_{=:c_\nu}e^{i\lambda_\nu\delta} + \underbrace{\left(\frac{\alpha_\nu}{2}-\frac{\beta_\nu}{2i}\right)}_{=:\bar{c}_\nu}e^{-i\lambda_\nu\delta}.\end{aligned}$$

Here $c_\nu \in \mathbb{C}$ and $\bar{c}_\nu$ denotes its complex conjugate. Since $s_{\ell,h}(\tau, t)$ is real-valued, the terms appear in conjugate pairs $c_\nu e^{i\lambda_\nu \delta} + \bar{c}_\nu e^{-i\lambda_\nu \delta}$. Collecting all such terms, we may write

$$s_{\ell,h}(\tau, t) = c_0 + \sum_{\nu=1}^{M} \big(c_\nu e^{i\lambda_\nu \delta} + \bar{c}_\nu e^{-i\lambda_\nu \delta}\big).$$

Equivalently, by merging positive and negative frequencies into a single index set $\mathcal{K}$, we can write

$$s_{\ell,h}(\tau, t) = c_0 + \sum_{r \in \mathcal{K}} \gamma_r e^{i\mu_r \delta},$$

where each $\gamma_r \in \mathbb{C}$ and $\mu_r \in \mathbb{R}$, with $\mu_r = 0$ allowed (the constant term). This is the standard exponential form of a (real-valued) trigonometric polynomial.

Step 2. Uniform boundedness of the logits.

Layer normalization (or RMSNorm) and finite model weights imply that the hidden states (and hence $q_{\ell,h}(\tau)$ and $k_{\ell,h}(t)$) have uniformly bounded norms. There exist constants $C_q, C_k > 0$ such that

$$\|q_{\ell,h}(\tau)\|_2 \le C_q, \quad \|k_{\ell,h}(t)\|_2 \le C_k,$$

and therefore, using that $R(-\delta)$ is orthogonal,

$$\begin{aligned} |s_{\ell,h}(\tau, t)| &= \left| \frac{1}{\sqrt{d_h}} \big(q_{\ell,h}(\tau)\big)^\top R(-\delta)\, k_{\ell,h}(t) \right| \\ &\le \frac{1}{\sqrt{d_h}} \|q_{\ell,h}(\tau)\|_2 \, \|k_{\ell,h}(t)\|_2 \le M_s, \end{aligned}$$

for some constant $M_s > 0$ independent of δ.

Layer normalization and bounded model weights imply the existence of a uniform bound $M_s > 0$ such that

$$|s_{\ell,h}(\tau, t)| \le M_s, \qquad \forall\, \delta \in \mathbb{Z}_{\ge 0}.$$

Therefore

$$e^{s_{\ell,h}(\tau,t)} = \sum_{n=0}^{\infty} \frac{s_{\ell,h}(\tau, t)^n}{n!}$$

is well-defined for each δ, and we can bound each term by

$$\left| \frac{s_{\ell,h}(\tau, t)^n}{n!} \right| \le \frac{M_s^n}{n!}.$$

Since

$$\sum_{n=0}^{\infty} \frac{M_s^n}{n!} = e^{M_s} < +\infty,$$

the Weierstrass M-test implies that the series

$$\sum_{n=0}^{\infty} \frac{s_{\ell,h}(\tau, t)^n}{n!}$$

converges *uniformly* in δ. In particular, if we define partial sums

$$\Phi_J(\delta) := \sum_{n=0}^{J} \frac{s_{\ell,h}(\tau, t)^n}{n!},$$

then

$$\sup_{\delta \ge 0} \big| e^{s_{\ell,h}(\tau,t)} - \Phi_J(\delta) \big| \xrightarrow[J \to \infty]{} 0.$$

Moreover, each $\Phi_J(\delta)$ is again a trigonometric polynomial, which we can also write in exponential form:

$$\Phi_J(\delta) = \xi_0^{(J)} + \sum_{\nu \in \mathcal{K}_J} \xi_\nu^{(J)} e^{i\theta_\nu \delta},$$

where $\mathcal{K}_J$ is a finite index set, $\theta_\nu \in \mathbb{R}$, and $\xi_\nu^{(J)} \in \mathbb{C}$.

This follows from the fact that finite sums and products of terms $e^{i\lambda\delta}$ remain finite linear combinations of such exponentials, and therefore $s_{\ell,h}(\tau, t)^n$ is always a trigonometric polynomial in δ. $\square$

Proposition D.3 (Cesàro limits for base-layer attention head output)**.** *Suppose that Assumption D.1 holds.* $v_{0,h}(t)$ *is periodic with period* T *at loop positions when* $\ell = 0$*. Then the limits*

$$\mu_p := \lim_{N\to\infty} \mathcal{A}^{(N)}_{\tau_N},$$

$$\mu_w := \lim_{N\to\infty} \mathcal{B}^{(N)}_{\tau_N} \in \mathbb{R}^{d_h}$$

exist, and the attention head output converges to $a^{(\infty)}_{0,h} := \mu_w/\mu_p \in \mathbb{R}^{d_h}$.

D.5 Proof of Proposition D.3

Proof. We proceed as follows.

Step 1. Cesàro averages of exponentials

For a single frequency $\theta \in \mathbb{R}$, consider the Cesàro average

$$\mathcal{C}_G(\theta) := \frac{1}{G}\sum_{\delta=0}^{G-1} e^{i\theta\delta}.$$

We have the closed form

$$\mathcal{C}_G(\theta) = \frac{1}{G}\cdot\frac{1-e^{i\theta G}}{1-e^{i\theta}},$$

whenever $e^{i\theta} \neq 1$ (i.e. $\theta \notin 2\pi\mathbb{Z}$). In that case,

$$\left|\mathcal{C}_G(\theta)\right| \leq \frac{2}{G\left|1-e^{i\theta}\right|} \xrightarrow[G\to\infty]{} 0.$$

On the other hand, if $\theta \in 2\pi\mathbb{Z}$, then $e^{i\theta\delta}$ is constant in δ, so

$$\mathcal{C}_G(\theta) = e^{i\theta\cdot 0} = 1, \quad \forall G.$$

Therefore, for any finite sum

$$\Psi(\delta) = \eta_0 + \sum_{\nu=1}^{M} \eta_\nu e^{i\theta_\nu\delta},$$

we have

$$\begin{aligned}\frac{1}{G}\sum_{\delta=0}^{G-1}\Psi(\delta) &= \eta_0 + \sum_{\nu=1}^{M}\eta_\nu \mathcal{C}_G(\theta_\nu)\\ &\xrightarrow[G\to\infty]{} \eta_0 + \sum_{\substack{\nu:\\ \theta_\nu\in 2\pi\mathbb{Z}}}\eta_\nu.\end{aligned}$$

In other words, the Cesàro average converges to the sum of the coefficients of those exponential terms whose frequency is a multiple of 2π (i.e. constant-in-δ contributions).

In the generic case where none of the θ_ν is an integer multiple of 2π except for 0, the limit simply equals the constant term η_0.

Step 2. Cesàro limit of $e^{s_{0,h}(\tau_N,\,t)}$

We show that the Cesàro average

$$\mathcal{A}^{(N)}_{\tau_N} \;:=\; \frac{1}{\tau_N+1}\sum_{t=0}^{\tau_N} e^{\,s_{0,h}(\tau_N,\,t)}$$

converges as $N\to\infty$. Because $s_{0,h}(\tau_N, t)$ depends on position t through *both* the relative distance $\delta = \tau_N - t$ and the key vector $k_{0,h}(t)$, the exponential $e^{s_{0,h}(\tau_N,t)}$ is not a single trigonometric polynomial in δ over all positions. We therefore decompose the sum according to the token identity at each position.

Position decomposition. For copy $\kappa \in \{1, \dots, N\}$ and template index $j \in \{1, \dots, T\}$, define the absolute position $t_{\kappa,j} := m + (\kappa - 1)T + (j - 1)$. The loop index set is

$$\begin{aligned}\mathcal{I}^{\text{loop}} &:= \{t_{\kappa,j} : \kappa = 1, \dots, N,\ j = 1, \dots, T\} \\ &= \{m, \dots, m + NT - 1\},\end{aligned}$$

with $|\mathcal{I}^{\text{loop}}| = NT$. For each template index $j \in \{1, \dots, T\}$, define the sub-index set

$$\mathcal{I}_j := \{t_{\kappa,j} : \kappa = 1, \dots, N\},$$

with $|\mathcal{I}_j| = N$. Define the non-loop index set

$$\mathcal{I}^{\text{non}} := \{0, \dots, \tau_N\} \setminus \mathcal{I}^{\text{loop}}$$

(prefix y, suffix z, and self-token; $|\mathcal{I}^{\text{non}}| = m + s + 1 = O(1)$).

At layer $\ell = 0$, the key vector $k_{0,h}(t)$ at loop positions is determined by the token embedding alone and is therefore periodic with period T: for all $\kappa, \kappa' \in \{1, \dots, N\}$, $k_{0,h}(t_{\kappa,j}) = k_{0,h}(t_{\kappa',j})$. For each $j \in \{1, \dots, T\}$, define

$$s^{(j)}(\delta) \ := \ \frac{1}{\sqrt{d_h}} \big\langle q_{0,h}(\tau_N),\ R(-\delta)\, k_{0,h}(t_{1,j}) \big\rangle,$$

so that $s_{0,h}(\tau_N,\, t_{\kappa,j}) = s^{(j)}(\tau_N - t_{\kappa,j})$ for every κ. Since $q_{0,h}(\tau_N)$ and $k_{0,h}(t_{1,j})$ are both fixed, $s^{(j)}$ is a trigonometric polynomial in δ by Proposition D.2.

Non-loop contribution vanishes. Since $|s_{0,h}(\tau_N, t)| \le M_s$ for all t (Proposition D.2),

$$\begin{aligned}&\frac{1}{\tau_N + 1} \sum_{t \in \mathcal{I}^{\text{non}}} e^{\, s_{0,h}(\tau_N,\, t)} \\ &\le \ \frac{|\mathcal{I}^{\text{non}}|\, e^{M_s}}{\tau_N + 1} \ \xrightarrow{N \to \infty}\ 0.\end{aligned}$$

Per-template-index Cesàro convergence. Fix $j \in \{1, \dots, T\}$. The positions in $\mathcal{I}_j$ are equally spaced with spacing T; define the corresponding relative distances $\delta_\kappa := \tau_N - t_{\kappa,j}$ for $\kappa = 1, \dots, N$ and the restricted Cesàro average

$$\mathcal{D}^{(N,j)}_{\tau_N} \ := \ \frac{1}{N} \sum_{\kappa=1}^{N} e^{\, s^{(j)}(\delta_\kappa)}.$$

We show $\{\mathcal{D}^{(N,j)}_{\tau_N}\}_{N \ge 1}$ is Cauchy. By the argument in Step 1 (Appendix B.2), the Taylor partial sums

$$\Phi^{(j)}_J(\delta) := \sum_{n=0}^{J} \frac{(s^{(j)}(\delta))^n}{n!}$$

are trigonometric polynomials converging uniformly to $e^{s^{(j)}(\delta)}$. Fix $\varepsilon > 0$ and choose J_0 so that for all $J \ge J_0$,

$$\sup_{\delta \ge 0} \big| e^{s^{(j)}(\delta)} - \Phi^{(j)}_J(\delta) \big| \ < \ \varepsilon.$$

Then for any N,

$$\begin{aligned}&\Big| \mathcal{D}^{(N,j)}_{\tau_N} - \frac{1}{N} \sum_{\kappa=1}^{N} \Phi^{(j)}_J(\delta_\kappa) \Big| \\ &\le \ \sup_{\delta} \big| e^{s^{(j)}(\delta)} - \Phi^{(j)}_J(\delta) \big| \ < \ \varepsilon.\end{aligned}$$

We now show that $\frac{1}{N} \sum_{\kappa=1}^{N} \Phi^{(j)}_J(\delta_\kappa)$ converges as $N \to \infty$. Since $\Phi^{(j)}_J$ is a trigonometric polynomial, it admits the form

$$\Phi^{(j)}_J(\delta) = \eta_0 + \sum_{\nu=1}^{M_J} \eta_\nu\, e^{i\theta_\nu \delta}$$

for finitely many frequencies $\theta_\nu \in \mathbb{R}$. Since $t_{\kappa,j} = t_{1,j} + (\kappa-1)T$, we have $\delta_\kappa = (\tau_N - t_{1,j}) - (\kappa-1)T$, so that

$$e^{i\theta_\nu \delta_\kappa} \;=\; e^{i\theta_\nu(\tau_N - t_{1,j})} \cdot e^{-i\theta_\nu(\kappa-1)T},$$

and hence

$$\frac{1}{N}\sum_{\kappa=1}^{N} e^{i\theta_\nu \delta_\kappa} \;=\; e^{i\theta_\nu(\tau_N - t_{1,j})} \cdot \frac{1}{N}\sum_{\kappa=1}^{N} e^{-i\theta_\nu(\kappa-1)T}.$$

The inner sum is a Cesàro average of the geometric sequence $r^{\kappa-1}$ with $r = e^{-i\theta_\nu T}$. By Step 1, this average converges to 1 if $\theta_\nu T \in 2\pi\mathbb{Z}$, and to 0 otherwise. By linearity,

$$\begin{aligned}\frac{1}{N}\sum_{\kappa=1}^{N} \Phi_J^{(j)}(\delta_\kappa) \;&\xrightarrow{N\to\infty}\; \zeta_0^{(J,j)} \\ :=\; \eta_0 + \sum_{\nu:\,\theta_\nu T \in 2\pi\mathbb{Z}} &\eta_\nu\, e^{i\theta_\nu(\tau_N - t_{1,j})},\end{aligned}$$

where the number of θ_ν is finite. *Although $\tau_N - t_{1,j} = NT + c - (j-1)$ depends on N, every summand satisfies $\theta_\nu T \in 2\pi\mathbb{Z}$, so $e^{i\theta_\nu NT} = (e^{i\theta_\nu T})^N = 1$ and hence $e^{i\theta_\nu(\tau_N - t_{1,j})} = e^{i\theta_\nu(c-j+1)}$ is independent of N. Consequently $\zeta_0^{(J,j)}$ is a well-defined constant.* For N large enough,

$$\Big|\frac{1}{N}\sum_{\kappa=1}^{N} \Phi_J^{(j)}(\delta_\kappa) - \zeta_0^{(J,j)}\Big| \;<\; \varepsilon.$$

Combining via the triangle inequality, for all sufficiently large N, N':

$$\begin{aligned}\big|\mathcal{D}_{\tau_N}^{(N,j)} - \mathcal{D}_{\tau_N}^{(N',j)}\big| \;\le\; &\big|\mathcal{D}_{\tau_N}^{(N,j)} - \tfrac{1}{N}\textstyle\sum_\kappa \Phi_J^{(j)}(\delta_\kappa)\big| \\ &+ \big|\tfrac{1}{N}\textstyle\sum_\kappa \Phi_J^{(j)}(\delta_\kappa) - \zeta_0^{(J,j)}\big| \\ &+ \big|\zeta_0^{(J,j)} - \tfrac{1}{N'}\textstyle\sum_\kappa \Phi_J^{(j)}(\delta_\kappa)\big| \\ &+ \big|\tfrac{1}{N'}\textstyle\sum_\kappa \Phi_J^{(j)}(\delta_\kappa) - \mathcal{D}_{\tau_N}^{(N',j)}\big| \;<\; 4\varepsilon.\end{aligned}$$

Therefore $\{\mathcal{D}_{\tau_N}^{(N,j)}\}$ is Cauchy and hence convergent. Denote the limit by

$$\mu_p^{(j)} := \lim_{N\to\infty} \mathcal{D}_{\tau_N}^{(N,j)}$$

. Since $e^{-M_s} \le e^{s^{(j)}(\delta)} \le e^{M_s}$, the limit satisfies $e^{-M_s} \le \mu_p^{(j)} \le e^{M_s}$.

Assembling $\mathcal{A}_{\tau_N}^{(N)}$. Since $\tau_N + 1 = m + NT + c + 1$ and $N/(\tau_N+1) \to 1/T$ as $N \to \infty$,

$$\begin{aligned}&\frac{1}{\tau_N+1}\sum_{t\in\mathcal{I}_j} e^{\,s_{0,h}(\tau_N,\,t)} \;= \\ &\frac{N}{\tau_N+1}\cdot \mathcal{D}_{\tau_N}^{(N,j)} \;\xrightarrow{N\to\infty}\; \frac{\mu_p^{(j)}}{T}.\end{aligned}$$

Summing over all template indices and adding the vanishing non-loop contribution:

$$\begin{aligned}\mathcal{A}_{\tau_N}^{(N)} \;=\; &\underbrace{\frac{1}{\tau_N+1}\sum_{t\in\mathcal{I}^{\mathrm{non}}} e^{\,s_{0,h}(\tau_N,\,t)}}_{\to\,0} \;+ \\ &\sum_{j=1}^{T} \underbrace{\frac{1}{\tau_N+1}\sum_{t\in\mathcal{I}_j} e^{\,s_{0,h}(\tau_N,\,t)}}_{\to\,\mu_p^{(j)}/T} \;\xrightarrow{N\to\infty}\; \sum_{j=1}^{T}\frac{\mu_p^{(j)}}{T}.\end{aligned}$$

Therefore the limit

$$\mu_p \;:=\; \lim_{N\to\infty} \mathcal{A}_{\tau_N}^{(N)} \;=\; \frac{1}{T}\sum_{j=1}^{T} \mu_p^{(j)}$$

exists and is finite. Moreover, $e^{-M_s} \le \mathcal{A}_{\tau_N}^{(N)} \le e^{M_s}$ for all N, so $\mu_p > 0$.

Step 3. Cesàro limit of $e^{s_{0,h}(\tau_N,\,t)}\, v_{0,h}(t)$

We now incorporate the value vectors and show that the value-weighted Cesàro average

$$\mathcal{B}_{\tau_N}^{(N)} \;:=\; \frac{1}{\tau_N+1}\sum_{t=0}^{\tau_N} e^{\,s_{0,h}(\tau_N,\,t)}\, v_{0,h}(t) \;\in\; \mathbb{R}^{d_h}$$

converges as $N \to \infty$. We decompose the sum using the partition $\{0,\dots,\tau_N\} = \mathcal{I}^{\text{non}} \cup \bigcup_{j=1}^{T} \mathcal{I}_j$ from Step 2.

Loop contribution. At layer $\ell = 0$, the value vectors at loop positions are periodic with period T: for all $\kappa, \kappa' \in \{1,\dots,N\}$, $v_{0,h}(t_{\kappa,j}) = v_{0,h}(t_{\kappa',j})$. Define the *template value vector*

$$v_j^{\text{tpl}} \;:=\; v_{0,h}(t_{1,j}) \;\in\; \mathbb{R}^{d_h}.$$

Since each $t \in \mathcal{I}_j$ shares the same value vector v_j^{tpl}, grouping by template index gives

$$\begin{aligned}
&\frac{1}{\tau_N+1}\sum_{t\in\mathcal{I}^{\text{loop}}} e^{\,s_{0,h}(\tau_N,\,t)}\, v_{0,h}(t) \\
&= \frac{1}{\tau_N+1}\sum_{j=1}^{T} v_j^{\text{tpl}} \sum_{t\in\mathcal{I}_j} e^{\,s_{0,h}(\tau_N,\,t)} \\
&= \sum_{j=1}^{T} \frac{N}{\tau_N+1}\, \mathcal{D}_{\tau_N}^{(N,j)}\, v_j^{\text{tpl}},
\end{aligned}$$

where $\mathcal{D}_{\tau_N}^{(N,j)}$ is the restricted Cesàro average defined in Step 2. By the convergence established in Step 2, $\mathcal{D}_{\tau_N}^{(N,j)} \to \mu_p^{(j)}$ and $N/(\tau_N+1) \to 1/T$, so

$$\begin{aligned}
&\frac{1}{\tau_N+1}\sum_{t\in\mathcal{I}^{\text{loop}}} e^{\,s_{0,h}(\tau_N,\,t)}\, v_{0,h}(t) \\
&\qquad\xrightarrow{N\to\infty} \sum_{j=1}^{T} \frac{\mu_p^{(j)}}{T}\, v_j^{\text{tpl}}.
\end{aligned}$$

Non-loop contribution. The non-loop positions $\mathcal{I}^{\text{non}}$ contain $m+n+1 = O(1)$ tokens, independent of N. Since logits and values are uniformly bounded ($|s_{0,h}(\tau_N,t)| \le M_s$ and $\|v_{0,h}(t)\| \le C_v$ for all t),

$$\begin{aligned}
&\left\| \frac{1}{\tau_N+1}\sum_{t\in\mathcal{I}^{\text{non}}} e^{\,s_{0,h}(\tau_N,\,t)}\, v_{0,h}(t) \right\| \;\le \\
&\qquad \frac{|\mathcal{I}^{\text{non}}|}{\tau_N+1}\, e^{M_s}\, C_v \;\xrightarrow{N\to\infty}\; 0.
\end{aligned}$$

Convergence of $\mathcal{B}_{\tau_N}^{(N)}$**.** Combining the two contributions:

$$\begin{aligned}
\mathcal{B}_{\tau_N}^{(N)} \;&=\; \frac{1}{\tau_N+1}\sum_{t\in\mathcal{I}^{\text{loop}}} e^{\,s_{0,h}(\tau_N,\,t)}\, v_{0,h}(t) \\
&\quad + \frac{1}{\tau_N+1}\sum_{t\in\mathcal{I}^{\text{non}}} e^{\,s_{0,h}(\tau_N,\,t)}\, v_{0,h}(t),
\end{aligned}$$

and hence the limit

$$\mu_w \;:=\; \lim_{N\to\infty} \mathcal{B}^{(N)}_{\tau_N} \;=\; \sum_{j=1}^{T} \frac{\mu_p^{(j)}}{T}\, v_j^{\mathrm{tpl}} \;\in\; \mathbb{R}^{d_h}$$

exists and is finite.

Step 4. Limiting attention head output

The output of attention head h at the target position τ_N is the ratio of the value-weighted sum to the softmax normalizer:

$$a^{(N)}_{\ell,h}(\tau_N) \;:=\; \frac{\displaystyle\sum_{t=0}^{\tau_N} e^{s_{\ell,h}(\tau_N,t)}\, v_{\ell,h}(t)}{\displaystyle\sum_{t=0}^{\tau_N} e^{s_{\ell,h}(\tau_N,t)}} \;=\; \frac{\mathcal{B}^{(N)}_{\tau_N}}{\mathcal{A}^{(N)}_{\tau_N}}.$$

Steps 2 and 3 established that $\mathcal{A}^{(N)}_{\tau_N} \to \mu_p$ with $\mu_p > 0$, and $\mathcal{B}^{(N)}_{\tau_N} \to \mu_w$, as $N \to \infty$. Therefore the attention head output converges to the finite limit

$$a^{(\infty)}_{\ell,h}(\tau_N) \;:=\; \lim_{N\to\infty} a^{(N)}_{\ell,h}(\tau_N) \;=\; \frac{\mu_w}{\mu_p} \;\in\; \mathbb{R}^{d_h}.$$

□

Theorem D.4 (Base-layer convergence at the prediction position)**.** *Suppose that Assumption D.1 holds. At layer $\ell = 0$, the limits $A_0^{(\infty)} := \lim_{N\to\infty} A_0^{(N)}(\tau_N)$, $F_0^{(\infty)} := \lim_{N\to\infty} F_0^{(N)}(\tau_N)$, and $E_1^{(\infty)} := E_0^{(\infty)} + A_0^{(\infty)} + F_0^{(\infty)}$ all exist and are finite.*

D.6 Proof of Theorem D.4

Proof. We establish convergence of each component at layer $\ell = 0$.

Step 1: Multi-head attention output $A_0^{(N)}(\tau_N)$. By Proposition D.3, each attention head $h \in \{0, 1, \ldots, H-1\}$ admits a finite limit

$$a^{(\infty)}_{0,h}(\tau_N) \;:=\; \lim_{N\to\infty} a^{(N)}_{0,h}(\tau_N) \;=\; \frac{\mu_w}{\mu_p} \;\in\; \mathbb{R}^{d_h}.$$

Let $W_O \in \mathbb{R}^{d_{\mathrm{model}} \times d_{\mathrm{model}}}$ denote the output projection matrix. The multi-head attention output at the target position is

$$A^{(N)}_0(\tau_N) \;=\; W_O\big[a^{(N)}_{0,0}(\tau_N);\; \ldots\; ;\; a^{(N)}_{0,H-1}(\tau_N)\big].$$

Since each head output converges and W_O is a fixed linear map,

$$A^{(\infty)}_0(\tau_N) \;:=\; \lim_{N\to\infty} A^{(N)}_0(\tau_N) \;=$$
$$W_O\big[a^{(\infty)}_{0,0}(\tau_N);\; \ldots\; ;\; a^{(\infty)}_{0,H-1}(\tau_N)\big]$$

exists and is finite.

Step 2: Base-layer input $E_0^{(N)}(\tau_N)$. At layer $\ell = 0$, the residual stream $E_0^{(N)}(\tau_N)$ is determined by the token embedding at the target position τ_N. In a standard decoder-only model with RoPE applied inside the attention block, the embedding layer does not depend on context length, so $E_0^{(N)}(\tau_N)$ is independent of N. In particular,

$$E^{(\infty)}_0(\tau_N) \;:=\; \lim_{N\to\infty} E^{(N)}_0(\tau_N) \;=\; E^{(N)}_0(\tau_N)$$

is trivially well-defined.

Step 3: FFN output $F_0^{(N)}(\tau_N)$. Following the pre-norm update, define the post-attention residual as:

$$E_{1/2}^{(N)}(\tau_N) \;:=\; E_0^{(N)}(\tau_N) + A_0^{(N)}(\tau_N).$$

By Steps 1 and 2, $E_{1/2}^{(N)}(\tau_N)$ converges:

$$\begin{aligned} E_{1/2}^{(\infty)}(\tau_N) \;&:=\; \lim_{N\to\infty} E_{1/2}^{(N)}(\tau_N) \\ &=\; E_0^{(\infty)}(\tau_N) + A_0^{(\infty)}(\tau_N). \end{aligned}$$

The FFN output is defined as

$$F_0^{(N)}(\tau_N) := \mathrm{FFN}_0\big(\mathrm{LN}_0^{\mathrm{ffn}}(E_{1/2}^{(N)}(\tau_N))\big).$$

Under Assumption D.1, $\mathrm{LN}_0^{\mathrm{ffn}}$ is continuous on bounded sets, and FFN_0 is continuous on bounded sets. Since the residual streams are uniformly bounded (Assumption D.1 (i)), the convergence $E_{1/2}^{(N)}(\tau_N) \to E_{1/2}^{(\infty)}(\tau_N)$ propagates through both maps:

$$\begin{aligned} F_0^{(\infty)}(\tau_N) \;&:=\; \lim_{N\to\infty} F_0^{(N)}(\tau_N) \;= \\ &\mathrm{FFN}_0\Big(\mathrm{LN}_0^{\mathrm{ffn}}\big(E_{1/2}^{(\infty)}(\tau_N)\big)\Big) \end{aligned}$$

exists and is finite.

Step 4: Next-layer residual $E_1^{(N)}(\tau_N)$. Applying the residual connection:

$$\begin{aligned} E_1^{(N)}(\tau_N) \;&=\; E_{1/2}^{(N)}(\tau_N) + F_0^{(N)}(\tau_N) \\ =\; E_0^{(N)}(\tau_N) + A_0^{(N)}(\tau_N) &+ F_0^{(N)}(\tau_N). \end{aligned}$$

Since all three summands converge as $N \to \infty$, so does their sum:

$$\begin{aligned} E_1^{(\infty)}(\tau_N) \;&:=\; \lim_{N\to\infty} E_1^{(N)}(\tau_N) \\ =\; E_0^{(\infty)}(\tau_N) + A_0^{(\infty)}(\tau_N) &+ F_0^{(\infty)}(\tau_N), \end{aligned}$$

which is finite. This completes the proof. □

Proposition D.5 (Base-layer asymptotic periodicity at loop positions)**.** *Suppose that Assumption D.1 holds. For any $\varepsilon > 0$, there exists N_0 such that whenever $N \geq 2N_0$, the last $N - N_0$ loop copies are ε-periodic: any two positions $t_{\kappa,j}, t_{\kappa',j}$ in copies $\kappa, \kappa' > N_0$ sharing the same template index $j \in \{1, \ldots, T\}$ satisfy $\|E_1^{(N)}(t_{\kappa,j}) - E_1^{(N)}(t_{\kappa',j})\| < \varepsilon$. In particular, the number of such ε-periodic copies grows without bound as $N \to \infty$.*

D.7 Proof of Proposition D.5

Proof. Fix $\varepsilon > 0$. We show that for all sufficiently late loop copies, $E_1^{(N)}$ is ε-periodic at positions sharing the same template index, thereby preserving the $y\|x\|x\|\cdots\|x\|z$ loop structure required for the Cesàro argument at layer $\ell = 1$.

Step 1: Loop-position notation and relation to τ_N. For copy $\kappa \in \{1, \ldots, N\}$ and template index $j \in \{1, \ldots, T\}$, define the absolute position within the loop region:

$$t_{\kappa,j} \;:=\; m + (\kappa - 1)T + (j - 1).$$

These positions satisfy $m \leq t_{\kappa,j} < m + NT$ and thus lie strictly *inside* the repeated loop region. By contrast, the target position $\tau_N = m + NT + c \geq m + NT$ lies in the suffix z, *after* all N loop copies. In particular, every loop position $t_{\kappa,j}$ precedes τ_N.

At layer $\ell = 0$, the residual stream at any loop position is determined by the token embedding alone which depends only on the template index j and is independent of the copy index κ. In particular, $E_0^{(N)}(t_{\kappa,j}) = E_0^{(N)}(t_{\kappa',j})$ for all κ, κ'.

Step 2: Head-wise convergence at loop positions. Fix a head h. At position $t_{\kappa,j}$, the pre-RoPE query is $q_{0,h}(t_{\kappa,j})$ which is the same for every copy κ. The attention head output at $t_{\kappa,j}$ is

$$a_{0,h}^{(N)}(t_{\kappa,j}) \;=\; \frac{\displaystyle\sum_{i=0}^{t_{\kappa,j}} e^{s_{0,h}(t_{\kappa,j},\,i)}\, v_{0,h}(i)}{\displaystyle\sum_{i=0}^{t_{\kappa,j}} e^{s_{0,h}(t_{\kappa,j},\,i)}}.$$

The preceding positions $i \le t_{\kappa,j}$ decompose into:

- *Loop positions from copies* $1,\dots,\kappa{-}1$: contributing $(\kappa{-}1)T$ tokens with exactly periodic key/value structure and period T;
- *Non-loop positions*: the prefix y (m tokens) and the partial copy κ up to template index j (j tokens including self), totalling $m + j = \mathcal{O}(1)$ tokens independent of κ.

Since $t_{\kappa,j} + 1 = m + (\kappa{-}1)T + j$ grows linearly in κ, this setting satisfies all structural conditions required by the Cesàro convergence argument of Proposition D.3 (Appendix D.5, Steps 2–4). We verify the two key prerequisites explicitly:

(a) *Fixed query vector.* At layer $\ell = 0$, the pre-RoPE query at any loop position $t_{\kappa,j}$ is $q_{0,h}(t_{\kappa,j}) = W_q\,\mathrm{LN}_0^{\mathrm{attn}}(E_0^{(N)}(t_{\kappa,j})) = W_q\,\mathrm{LN}_0^{\mathrm{attn}}(\mathrm{embed}(x_j))$, which depends only on the template token x_j and is independent of the copy index κ and the total repetition count N. Hence the trigonometric polynomial structure of the logit $s_{0,h}(t_{\kappa,j},\,\cdot\,)$ has fixed coefficients for each template index j.

(b) *Cesàro-amenable loop structure.* Among the $(\kappa{-}1)T$ loop tokens preceding $t_{\kappa,j}$, the key and value vectors are exactly periodic with period T (since they are determined by token embeddings at $\ell = 0$). As $\kappa \to \infty$ (within $N \ge \kappa$), the number of preceding loop copies grows without bound, so the loop contribution forms a Cesàro average of trigonometric-polynomial exponentials that stabilises, while the $\mathcal{O}(1)$ non-loop terms (prefix y and the partial current copy) vanish.

Therefore, for each template index j and head h, the limit

$$a_{0,h}^{(\infty,j)} \;:=\; \lim_{\kappa\to\infty} a_{0,h}^{(N)}(t_{\kappa,j})$$

exists and depends on j (through the query $W_q\,\mathrm{embed}(x_j)$ and the template key/value vectors) but not on κ.

Consequently, for each h and j, there exists $\kappa_0(h,j)$ such that for all $\kappa \ge \kappa_0(h,j)$,

$$\big\|a_{0,h}^{(N)}(t_{\kappa,j}) - a_{0,h}^{(\infty,j)}\big\| \;<\; \frac{\varepsilon}{4H\,\|W_O\|\,(1+\Lambda)},$$

where $\Lambda > 0$ is the Lipschitz constant specified in Step 4. Set

$$\kappa_0 \;:=\; \max_{h\in\{0,\dots,H-1\},\; j\in\{1,\dots,T\}} \kappa_0(h,j),$$

which is finite.

Step 3: Multi-head attention near-periodicity. For any two copies $\kappa, \kappa' > \kappa_0$ with the same template index j, the triangle inequality gives for each head h:

$$\big\|a_{0,h}^{(N)}(t_{\kappa,j}) - a_{0,h}^{(N)}(t_{\kappa',j})\big\| \;<\; \frac{\varepsilon}{2H\,\|W_O\|\,(1+\Lambda)}.$$

Since $A_0^{(N)}(t) = W_O\big[a_{0,0}^{(N)}(t);\dots;a_{0,H-1}^{(N)}(t)\big]$, concatenation over H heads and the linear map W_O yield

$$\begin{aligned}&\big\|A_0^{(N)}(t_{\kappa,j}) - A_0^{(N)}(t_{\kappa',j})\big\| \\ \le\; \|W_O\|\cdot H\cdot\frac{\varepsilon}{2H\,\|W_O\|\,(1+\Lambda)} \;&=\; \frac{\varepsilon}{2(1+\Lambda)}.\end{aligned}$$

Step 4: Propagation through FFN to $E_1^{(N)}$. Since $E_0^{(N)}(t_{\kappa,j}) = E_0^{(N)}(t_{\kappa',j})$ (Step 1), the post-attention residuals satisfy

$$\begin{aligned} &\left\|E_{1/2}^{(N)}(t_{\kappa,j}) - E_{1/2}^{(N)}(t_{\kappa',j})\right\| \\ =\ &\left\|A_0^{(N)}(t_{\kappa,j}) - A_0^{(N)}(t_{\kappa',j})\right\| \ <\ \frac{\varepsilon}{2(1+\Lambda)}. \end{aligned}$$

Under Assumption D.1, $\mathrm{LN}_0^{\mathrm{ffn}}$ and FFN_0 are Lipschitz on the bounded set containing all residual streams. Let Λ_{LN} and Λ_{FFN} denote the respective Lipschitz constants, and set $\Lambda := \Lambda_{\mathrm{FFN}} \cdot \Lambda_{\mathrm{LN}}$. Then

$$\begin{aligned} &\left\|F_0^{(N)}(t_{\kappa,j}) - F_0^{(N)}(t_{\kappa',j})\right\| \ \le \\ \Lambda\left\|E_{1/2}^{(N)}(t_{\kappa,j}) - E_{1/2}^{(N)}(t_{\kappa',j})\right\| \ &<\ \frac{\Lambda\,\varepsilon}{2(1+\Lambda)}. \end{aligned}$$

Combining via the residual connection $E_1^{(N)}(t) = E_{1/2}^{(N)}(t) + F_0^{(N)}(t)$:

$$\begin{aligned} &\left\|E_1^{(N)}(t_{\kappa,j}) - E_1^{(N)}(t_{\kappa',j})\right\| \ \le \\ \frac{\varepsilon}{2(1+\Lambda)} + \frac{\Lambda\,\varepsilon}{2(1+\Lambda)} \ &=\ \frac{\varepsilon}{2} \ <\ \varepsilon. \end{aligned}$$

Step 5: Conclusion and preservation of loop structure. Setting $N_0 := \kappa_0$, we have shown that whenever $N \ge 2N_0$, any two copies $\kappa, \kappa' > N_0$ sharing the same template index $j \in \{1, \dots, T\}$ satisfy

$$\left\|E_1^{(N)}(t_{\kappa,j}) - E_1^{(N)}(t_{\kappa',j})\right\| \ <\ \varepsilon.$$

The number of such ε-periodic copies is $N - N_0$. Since N_0 depends only on ε (and the fixed model parameters) but not on N, we may take $N \to \infty$ to obtain

$$N - N_0 \ \xrightarrow[N\to\infty]{}\ \infty.$$

That is, for any prescribed $\varepsilon > 0$, the number of ε-periodic loop copies can be made arbitrarily large by choosing N sufficiently large.

This guarantees that the input to layer $\ell = 1$ preserves the $y\|x\|x\|\cdots\|x\|z$ structure with unboundedly many near-periodic loop copies preceding the target τ_N. Since $\tau_N > m + NT$ remains in the suffix after all loop copies, the target at layer $\ell = 1$ is preceded by unboundedly many near-periodic repetitions of the template block as $N \to \infty$. This is precisely the structural condition required by the Cesàro convergence argument (Appendix D.5) to be reapplied at layer $\ell = 1$, enabling the layer-wise induction in Theorem D.6. □

Theorem D.6 (Full Transformer convergence)**.** *Suppose that Assumption D.1 holds. Then for every $\ell \in \{0, 1, \dots, L\}$, the limit $E_\ell^{(\infty)} := \lim_{N\to\infty} E_\ell^{(N)}(\tau_N)$ exists and is finite. In particular,* $\lim_{N\to\infty} E_L^{(N)}(\tau_N) = E_L^{(\infty)}$.

D.8 Proof of Theorem D.6

Proof. We prove by induction on layers $\ell = 0, 1, \dots, L$ that the following two properties hold simultaneously:

- **($\mathbf{C}_\ell$)** *Convergence at the target:* $E_\ell^{(N)}(\tau_N) \xrightarrow[N\to\infty]{} E_\ell^{(\infty)}$ exists and is finite.
- **($\mathbf{P}_\ell$)** *Asymptotic periodicity at loop positions:* For any $\varepsilon > 0$, there exists N_0 (depending on ε and ℓ but not on N) such that whenever $N \ge 2N_0$, the last $N - N_0$ loop copies are ε-periodic under $E_\ell^{(N)}$: for any $\kappa, \kappa' > N_0$ and $j \in \{1, \dots, T\}$, $\|E_\ell^{(N)}(t_{\kappa,j}) - E_\ell^{(N)}(t_{\kappa',j})\| < \varepsilon$.

Base case ($\ell = 0$). At layer $\ell = 0$, the residual stream is given by token embeddings. Since the suffix token at position τ_N is always z_{c+1} (independent of N), $E_0^{(N)}(\tau_N) = \mathrm{embed}(z_{c+1})$ is independent of N, so **($\mathbf{C}_0$)** holds trivially. At loop positions, $E_0^{(N)}(t_{\kappa,j}) = \mathrm{embed}(x_j)$ depends only on the template index j and not on the copy index κ, so **($\mathbf{P}_0$)** holds with $N_0 = 1$ and ε-periodicity replaced by exact periodicity.

Induction step ($\ell \to \ell + 1$). Assume **($\mathbf{C}_\ell$)** and **($\mathbf{P}_\ell$)** hold at layer ℓ. We establish **($\mathbf{C}_{\ell+1}$)** and **($\mathbf{P}_{\ell+1}$)**.

*Step 1: Attention head convergence at τ_N (**($\mathbf{P}_\ell$) $\Rightarrow$ head limits**).* By **($\mathbf{P}_\ell$)**, the input to layer ℓ's attention block has unboundedly many near-periodic loop copies preceding τ_N. Fix a head h. We show that the attention head output

$$a_{\ell,h}^{(N)}(\tau_N) \;=\; \frac{\displaystyle\sum_{t=0}^{\tau_N} e^{s_{\ell,h}(\tau_N,t)}\, v_{\ell,h}(t)}{\displaystyle\sum_{t=0}^{\tau_N} e^{s_{\ell,h}(\tau_N,t)}}$$

converges as $N \to \infty$. At layer $\ell > 0$, the key and value vectors at loop positions are no longer exactly periodic but only approximately so; we establish convergence via a perturbation–reduction argument.

Step 1a (Reference template construction). Fix an arbitrary $\varepsilon' > 0$. By **($\mathbf{P}_\ell$)**, there exists an integer κ_0 such that whenever $N \geq 2\kappa_0$, every two loop copies $\kappa, \kappa' > \kappa_0$ sharing the same template index $j \in \{1, \ldots, T\}$ satisfy

$$\big\|E_\ell^{(N)}(t_{\kappa,j}) - E_\ell^{(N)}(t_{\kappa',j})\big\| \;<\; \varepsilon'. \tag{26}$$

For each j, fix an arbitrary reference copy $\kappa_{\mathrm{ref}} > \kappa_0$ and define the *reference key and value vectors*

$$\begin{aligned} k_{\ell,h}^{\mathrm{ref}}(j) \;&:=\; k_{\ell,h}(t_{\kappa_{\mathrm{ref}},j}), \\ v_{\ell,h}^{\mathrm{ref}}(j) \;&:=\; v_{\ell,h}(t_{\kappa_{\mathrm{ref}},j}). \end{aligned}$$

Because the maps $E_\ell^{(N)}(t) \mapsto k_{\ell,h}(t)$ and $E_\ell^{(N)}(t) \mapsto v_{\ell,h}(t)$ are Lipschitz on bounded sets (Assumption D.1), there exist constants $L_k, L_v > 0$ (depending only on the model weights and the uniform bound C) such that for all $\kappa > \kappa_0$,

$$\begin{aligned} \big\|k_{\ell,h}(t_{\kappa,j}) - k_{\ell,h}^{\mathrm{ref}}(j)\big\| \;&\leq\; L_k\, \varepsilon', \\ \big\|v_{\ell,h}(t_{\kappa,j}) - v_{\ell,h}^{\mathrm{ref}}(j)\big\| \;&\leq\; L_v\, \varepsilon'. \end{aligned} \tag{27}$$

Similarly, by **($\mathbf{C}_\ell$)**, $E_\ell^{(N)}(\tau_N)$ converges as $N \to \infty$. Since the map $E_\ell^{(N)}(\tau_N) \mapsto q_{\ell,h}(\tau_N) = W_q\, \mathrm{LN}_\ell^{\mathrm{attn}}(E_\ell^{(N)}(\tau_N))$ is continuous on bounded sets (as a composition of the Lipschitz map $\mathrm{LN}_\ell^{\mathrm{attn}}$ and the linear map W_q; Assumption D.1), there exist an integer $N_q \geq 1$ and a vector $q_{\ell,h}^{\mathrm{ref}} \in \mathbb{R}^{d_h}$ such that for all $N \geq N_q$,

$$\big\|q_{\ell,h}(\tau_N) - q_{\ell,h}^{\mathrm{ref}}\big\| \;<\; \varepsilon'. \tag{28}$$

Step 1b (Reference logit and perturbation bound). Define the *reference logit* at loop position $t_{\kappa,j}$ by

$$s_{\ell,h}^{\mathrm{ref}}(\tau_N, t_{\kappa,j}) \;:=\; \frac{\big\langle q_{\ell,h}^{\mathrm{ref}},\; R\big(-(\tau_N - t_{\kappa,j})\big)\, k_{\ell,h}^{\mathrm{ref}}(j)\big\rangle}{\sqrt{d_h}}\,.$$

Since $R(\cdot)$ is orthogonal and all query and key vectors are uniformly bounded by Assumption D.1 (denote the bounds C_q, C_k), we obtain for every $\kappa > \kappa_0$ and $N \geq N_q$:

$$\begin{aligned} &\big|s_{\ell,h}(\tau_N, t_{\kappa,j}) - s_{\ell,h}^{\mathrm{ref}}(\tau_N, t_{\kappa,j})\big| \;\leq \\ &\qquad \frac{1}{\sqrt{d_h}}\big(C_q\, L_k + C_k\big)\, \varepsilon' \;=:\; \eta\, \varepsilon'. \end{aligned} \tag{29}$$

Here the term $C_q L_k \, \varepsilon'/\sqrt{d_h}$ bounds the key perturbation equation 27 with query fixed, and $C_k \, \varepsilon'/\sqrt{d_h}$ bounds the query perturbation equation 28 with key fixed. Since all logits are bounded by M_s (Proposition D.2), the mean-value theorem gives

$$\left| e^{s_{\ell,h}(\tau_N, t_{\kappa,j})} - e^{s^{\mathrm{ref}}_{\ell,h}(\tau_N, t_{\kappa,j})} \right| \;\le\; e^{M_s} \, \eta \, \varepsilon', \tag{30}$$

for all $\kappa > \kappa_0, \; N \ge N_q$.

Key structural observation. For fixed j, the reference logit $s^{\mathrm{ref}}_{\ell,h}(\tau_N, t_{\kappa,j})$ depends on κ only through the relative distance $\delta = \tau_N - t_{\kappa,j}$, since both $q^{\mathrm{ref}}_{\ell,h}$ and $k^{\mathrm{ref}}_{\ell,h}(j)$ are independent of κ. Therefore $s^{\mathrm{ref}}_{\ell,h}(\tau_N, t_{\kappa,j})$ is a trigonometric polynomial in δ with fixed coefficients. Similarly, $v^{\mathrm{ref}}_{\ell,h}(j)$ is constant over κ. Thus the pair $(s^{\mathrm{ref}}_{\ell,h}, \, v^{\mathrm{ref}}_{\ell,h})$ satisfies the same structural hypotheses as the base-layer case ($\ell = 0$): the logit is a finite trigonometric polynomial in δ and the value vectors are exactly periodic with period T.

Step 1c (Cesàro convergence of $\mathcal{A}^{(N)}_{\tau_N}$). We show that

$$\mathcal{A}^{(N)}_{\tau_N} := \frac{1}{\tau_N + 1} \sum_{t=0}^{\tau_N} e^{s_{\ell,h}(\tau_N, t)}$$

is a Cauchy sequence in N. Partition $\{0, \ldots, \tau_N\}$ into three index sets:

- $\mathcal{I}^{\mathrm{non}}$: non-loop positions (prefix y and suffix z), $|\mathcal{I}^{\mathrm{non}}| = m + s + 1 = O(1)$;
- $\mathcal{I}^{\mathrm{early}}$: loop copies $\kappa \le \kappa_0$, $|\mathcal{I}^{\mathrm{early}}| = \kappa_0 T = O(1)$;
- $\mathcal{I}^{\mathrm{late}}$: loop copies $\kappa > \kappa_0$, $|\mathcal{I}^{\mathrm{late}}| = (N - \kappa_0)T$.

Non-loop and early-loop contributions vanish. Each exponential is bounded by e^{M_s} and $|\mathcal{I}^{\mathrm{non}} \cup \mathcal{I}^{\mathrm{early}}| = O(1)$, so

$$\begin{aligned} &\frac{1}{\tau_N + 1} \sum_{t \in \mathcal{I}^{\mathrm{non}} \cup \mathcal{I}^{\mathrm{early}}} e^{s_{\ell,h}(\tau_N, t)} \;\le \\ &\frac{(m + s + 1 + \kappa_0 T) \, e^{M_s}}{\tau_N + 1} \;\xrightarrow{N \to \infty}\; 0. \end{aligned} \tag{31}$$

Late-loop contribution: comparison with reference. Define the *reference normalizer*

$$\widetilde{\mathcal{A}}^{(N)}_{\tau_N} \;:=\; \frac{1}{\tau_N + 1} \sum_{t \in \mathcal{I}^{\mathrm{late}}} e^{s^{\mathrm{ref}}_{\ell,h}(\tau_N, t)}.$$

By equation 30, the per-position error is at most $e^{M_s} \eta \varepsilon'$, so

$$\begin{aligned} &\left| \frac{1}{\tau_N + 1} \sum_{t \in \mathcal{I}^{\mathrm{late}}} e^{s_{\ell,h}(\tau_N, t)} \;-\; \widetilde{\mathcal{A}}^{(N)}_{\tau_N} \right| \\ &\qquad \le \; \frac{|\mathcal{I}^{\mathrm{late}}|}{\tau_N + 1} \, e^{M_s} \eta \varepsilon' \;\le\; e^{M_s} \eta \varepsilon'. \end{aligned} \tag{32}$$

Reference normalizer converges. By the structural observation above, $\widetilde{\mathcal{A}}^{(N)}_{\tau_N}$ is a Cesàro average of $e^{s^{\mathrm{ref}}_{\ell,h}(\tau_N, \cdot)}$ over positions with exactly periodic key/value vectors and trigonometric-polynomial logits with fixed coefficients. We verify that this reduces to the base-layer Cesàro setting (Appendix D.5, Steps 1–2) via an explicit index substitution. For each template index j, the late-loop positions are $t_{\kappa,j}$ with $\kappa \in \{\kappa_0 + 1, \ldots, N\}$. Introduce the reversed index $\kappa' := N - \kappa + 1$, ranging over $\{1, \ldots, N - \kappa_0\}$. The relative distance becomes $\delta_\kappa = \tau_N - t_{\kappa,j} = \kappa' T + c_j$, where $c_j := c - (j-1)$ is a constant that depends only on the suffix offset c and the template index j but *not* on N or κ. Since $q^{\mathrm{ref}}_{\ell,h}$ and $k^{\mathrm{ref}}_{\ell,h}(j)$ are both fixed, the reference logit $s^{\mathrm{ref}}_{\ell,h}(\tau_N, t_{\kappa,j}) = s^{\mathrm{ref},j}(\kappa' T + c_j)$

is a trigonometric polynomial in κ' with coefficients that are *independent of* N. Therefore, grouping the sum by template index gives

$$\widetilde{\mathcal{A}}^{(N)}_{\tau_N} \;=\; \frac{N-\kappa_0}{\tau_N+1}\cdot \frac{1}{N-\kappa_0}\sum_{j=1}^{T}\sum_{\kappa'=1}^{N-\kappa_0} e^{\,s^{\mathrm{ref},j}(\kappa' T+c_j)},$$

where the inner (double) Cesàro average is over an N-independent function evaluated along an arithmetic progression with spacing T, exactly as in the base-layer argument. As $N\to\infty$, the prefactor satisfies $(N-\kappa_0)/(\tau_N+1)\to 1/T$ and the inner average converges by the same trigonometric Cesàro limit (Appendix D.5, Steps 1–2). Hence $\widetilde{\mathcal{A}}^{(N)}_{\tau_N}$ converges as $N\to\infty$.

$\mathcal{A}^{(N)}_{\tau_N}$ is Cauchy. Fix any $\varepsilon>0$. Choose ε' small enough that $e^{M_s}\eta\,\varepsilon'<\varepsilon/5$. For this ε', let κ_0 and N_q be the integers furnished above. By equation 31, there exists $N_1\ge\max(2\kappa_0,N_q)$ such that for all $N\ge N_1$ the non-loop and early-loop contribution is less than $\varepsilon/5$. By the convergence of $\widetilde{\mathcal{A}}^{(N)}_{\tau_N}$, there exists $N_2\ge N_1$ such that for all $N,N'\ge N_2$, $|\widetilde{\mathcal{A}}^{(N)}_{\tau_N}-\widetilde{\mathcal{A}}^{(N')}_{\tau_N}|<\varepsilon/5$.

For any $N,N'\ge N_2$, write

$$\mathcal{A}^{(N)}_{\tau_N}=P^{(N)}+Q^{(N)},$$

where $P^{(N)}$ collects the non-loop and early-loop contribution and $Q^{(N)}$ the late-loop contribution. Because $N,N'\ge N_2\ge\max(2\kappa_0,N_q)$, the same κ_0 applies to both and the same reference vectors $(q^{\mathrm{ref}}_{\ell,h},\,k^{\mathrm{ref}}_{\ell,h}(j),\,v^{\mathrm{ref}}_{\ell,h}(j))$ are used for both N and N'. Hence

$$\begin{aligned}\big|\mathcal{A}^{(N)}_{\tau_N}-\mathcal{A}^{(N')}_{\tau_N}\big| \;\le\; &\underbrace{|P^{(N)}|}_{<\varepsilon/5}+\underbrace{|Q^{(N)}-\widetilde{\mathcal{A}}^{(N)}_{\tau_N}|}_{\le e^{M_s}\eta\varepsilon'<\varepsilon/5}+\\ &\underbrace{|\widetilde{\mathcal{A}}^{(N)}_{\tau_N}-\widetilde{\mathcal{A}}^{(N')}_{\tau_N}|}_{<\varepsilon/5}+\underbrace{|\widetilde{\mathcal{A}}^{(N')}_{\tau_N}-Q^{(N')}|}_{<\varepsilon/5}+\underbrace{|P^{(N')}|}_{<\varepsilon/5}<\varepsilon.\end{aligned}$$

Therefore $\{\mathcal{A}^{(N)}_{\tau_N}\}_{N\ge1}$ is Cauchy and hence convergent. Moreover, $e^{-M_s}\le\mathcal{A}^{(N)}_{\tau_N}\le e^{M_s}$ for all N, so the limit

$$\mu_p \;:=\; \lim_{N\to\infty}\mathcal{A}^{(N)}_{\tau_N}$$

is finite and strictly positive.

Step 1d (Cesàro convergence of $\mathcal{B}^{(N)}_{\tau_N}$). An analogous argument establishes convergence of the value-weighted Cesàro average $\mathcal{B}^{(N)}_{\tau_N}:=\frac{1}{\tau_N+1}\sum_{t=0}^{\tau_N}e^{s_{\ell,h}(\tau_N,t)}\,v_{\ell,h}(t)$. Define the *reference numerator*

$$\widetilde{\mathcal{B}}^{(N)}_{\tau_N} \;:=\; \frac{1}{\tau_N+1}\sum_{j=1}^{T}v^{\mathrm{ref}}_{\ell,h}(j)\sum_{\kappa=\kappa_0+1}^{N}e^{\,s^{\mathrm{ref}}_{\ell,h}(\tau_N,\,t_{\kappa,j})}.$$

For $i=t_{\kappa,j}$ with $\kappa>\kappa_0$ and $N\ge N_q$, using $ab-a'b'=(a-a')b+a'(b-b')$ we bound the per-position error:

$$\begin{aligned}&\big\|e^{s_{\ell,h}(\tau_N,t)}\,v_{\ell,h}(t)-e^{s^{\mathrm{ref}}_{\ell,h}(\tau_N,i)}\,v^{\mathrm{ref}}_{\ell,h}(j)\big\|\\ &\le\big|e^{s_{\ell,h}(\tau_N,t)}-e^{s^{\mathrm{ref}}_{\ell,h}(\tau_N,i)}\big|\,\|v_{\ell,h}(t)\|\\ &\quad+e^{s^{\mathrm{ref}}_{\ell,h}(\tau_N,i)}\,\|v_{\ell,h}(t)-v^{\mathrm{ref}}_{\ell,h}(j)\|\\ &\le e^{M_s}\eta\varepsilon'\cdot C_v+e^{M_s}\cdot L_v\,\varepsilon'=\Gamma\,\varepsilon'.\end{aligned}\tag{33}$$

where $C_v:=\sup_{N,i}\|v_{\ell,h}(t)\|<\infty$ (Assumption D.1) and $\Gamma:=e^{M_s}(\eta C_v+L_v)$. Summing over $\mathcal{I}^{\mathrm{late}}$ and dividing by τ_N+1:

$$\left\|\frac{1}{\tau_N+1}\sum_{t\in\mathcal{I}^{\mathrm{late}}}e^{s_{\ell,h}(\tau_N,t)}\,v_{\ell,h}(t)\;-\;\widetilde{\mathcal{B}}^{(N)}_{\tau_N}\right\|\;\le\;\Gamma\,\varepsilon'.\tag{34}$$

The reference numerator $\widetilde{\mathcal{B}}^{(N)}_{\tau_N}$ has exactly periodic values and trigonometric-polynomial logits with fixed coefficients, so it converges by the base-layer argument (Appendix D.5, Step 3). The non-loop and early-loop contributions vanish as in equation 31. The same Cauchy-sequence argument as in Step 1c (with $e^{M_s}\eta\varepsilon'$ replaced by $\Gamma\varepsilon'$) then yields

$$\mu_w \;:=\; \lim_{N\to\infty} \mathcal{B}^{(N)}_{\tau_N} \;\in\; \mathbb{R}^{d_h}, \qquad \text{finite.}$$

Step 1e (Head output convergence). Since $\mu_p > 0$, the ratio converges:

$$a^{(\infty)}_{\ell,h}(\tau_N) \;:=\; \lim_{N\to\infty} a^{(N)}_{\ell,h}(\tau_N) \;=\; \frac{\mu_w}{\mu_p} \;\in\; \mathbb{R}^{d_h}.$$

This establishes the head-wise limit at layer ℓ for every head h, completing Step 1.

Step 2: Multi-head aggregation and FFN ($(\mathbf{C}_\ell) \Rightarrow (\mathbf{C}_{\ell+1})$). Concatenating over heads and applying the output projection:

$$A^{(\infty)}_\ell(\tau_N) \;:=\; \lim_{N\to\infty} A^{(N)}_\ell(\tau_N) \;=$$
$$W_O\big[a^{(\infty)}_{\ell,0}(\tau_N);\; \ldots \;;\; a^{(\infty)}_{\ell,H-1}(\tau_N)\big].$$

By $(\mathbf{C}_\ell)$, $E^{(N)}_\ell(\tau_N) \to E^{(\infty)}_\ell(\tau_N)$, so the post-attention residual converges:

$$E^{(N)}_{\ell+1/2}(\tau_N) \;=\; E^{(N)}_\ell(\tau_N) + A^{(N)}_\ell(\tau_N)$$
$$\xrightarrow[N\to\infty]{} E^{(\infty)}_\ell(\tau_N) + A^{(\infty)}_\ell(\tau_N) \;=:\; E^{(\infty)}_{\ell+1/2}(\tau_N).$$

Under Assumption D.1, $\mathrm{LN}^{\mathrm{ffn}}_\ell$ and FFN_ℓ are continuous on bounded sets, so

$$F^{(\infty)}_\ell(\tau_N) \;:=\; \lim_{N\to\infty} F^{(N)}_\ell(\tau_N)$$
$$=\; \mathrm{FFN}_\ell\Big(\mathrm{LN}^{\mathrm{ffn}}_\ell\big(E^{(\infty)}_{\ell+1/2}(\tau_N)\big)\Big)$$

exists. Applying the residual connection:

$$E^{(\infty)}_{\ell+1}(\tau_N) \;:=\; \lim_{N\to\infty} E^{(N)}_{\ell+1}(\tau_N)$$
$$=\; E^{(\infty)}_{\ell+1/2}(\tau_N) + F^{(\infty)}_\ell(\tau_N),$$

establishing $(\mathbf{C}_{\ell+1})$.

Step 3: Asymptotic periodicity at layer $\ell+1$ ($(\mathbf{P}_\ell) \Rightarrow (\mathbf{P}_{\ell+1})$). Fix $\varepsilon > 0$. We show that $E^{(N)}_{\ell+1}$ is ε-periodic at late loop copies by repeating the perturbation–reduction argument of Step 1 at loop positions rather than at the target τ_N, and then propagating the resulting near-periodicity of attention outputs through the multi-head aggregation and FFN exactly as in Steps 2–4 of Appendix D.7.

At $\ell = 0$, the proof of Proposition D.5 used two facts:

(i) $E^{(N)}_0(t_{\kappa,j}) = E^{(N)}_0(t_{\kappa',j})$ exactly for all κ, κ' (identical token embeddings);

(ii) the query, key, and value vectors at loop positions are exactly periodic with period T.

At layer ℓ, $(\mathbf{P}_\ell)$ replaces both (i) and (ii) with ε'-approximate versions: for $\kappa, \kappa' > \kappa_0$,

$$\big\|E^{(N)}_\ell(t_{\kappa,j}) - E^{(N)}_\ell(t_{\kappa',j})\big\| \;<\; \varepsilon'.$$

Step 3a (Attention head near-periodicity at loop positions). Fix a head h and a template index j. Consider a loop position $t_{\kappa,j}$ with κ large. The attention head output $a^{(N)}_{\ell,h}(t_{\kappa,j})$ sums over all preceding positions $i \le t_{\kappa,j}$, of which $(\kappa-1)T$ are loop tokens from copies $1, \ldots, \kappa-1$ and $\mathcal{O}(1)$ are non-loop tokens.

We apply the perturbation–reduction technique from Step 1 with $t_{\kappa,j}$ playing the role of τ_N:

- *Query perturbation.* The query vector at $t_{\kappa,j}$ is $q_{\ell,h}(t_{\kappa,j}) = W_q \,\mathrm{LN}_\ell^{\mathrm{attn}}(E_\ell^{(N)}(t_{\kappa,j}))$. By **($\mathbf{P}_\ell$)**, for any two late copies $\kappa, \kappa' > \kappa_0$ with the same template index j, the representations $E_\ell^{(N)}(t_{\kappa,j})$ are within ε', so the query vectors are within $L_q\,\varepsilon'$ (where L_q is the Lipschitz constant of $W_q \circ \mathrm{LN}_\ell^{\mathrm{attn}}$). Fix a reference query $q_{\ell,h}^{\mathrm{ref},j} := q_{\ell,h}(t_{\kappa_{\mathrm{ref}},j})$ for some $\kappa_{\mathrm{ref}} > \kappa_0$.
- *Key/value perturbation and reference logit.* For the preceding loop positions, the same reference key/value vectors $k_{\ell,h}^{\mathrm{ref}}(\cdot)$, $v_{\ell,h}^{\mathrm{ref}}(\cdot)$ and the same perturbation bounds equation 27 apply. The reference logit $s_{\ell,h}^{\mathrm{ref}}(t_{\kappa,j},\, t_{\kappa'',j'}) = \frac{1}{\sqrt{d_h}}\langle q_{\ell,h}^{\mathrm{ref},j},\, R(-(t_{\kappa,j} - t_{\kappa'',j'}))\, k_{\ell,h}^{\mathrm{ref}}(j')\rangle$ is a trigonometric polynomial in the relative distance with fixed coefficients (independent of κ and N), since both $q_{\ell,h}^{\mathrm{ref},j}$ and $k_{\ell,h}^{\mathrm{ref}}(j')$ are fixed.
- *Cesàro convergence.* The same index-reversal argument as in Step 1c shows that the reference normalizer and reference numerator at $t_{\kappa,j}$ each converge as $\kappa \to \infty$ (within $N \geq \kappa$), with convergence rate controlled by the perturbation level ε'.

Combining these, for each head h and template index j, the attention head output at late loop positions stabilises: there exists $\kappa_1(h,j)$ (independent of N) such that for all $\kappa, \kappa' > \kappa_1(h,j)$,

$$\left\| a_{\ell,h}^{(N)}(t_{\kappa,j}) - a_{\ell,h}^{(N)}(t_{\kappa',j}) \right\| \;<\; \frac{\varepsilon}{4H\,\|W_O\|\,(1+\Lambda_\ell)},$$

where $\Lambda_\ell := \Lambda_{\mathrm{FFN},\ell} \cdot \Lambda_{\mathrm{LN},\ell}$ is the composed Lipschitz constant at layer ℓ.

Step 3b (Multi-head aggregation and FFN propagation). Setting $\kappa_0(\ell{+}1) := \max_{h,j} \kappa_1(h,j)$ and proceeding exactly as in Steps 3–4 of Appendix D.7, concatenation over heads, application of W_O, and propagation through the Lipschitz maps $\mathrm{LN}_\ell^{\mathrm{ffn}}$ and FFN_ℓ yield

$$\left\| E_{\ell+1}^{(N)}(t_{\kappa,j}) - E_{\ell+1}^{(N)}(t_{\kappa',j}) \right\| \;<\; \varepsilon$$

for all κ, κ' beyond $\kappa_0(\ell{+}1)$, which is independent of N. The number of such ε-periodic copies is $N - \kappa_0(\ell{+}1) \to \infty$ as $N \to \infty$, establishing **($\mathbf{P}_{\ell+1}$)**.

Conclusion. By induction, **($\mathbf{C}_\ell$)** and **($\mathbf{P}_\ell$)** hold for all $\ell = 0, 1, \ldots, L$. In particular, **($\mathbf{C}_L$)** gives

$$E_L^{(\infty)} \;:=\; \lim_{N\to\infty} E_L^{(N)}(\tau_N)$$

exists and is finite, completing the proof. □

D.9 Deriving Component-Level Convergence

We now derive the convergence of each component output $O_c(N)$ from Theorem D.6.

Proof of Lemma 3.4. We treat the two component types separately.

Attention heads. Fix a layer ℓ and a head h. By Theorem D.6, the head-wise limit $a_{\ell,h}^{(\infty)}(\tau_N) := \lim_{N\to\infty} a_{\ell,h}^{(N)}(\tau_N)$ exists and is finite. Since $a_{\ell,h}(N) = a_{\ell,h}^{(N)}(\tau_N)$ and $W_{O,\ell}^{(h)}$ is a fixed linear map, we have

$$O_{\ell,h}^{A}(\infty) \;:=\; \lim_{N\to\infty} O_{\ell,h}^{A}(N) \;=\; W_{O,\ell}^{(h)}\, a_{\ell,h}^{(\infty)}(\tau_N),$$

which exists and is finite.

FFN neurons. Fix a layer ℓ and a neuron index r. By Theorem D.6, the post-attention residual limit $E_{\ell+\frac{1}{2}}^{(\infty)} := \lim_{N\to\infty} E_{\ell+\frac{1}{2}}^{(N)}(\tau_N)$ exists and is finite. Under Assumption D.1, $\mathrm{LN}_\ell^{\mathrm{ffn}}$ is continuous on bounded sets. Since the residual streams are uniformly bounded (Assumption D.1(i)), the composition

$$\mathrm{LN}_\ell^{\mathrm{ffn}}\big(E_{\ell+\frac{1}{2}}^{(N)}(\tau_N)\big) \;\xrightarrow{N\to\infty}\; \mathrm{LN}_\ell^{\mathrm{ffn}}\big(E_{\ell+\frac{1}{2}}^{(\infty)}\big)$$

converges. The up-projection W_ℓ^{up} is a fixed linear map and the element-wise nonlinearity σ is continuous, so the neuron activation $m_{\ell,r}(N) = [\sigma(W_\ell^{\mathrm{up}}\,\mathrm{LN}_\ell^{\mathrm{ffn}}(E_{\ell+\frac{1}{2}}^{(N)}(\tau_N)))]_r$ converges:

$$m_{\ell,r}(\infty) \;:=\; \lim_{N\to\infty} m_{\ell,r}(N)$$

exists and is finite for every r. Since $w_{\ell,r}^{\mathrm{down}}$ is a fixed vector, we obtain

$$O_{\ell,r}^{F}(\infty) \;:=\; \lim_{N\to\infty} O_{\ell,r}^{F}(N) \;=\; w_{\ell,r}^{\mathrm{down}}\, m_{\ell,r}(\infty),$$

which exists and is finite.

Since both cases yield finite limits, for every component c, $O_c(\infty) := \lim_{N\to\infty} O_c(N)$ exists and is finite. □