

Tracing Contextual Override: Mediated Causal Attribution in Language Models

Zongye Hu[♦] Weiqing Luo[♦] Haofeng Zhang[♣] Ziyi Huang[♦]

[♦]Arizona State University [♣]Morgan Stanley

{weiqing2,zongyehu,zhuan236}@asu.edu hz2553@columbia.edu

Abstract

Contextual override, the tendency of large language models to shift away from otherwise correct predictions under misleading prompt evidence, poses a fundamental risk to reliable deployment in context-rich environments. Despite its prevalence, the internal mechanism behind this shift remains poorly understood, particularly the causal structure underlying the override process. To bridge this gap, we formulate contextual override as an internal mediation process that attributes evidence-induced answer-preference shifts to fine-grained model components. Building on this formulation, we introduce a theoretically grounded mediated-effect framework that disentangles evidence sensitivity from causal mediation, allowing override-mediating components to be precisely identified and their causal role verified through targeted ablation. We empirically validate our framework across six models and three benchmarks, demonstrating consistent support for our mediation account across diverse architectures and task settings. Our study reveals that *contextual override is routed through sparse, structurally localized pathways, rather than being diffusely entangled with the model’s internal knowledge*. These findings provide a mechanistic diagnostic for auditing contextual override and a targeted interventional test for validating the identified pathways, rather than a general-purpose defense mechanism. We release our code at https://github.com/abhorizon03/Contextual_Override.

1 Introduction

Modern language models are increasingly deployed in context-rich environments, with prompt-provided evidence playing a central role in shaping their predictions (Dong et al., 2024; Lou et al., 2024; Luo et al., 2026). Yet such evidence is often neither curated nor guaranteed to be truthful, and even a single misleading passage can steer a model

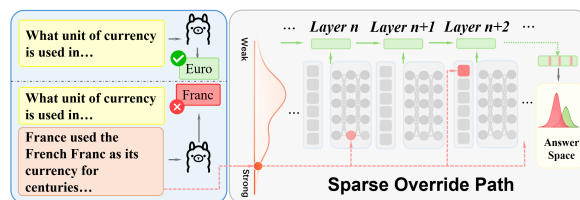


Figure 1: **Overview of contextual override and sparse override pathways.** **Left:** Misleading prompt evidence triggers contextual override, shifting the model from a correct prediction (Euro) to an incorrect alternative (Franc). **Right:** The mediated effect score distribution (vertical axis) shows that external-context sensitivity concentrates in a small number of components (red) while the majority of components (gray) exhibit negligible mediated effect scores. These components form a *sparse* override pathway that shifts the answer distribution from the correct answer (green) toward the context-supported alternative (red). Illustration is based on Qwen3-8B with MINTAKA.

toward an incorrect answer despite encoded knowledge of the correct one (Longpre et al., 2021; Xie et al., 2024; Yu et al., 2024; Xu et al., 2024; Hu et al., 2026). This phenomenon, which we refer to as *contextual override*, creates a direct channel for injecting misinformation and manipulating model outputs at scale. Dialogue and interactive settings further amplify this risk, as misleading cues can be repeatedly introduced within the prompt and exert cumulative influence on the model’s prediction (Li et al., 2025c; Yang et al., 2025; Li et al., 2025b; Cheng’an Wei et al., 2024). Mitigating contextual override at its source demands mechanistic insight beyond behavioral observation: *Which components shift the model’s answer preference toward the context-supported alternative?*

Existing studies of contextual override have been largely empirical, documenting model behavior under misleading or conflicting context without explaining the internal mechanism behind the shift (Zhou et al., 2023; Bi et al., 2025a,b; Wang et al., 2025; Cattan et al., 2025). Prompt-level defenses, filtering strategies, and post-hoc verification

can reduce exposure to misleading evidence, but they do not reveal how such evidence is represented or propagated once it enters the model (Yi et al., 2025; Chen et al., 2025). Model-level interventions such as fine-tuning, model editing, or broad pruning can alter behavior, but they often lack the granularity needed to isolate the computations responsible for the answer shift and may damage the model’s underlying knowledge (Tighidet et al., 2024; Minder et al., 2025; Ma et al., 2026). Consequently, contextual override remains largely characterized as an input-output failure rather than as an internal causal process. However, prompt evidence can either enhance reasoning or override internal knowledge, making it important to understand how context shapes model predictions (Xie et al., 2024; Longpre et al., 2021; Lewis et al., 2020).

To address this question, we develop a mechanistic framework for identifying and intervening on the fine-grained internal pathways that mediate contextual override, linking behavioral shifts in answer preference to subcomponent-level attribution and targeted causal validation. As illustrated in Figure 1, we instantiate this framework in a controlled setting where a model initially prefers the correct answer but shifts toward an incorrect alternative after misleading evidence is added to the prompt. This setting isolates contextual override from ordinary knowledge failure, ensuring that the observed shift reflects the influence of external evidence rather than the model’s inability to answer the question. It also captures a central risk of context-rich deployment: external evidence can silently displace knowledge the model would otherwise rely on. This makes the failure difficult to detect from standard capability evaluations alone, since a model may perform well in isolation yet still be redirected by misleading context at inference time (Shi et al., 2023; Liu et al., 2024a; Yu et al., 2024). Understanding contextual override therefore requires tracing how external evidence changes the model’s answer preference inside the model, rather than relying solely on behavioral diagnostics at the output level (Geva et al., 2023).

Our analysis reveals that contextual override is not a diffuse, model-wide response but is concentrated in a *sparse* set of components, as shown in Figure 1 (right). By separating evidence sensitivity from causal relevance, our framework distinguishes components that merely react to external evidence from those that actually transmit its influence to the model’s final prediction. We formalize

this distinction through a *mediated effect score* that combines each component’s evidence-induced activation change with its downstream influence on the answer preference. Guided by this score, we find that targeted pruning affecting only 2.0% of parameters on average restores the model’s original correct preference in over 73% of overridden examples. These results indicate that contextual override is not a global degradation of model behavior, but a localized process routed through identifiable components that can be selectively neutralized.

Beyond mechanistic insight, our findings yield several practical implications. First, the mediated effect score offers a theoretically grounded diagnostic for model vulnerability, identifying the internal components most susceptible to contextual manipulation and potentially providing a fine-grained signal for red-teaming and safety evaluation. Second, the sparsity and locality of the override pathway suggest that mitigation can move beyond output-level verification toward internal diagnosis and targeted control. Third, we use targeted suppression as interventional validation: selectively reversing override while preserving clean-input behavior supports the identified components as mediators of contextual influence. Our contributions can be summarized as follows:

- **Problem Formulation.** We propose a mediation theory of contextual override, casting the phenomenon as an internal causal pathway from misleading external evidence to answer-preference shift. This perspective distinguishes context-sensitive components from causally relevant mediators, turning override from a behavioral failure into a traceable internal causal process.
- **Fine-Grained Mediated-Effect Attribution and Intervention.** We develop the first mediated-effect framework for fine-grained causal attribution and targeted intervention, isolating external-context pathways from the model’s internal knowledge. The framework enables selective suppression of override-mediating components at the prediction position, shifting mitigation from coarse behavioral correction to targeted mechanistic intervention.
- **Sparse Override Pathways.** We empirically validate our framework across six models and three benchmarks, demonstrating that override-mediating components can be identified and selectively suppressed. Our analysis reveals that the override pathway is sparse, structurally localized,

and largely separable from the model’s own knowledge representations.

2 Related Work

Knowledge Conflict under External Evidence.

A related area is knowledge conflict, which studies model behavior when contextual and parametric signals disagree (Xu et al., 2024; Amugongo et al., 2025; Longpre et al., 2021; Wang et al., 2023; Xie et al., 2024). While this line of work has proposed various mitigation strategies, such as alignment training and conflict-aware generation (Zhou et al., 2023; Bi et al., 2025b; Wang et al., 2025; Cattani et al., 2025), these approaches mainly address the behavioral outcome without attributing the shift to individual model components. Moreover, most existing studies remain empirical, focusing on whether models follow context or memory, while leaving the internal causal process behind evidence-induced answer shifts largely unexplained (Kortukov et al., 2024; Marjanović et al., 2024; Liu et al., 2024b; Hagström et al., 2025).

Mechanistic Analyses of Contextual Influence.

Recent work has started to investigate knowledge conflict through a mechanistic lens, focusing on the internal processes by which models balance parametric knowledge against external contextual evidence (Wang et al., 2024; Li et al., 2025a; Haklay et al., 2025). While powerful, they are often applied to coarse representational units, such as hidden activations, subspaces, attention representations, or layer-wise hidden-state edits (Tighidet et al., 2024; Minder et al., 2025; Ma et al., 2026). Such granularity makes it difficult to isolate the positional mediators responsible for answer shifts under conflict, calling for a more fine-grained, position-aware causal account of contextual override.

Attribution of External-Evidence Pathways.

Existing methods rarely separate external-evidence sensitivity from internal-knowledge insufficiency, a distinction crucial for conflict mitigation (Zhang et al., 2024; Sun et al., 2025a; Sui et al., 2025). This raises an attribution problem over answer shifts and motivates diagnosis of external-evidence pathways before targeted intervention (Sun et al., 2025b; Zhang and Zhang, 2025).

3 Causal Attribution and Targeted Intervention for Contextual Override

In this section, we study the mediation of external contextual evidence through internal model components. We focus on cases in which external evidence shifts the model from the correct answer to a supported alternative, and identify the components that transmit this influence to the final prediction. We first define an answer-shift metric to measure this preference change, then introduce a component-level mediated-effect score that combines each component’s evidence-induced activation change with its downstream effect on the final answer preference. Since natural-language evidence is discrete, we further instantiate the evidence path with repeated external evidence and validate the identified components through targeted pruning at the prediction position. An overview of the framework is provided in Figure 2. Detailed notation is provided in Appendix C.1, and all proofs are provided in Appendix C.2 and C.3.

3.1 Problem Setup and Notation

We begin with a continuous formulation that provides the theoretical basis for our component-level attribution score. Consider a language model processing a query y accompanied by external evidence supporting an alternative answer. Let $e_a \in \mathcal{V}$ denote the wrong answer promoted by the external evidence and $e_b \in \mathcal{V}$ the correct answer, where $\mathcal{V}$ is the token vocabulary. We define the **answer-shift metric** as the logit difference

$$g_{a,b} = \langle \Delta_{e_a, e_b}, E \rangle, \quad (1)$$

where $\Delta_{e_a, e_b} = W_U[e_a] - W_U[e_b]$, with W_U the unembedding matrix and E the final-layer residual representation at the prediction position. A positive $g_{a,b}$ indicates preference for the wrong answer e_a .

We consider two types of **components** c at the prediction position: (i) attention heads, whose residual-space contribution is $O_{\ell, h}^A = W_{O, \ell}^{(h)} a_{\ell, h}$, and (ii) FFN neurons, with scalar output $O_{\ell, r}^F = w_{\ell, r}^{\text{down}} m_{\ell, r}$. We use O_c as the output of component c generically when the component type is clear from context. Our central question is: *which components c are causally responsible for shifting $g_{a,b}$ when external evidence is introduced?*

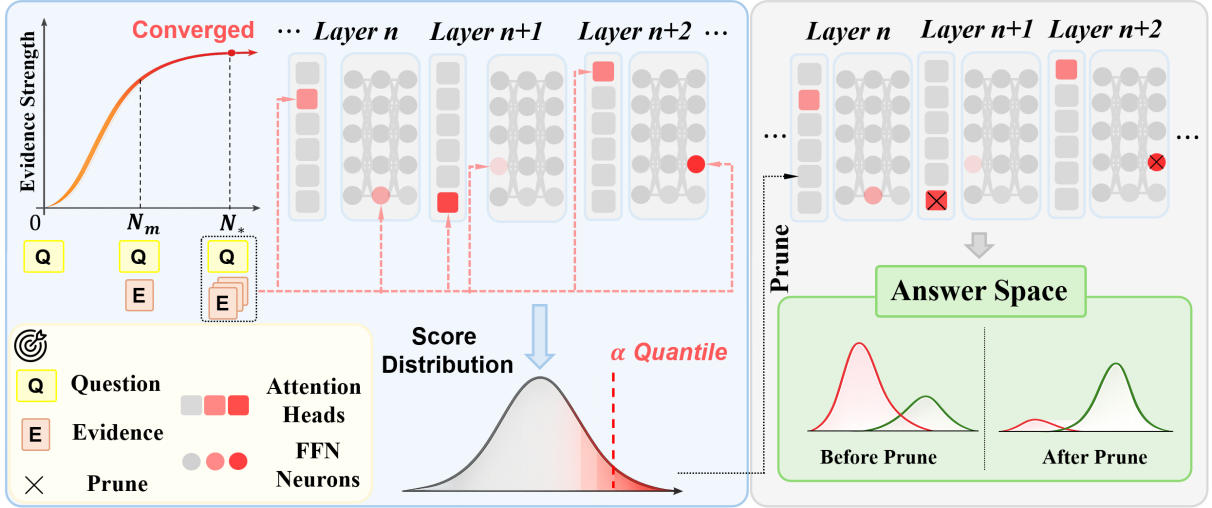


Figure 2: **Overview of the mediated-effect attribution and intervention framework.** **Left: Mediated-effect attribution.** Repeated external evidence serves as a discrete proxy for evidence strength, with component outputs saturating at N^* . Mediated effect scores are computed for each attention head and FFN neuron, and components exceeding the α -quantile threshold are selected as override mediators. **Right: Targeted intervention.** Selected components ($\times$) are zeroed out exclusively at the prediction position, shifting the answer distribution from the context-supported alternative (red) back toward the correct answer (green).

3.2 Theoretical Formulation of Continuous External Mediated Effect

To ground our approach theoretically, we first define the mediated effect in a continuous domain where $\lambda \geq 0$ parameterizes external evidence strength.

Definition 3.1 (Cumulative Mediated Effect). *The cumulative first-order effect of component c on $g_{a,b}$ over $[\lambda_0, \lambda_1]$ is defined as*

$$I_c(\lambda_0, \lambda_1) = \int_{\lambda_0}^{\lambda_1} \langle \nabla_{O_c} g_{a,b}(\lambda), \dot{O}_c(\lambda) \rangle d\lambda. \quad (2)$$

This integral accumulates the sensitivity of the answer-shift metric to each component’s trajectory as evidence strength varies. However, evaluating this integral requires access to the full path $O_c(\lambda)$, which is impractical. Thus, we impose a local stability condition on the component trajectory and downstream gradient to obtain a practical estimator.

Assumption 3.2 (Regularity of Component Trajectories). Suppose that for each component c : 1) Directional Stability. The velocity $\dot{O}_c(\lambda)$ is approximately aligned with the total displacement $O_c(\lambda_1) - O_c(\lambda_0)$ for a.e. $\lambda \in [\lambda_0, \lambda_1]$. 2) Gradient Stability. The downstream gradient $\nabla_{O_c} g_{a,b}(\lambda)$ is approximately constant over $[\lambda_0, \lambda_1]$.

Under this assumption, the path integral can be approximated by the inner product between the

endpoint displacement of the component and an anchor-point downstream gradient.

Proposition 3.3. *Under Assumption 3.2, the cumulative mediated effect (Definition 3.1) admits the practical version*

$$I_c(\lambda_0, \lambda_1) = \langle \hat{G}_c, O_c(\lambda_1) - O_c(\lambda_0) \rangle, \quad (3)$$

where $\hat{G}_c = \nabla_{O_c} g_{a,b}(\lambda_m)$ is the downstream gradient evaluated at anchor points $\lambda_m \in [\lambda_0, \lambda_1]$.

Proposition 3.3 motivates a simple attribution principle. A component receives a large positive mediated-effect score when two conditions hold: its output changes substantially along the external-evidence path, and this change is aligned with the downstream direction that increases preference for the externally supported answer.

3.3 Discrete Realization under Repeated External Evidence

While Proposition 3.3 provides a clean endpoint formula, a continuous λ is practically intractable. Language models operate on discrete token sequences, so a single piece of evidence cannot be fractionally modulated (*inaccessibility*). Furthermore, the abstract influence strength is jointly determined by semantic content, position, and model internals, making it impossible to calibrate multiple evidence inputs onto a shared continuous axis (*non-identifiability*). We therefore adopt a discrete

instantiation that preserves the structure of Proposition 3.3 while being exactly computable.

We use the integer repetition count $N \geq 0$ as a discrete proxy for evidence strength, and reuse $O_c(\cdot)$ and $g_{a,b}(\cdot)$ from the continuous formulation with N in place of λ for notational continuity. Concretely, given a query y , an external evidence template x asserting wrong answer e_a , and a suffix z , we construct the input $u^{(N)} = y \| x^N \| z$ by concatenating N copies of x , and write $O_c(N)$ for the output of component c at the prediction position.

To ensure endpoint displacement captures the full external-evidence effect, we require component outputs to converge under increasing repetition:

Lemma 3.4 (Asymptotic Convergence). *For each component c , the limit $O_c(\infty) := \lim_{N \rightarrow \infty} O_c(N)$ exists and is finite. (See Appendix C.3 for proof and Appendix C.4 for empirical validation)*

Lemma 3.4 guarantees that each component’s output saturates, so the displacement $O_c(N^*) - O_c(0)$ for sufficiently large N^* captures the exhaustive effect of external evidence on component c . Combining Proposition 3.3 with Lemma 3.4 yields the **Discrete Mediated Effect Score**, our primary working metric:

Corollary 3.5 (Discrete Mediated Effect Score). *Under Assumption 3.2 and Lemma 3.4, the exhaustive mediated effect of component c is approximated by*

$$I_c^D = \langle \hat{G}_c, O_c(N^*) - O_c(0) \rangle, \quad (4)$$

where N^* is a sufficiently large number and $\hat{G}_c = \nabla_{O_c} g_{a,b}(N_m)$ is the downstream gradient at anchor points $N_m \in [0, N^*]$.

The score I_c^D decomposes into two interpretable factors: the *displacement* $O_c(N^*) - O_c(0)$, which measures how much component c changes under external evidence, and the *gradient* $\hat{G}_c^b$, which measures how sensitive the answer-shift metric is to changes in component c . A large positive I_c^D identifies components that both respond strongly to external evidence and influence the model’s answer preference toward the externally supported answer.

3.4 Targeted Identification and Pruning of Mediating Components

Building on the mediated effect scores in Corollary 3.5, we design a pruning procedure with three key properties: *causal grounding*, as it targets the exact quantity measured by I_c^D ; *position specificity*,

with interventions restricted to the prediction position to avoid cascading confounds; and *fine granularity*, through prediction-position interventions on individual attention heads and FFN neurons rather than removing entire layers or modules.

Component Selection. Let $\mathcal{Q}$ denote the evaluation dataset. For each component c at layer ℓ , we compute $I_c^D(q)$ for every $q \in \mathcal{Q}$ and obtain the median score $\tilde{I}_c^D = \text{median}_{q \in \mathcal{Q}} I_c^D(q)$. Within each layer ℓ , we rank all components by $\tilde{I}_c^D$ and select those exceeding the α -quantile as external-evidence-mediating components:

$$\mathcal{C}_\ell^+ = \left\{ c \in \ell \mid \tilde{I}_c^D \geq Q_\alpha \left(\{ \tilde{I}_{c'}^D \}_{c' \in \ell} \right) \right\}. \quad (5)$$

Prediction-Position Pruning. Given prompts $\mathcal{P}$ and denoting the model’s forward pass by f , the pruned output is

$$\hat{y} = f(\mathcal{P} \mid \text{do}(O_c = 0, \forall c \in \bigcup_\ell \mathcal{C}_\ell^+)), \quad (6)$$

where the intervention zeroes out all selected components exclusively at the prediction position.

Crucially, this intervention does not remove an attention head or neuron from the model entirely: a selected component remains active at all other token positions and is silenced only at the position where the output logits are computed. The effective intervention unit is therefore *component* $\times$ *position*, which is strictly finer-grained than head-level or layer-level pruning and affects only 2.0% of parameters on average, as detailed in Appendix B.1.

4 Experiments

Our study introduces the first fine-grained mediation framework for contextual override. We hypothesize that contextual override is mediated by a sparse and identifiable set of components, and conduct experiments on *mechanism validation* rather than task-level mitigation performance. We thus evaluate causal effectiveness, knowledge specificity, component-selection validity, and generalization to question-specific persuasive contexts.

4.1 Experimental Setup

Benchmarks and Models. We evaluate on three benchmark subsets: MMLU (Hendrycks et al., 2021) (closed, 874 QA), OpenBookQA (Mihaylov et al., 2018) (closed, 1,769 QA), and MINTAKA (Sen et al., 2022) (open, 370 QA), ensuring breadth of evaluation. Details are provided in Appendix A.1. We evaluate six LLMs spanning

Table 1: **Prediction stability after mediated effect-based pruning under external-context override.** Values report the percentage of override examples that maintain the original correct preference $p_{e_b} > p_{e_a}$ across repetition lengths $N \in \{1000, 1500, 2000\}$.

Model	Benchmarks								
	MMLU			OBQA			MINTAKA		
	1000	1500	2000	1000	1500	2000	1000	1500	2000
Qwen3-1.7B	84.2	85.5	88.3	87.6	88.4	90.9	73.4	79.0	75.5
Qwen3-4B	83.2	84.3	87.4	66.1	66.9	75.0	66.7	69.1	71.6
Qwen3-8B	90.8	91.6	92.1	96.0	96.4	95.1	84.6	83.8	82.9
Gemma3-1b-it	62.5	61.4	61.2	63.8	63.0	63.6	63.8	70.9	58.9
Gemma3-4b-it	88.5	90.0	88.9	89.0	90.7	90.0	83.7	76.0	77.4
Gemma3-12b-it	74.8	58.9	52.3	52.3	34.3	25.3	94.9	95.6	88.0
Mean	80.7	78.6	78.4	75.8	73.3	73.3	77.9	79.1	75.7

two distinct model families and a range of capacities, including Qwen3-1.7B, Qwen3-4B, Qwen3-8B (Team, 2025b), Gemma3-1b-it, Gemma3-4b-it, and Gemma3-12b-it (Team, 2025a). This selection covers small (<3B), medium (3B-6B), and large (>6B) scales within the practical compute budget of a single NVIDIA A100 GPU, while the training-free nature of our procedure allows direct extension to larger or architecturally distinct models when resources permit.

Override Setting. For each question, we append repeated assertions of a distractor answer to the original prompt, following Appendix A.4. Our main analysis focuses on cases where the model initially prefers the correct answer ($p_{e_b} > p_{e_a}$) but shifts toward the distractor under contextual prompting, isolating instances where external context overrides the model’s original preference.

Mediated Effect Computation and Pruning. We compute component-level mediated effect scores I_c^D (Eq. 4) for attention heads and FFN neurons using the displacement $O_c(N^*) - O_c(0)$ from the clean input ($N = 0$) to a large repetition count N^* , and the downstream gradient $\hat{G}_c$ at an intermediate anchor N_m . Scores are aggregated by per-component medians across the evaluation set, and components above the layer-wise $\alpha = 0.995$ quantile are pruned (Eq. 5). Pruning is applied only at the prediction position by zeroing selected components through dynamic masking on output-projection pre-hooks. Details are in Appendix A.2.

Evaluation Metrics. We report three metrics. *Prediction stability* is the fraction of originally overridden examples whose correct answer preference is recovered after pruning under the same context-

tual prompt. *Balanced accuracy* averages clean-input retention for initially correct examples and recovery for initially incorrect examples ($N = 0$), with 50% indicating a knowledge-neutral intervention. *Recovery rate* is the fraction of initially incorrect predictions corrected by pruning. Formal definitions are in Appendix A.3.

4.2 Causal Validation on External-Context Override

We first evaluate our method for localizing external-evidence-sensitive components. In Table 1, we report the fraction of questions in this subset that maintain $p_{e_b} > p_{e_a}$ after pruning, across repetition lengths $N \in \{1000, 1500, 2000\}$ with the targeted components selected when $N = 1000$, which is sufficiently large for convergence. Mediated effect-based pruning consistently restores the model’s original preference, with mean stability above 73% across all benchmarks and repetition lengths. Notably, this effect is achieved by pruning only 2.0% of parameters on average, detailed in Appendix B.1, since the intervention targets selected subcomponents within attention heads and FFN layers rather than removing entire heads or layers. Additionally, the effect persists at longer repetition lengths, suggesting that the selected components capture a stable external-context pathway rather than a prompt-length-specific artifact.

4.3 Specificity of Identified Components across External and Internal Pathways

We next evaluate whether our method localizes external-context influence without conflating it with the model’s internal answer preference. We run the pruned model on the original QA inputs without repeated external evidence and compare

Table 2: **Balanced accuracy on original QA inputs after pruning.** Values report balanced accuracy between retention of initially correct predictions and recovery of initially incorrect predictions after removing external-context-sensitive components. A value close to **50.0** indicates a knowledge-neutral intervention on the model’s internal answer preference.

Model	MMLU	OBQA	MINTAKA
Qwen3-1.7B	55.6	59.0	53.5
Qwen3-4B	57.5	61.9	54.7
Qwen3-8B	55.8	56.6	53.4
Gemma3-1b-it	47.9	49.8	50.6
Gemma3-4b-it	46.6	46.0	51.7
Gemma3-12b-it	59.2	54.1	55.1
Mean	53.8	54.6	53.2

Table 3: **Recovery rate for initially incorrect questions after pruning.** Values report the percentage of initially incorrect predictions that shift to the correct preference after removing external-context-sensitive components, providing a measure of unexpressed internal knowledge revealed by targeted pruning.

Model	MMLU	OBQA	MINTAKA
Qwen3-1.7B	38.2	44.8	36.9
Qwen3-4B	37.6	43.2	24.2
Qwen3-8B	48.9	35.3	15.8
Gemma3-1b-it	48.5	50.4	28.7
Gemma3-4b-it	31.3	27.2	30.9
Gemma3-12b-it	40.3	40.7	31.3
Mean	40.8	40.3	28.0

its answer preferences with the pre-pruning baseline, assessing whether the identified components are specific to the external-context pathway. To avoid inflated estimates from imbalanced outcome groups, we report balanced accuracy in Table 2. This metric averages retention of initially correct predictions and recovery of initially incorrect ones, with values close to 50% indicating a knowledge-neutral intervention. Across settings, balanced accuracy remains close to and slightly above this baseline, suggesting that the pruned components are largely orthogonal to internal knowledge and that their removal does not systematically degrade intrinsic predictive capacity. This supports the specificity of our method in targeting external-evidence-sensitive components without substantially interfering with internal knowledge pathways.

In Table 3, we report the recovery rate for initially incorrect questions to characterize the relationship between external-context-sensitive components and unexpressed internal knowledge. Across

Table 4: **Negative control with randomized component selection.** We report prediction stability and balanced accuracy after randomly pruning the same number of components as mediated effect-based pruning. The **Ours** row reports the corresponding results from proposed mediated effect-based pruning for comparison.

	Model	MMLU	OBQA	MINTAKA
Prediction Stability	Qwen3-1.7B	2.1	0.5	0.0
	Qwen3-4B	5.1	2.2	0.8
	Qwen3-8B	7.7	3.4	1.8
	Gemma3-1b-it	3.3	3.1	1.9
	Gemma3-4b-it	0.0	0.4	1.0
	Gemma3-12b-it	6.2	4.0	5.1
	Mean	4.1	2.3	1.8
Balanced Accuracy	Ours	80.7	75.8	77.9
	Qwen3-1.7B	54.0	55.6	52.5
	Qwen3-4B	51.3	51.7	51.5
	Qwen3-8B	53.1	52.2	50.1
	Gemma3-1b-it	52.9	51.2	54.7
	Gemma3-4b-it	57.3	56.3	51.7
	Gemma3-12b-it	53.7	53.6	53.2
	Mean	53.7	53.4	52.3
	Ours	53.8	54.6	53.2

settings, pruning yields a moderate recovery rate, suggesting that removing external-context-sensitive components can potentially reveal latent internal knowledge that was previously suppressed and provide a possible avenue for improving model accuracy through targeted pruning. This recovery pattern supports the selectivity of our method in targeting external evidence pathways while limiting interference with internal knowledge.

4.4 Diagnostic Control: Randomized Component Selection

We compare mediated effect-based pruning with random pruning using the same intervention size, separating the effect of component selection from the effect of pruning itself. Table 4 reports prediction stability and balanced accuracy after random pruning, with additional recovery-rate results provided in Appendix B.2. Random pruning leads to a sharp drop in prediction stability compared with our method, with mean stability falling from above 73% to below 5% across all benchmarks. This result shows that the reduction of external influence is attributable to the precise identification of external-sensitive components, reflecting the discriminative power of our method in isolating external evidence pathways. Meanwhile, mediated effect-based pruning achieves balanced accuracy

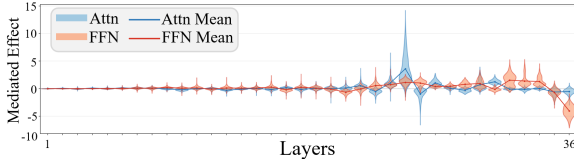


Figure 3: **Layer-wise mediated effect distributions on Qwen3-4B with OpenBookQA.** Mediated effect scores are aggregated over all evaluation questions and plotted separately for attention and FFN components. The distributions show that external-context sensitivity is not uniformly spread across the model, but concentrates in a small number of structured layer-module regions.

comparable to random pruning, indicating that the stability gain comes from targeted component selection while preserving a similar knowledge-neutral profile on original QA inputs.

4.5 Structural Sparsity of External-Context-Sensitive Components

We further characterize the structural pattern of mediated effect scores across layers and modules to reveal the distribution of external-context sensitivity. Figure 3 aggregates mediated effect scores over all OpenBookQA evaluation questions on Qwen3-4B and plots the layer-wise distributions for attention and FFN components. Most layers have mediated effect scores concentrated near zero, indicating that external-context sensitivity is not uniformly distributed across the model. Instead, clear deviations appear in a small number of middle-to-late layers. Attention components show a sharp localized peak, while FFN components exhibit a broader late-layer shift. This distributional pattern suggests that external-context sensitivity is concentrated in structured layer-module regions rather than spread uniformly across the network.

Figure 4 visualizes the head-level mediated effect pattern on Qwen3-4B. For each benchmark, we select a representative high-mediated-effect layer and visualize mediated effect scores across attention heads and questions. Rows denote heads, columns denote questions, and color intensity indicates effect strength. The heatmaps show that mediated effect is concentrated in a small number of heads rather than uniformly spread across all heads, suggesting a sparse head-level structure within the external-context-sensitive pathway. Additional structural patterns across models and benchmarks are provided in Appendix B.3.

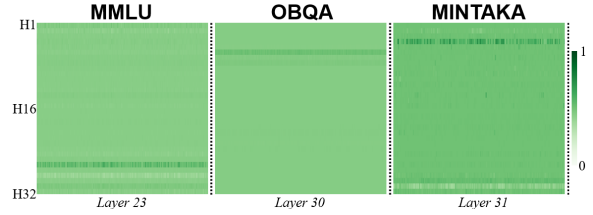


Figure 4: **Head-level mediated effect patterns on Qwen3-4B.** Mediated effect scores are visualized across attention heads and evaluation questions within one representative high-mediated effect layer for each benchmark. Rows correspond to attention heads, columns correspond to questions, and color intensity indicates mediated effect strength. Mediated effect is concentrated in a small number of heads, suggesting a sparse head-level structure within the external-context-sensitive pathway.

Table 5: **Prediction stability under question-specific persuasive contexts.** Values report the percentage of examples that maintain the correct preference after mediated effect-based pruning, evaluated on samples where distractor-targeted contextual interference flips the baseline prediction.

Model	MMLU	OBQA	MINTAKA
Qwen3-1.7B	58.1	53.8	52.7
Qwen3-4B	56.9	55.6	57.8
Qwen3-8B	59.4	53.6	53.1
Gemma3-1b-it	48.8	41.7	46.4
Gemma3-4b-it	60.0	60.6	83.7
Gemma3-12b-it	34.0	52.6	55.1
Mean	53.0	52.9	58.1

4.6 Generalization to Question-Specific Persuasive Contexts

To examine the practical implication of our method, we evaluate the identified components under natural contextual interference setting where external evidence is semantically tailored to individual questions, as detailed in Appendix A.4. In Table 5, we evaluate the pruned model on the subset of samples where this interference flips the baseline prediction under this general persuasive-context setting. Our mediated effect-based pruning consistently improves stability across settings, suggesting that the identified components capture a broader pathway for context-induced answer-preference shifts beyond a single repetition pattern, with potential for mitigating contextual interference in practice.

4.7 Retrieval Augmented Generation Case Studies

Finally, we provide qualitative examples showcasing that the proposed mediated-effect anal-

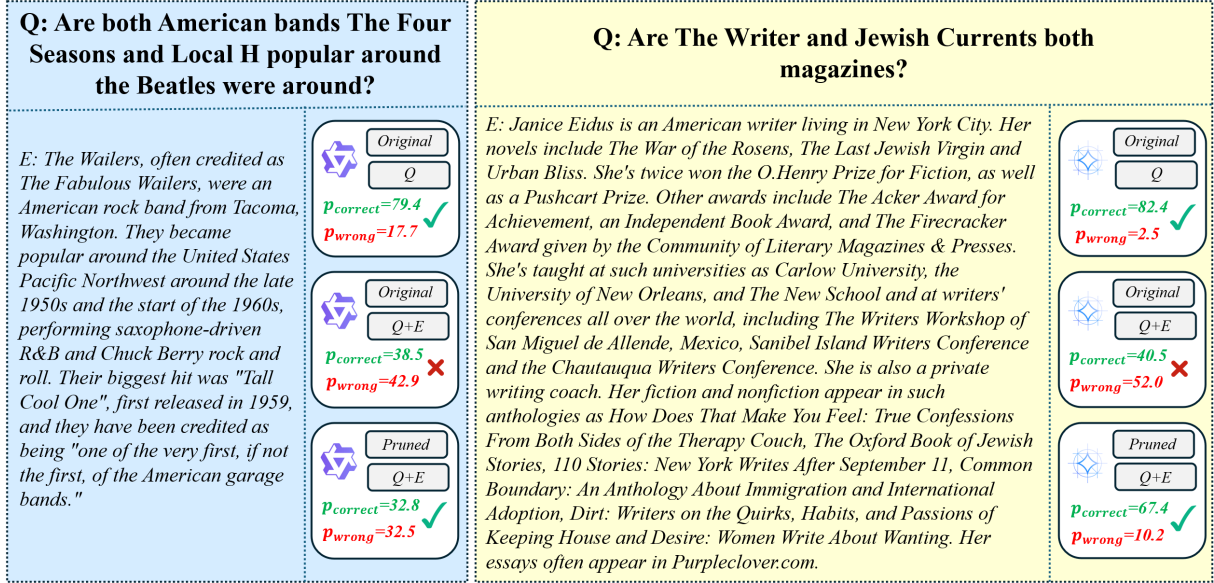


Figure 5: **Qualitative case studies on retrieval-induced contextual override.** **Left:** A HotpotQA case with Qwen3-4B, where the model initially predicts the correct answer under the question-only input but is shifted toward the wrong answer after adding realistic retrieved evidence. **Right:** A HotpotQA case with Gemma3-4b-it showing the same retrieval-induced override pattern. In both cases, p_{correct} and p_{wrong} denote the probabilities assigned to the original correct answer and the retrieved-evidence-biased wrong answer, respectively. Pruning the components identified by proposed framework restores the correct prediction under (Q+E).

ysis extends to practically deployed language-model paradigms, with retrieval-augmented generation (RAG) as a representative setting. We use BM25 (Harman, 1994) on HotpotQA (Yang et al., 2018) to induce natural RAG retrieval failures, where query-relevant but non-answer-supporting passages trigger contextual override and apply mediated-effect pruning to restore the model’s original correct prediction. Figure 5 shows that models recover their original correct preferences under our framework after retrieval-induced override, supporting our study’s practical relevance under incidental retrieval failures in realistic end-to-end RAG. These RAG examples highlight the practical value of our framework as an intervention-oriented diagnostic for identifying and reducing contextual override in deployed language-model systems even in realistic settings.

5 Conclusion

We introduce a theoretical framework for the mechanistic analysis of contextual override, casting the phenomenon as an internal mediation process over fine-grained model components. This perspective yields a principled distinction between evidence sensitivity and causal mediation, connecting the behavioral manifestation of override to its internal causal origins. Our study shows that contextual

override is not a diffuse architectural vulnerability but a localized causal process that can be precisely attributed and selectively suppressed. Through extensive experiments, we substantiate this mediation account and show that our framework complements existing behavioral characterizations with a mechanistic understanding of the override process.

Limitations

We study contextual override under controlled conditions isolating cases where misleading external evidence alters otherwise correct answer preferences. This supports causal analysis but does not cover all context-induced failures; extending the mediation framework to richer settings is an important future direction. Targeted suppression is primarily intended for causal validation and auditing. Although it suggests lightweight mitigation, deployment would require mechanisms to assess intervention appropriateness and balance override reduction with general model utility. We therefore regard it as complementary to existing safeguards rather than a standalone solution. Our mechanistic diagnosis aims to clarify contextual override. Without introducing new model capabilities or user-facing deployment mechanisms, we do not anticipate direct additional risks beyond standard concerns for interpretability tools.

References

- Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877.
- Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. 2025a. Context-dpo: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10280–10300.
- Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025b. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.
- Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roei Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. 2025. Dragged into conflicts: Detecting and addressing conflicting sources in search-augmented llms. *arXiv preprint arXiv:2506.08500*.
- Yulin Chen, Haoran Li, Yuan Sui, Yufei He, Yue Liu, Yangqiu Song, and Bryan Hooi. 2025. Can indirect prompt injection attacks be detected and removed? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18189–18206.
- Kai Chen Cheng'an Wei, Yue Zhao, Yujia Gong, Lu Xiang, and Shenchen Zhu. 2024. Context injection attacks on large language models. *arXiv preprint arXiv:2405.20234*.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.
- Lovisa Hagström, Sara Vera Marjanovic, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2025. A reality check on context utilisation for retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19691–19730.
- Tal Haklay, Hadas Orgad, David Bau, Aaron Mueller, and Yonatan Belinkov. 2025. Position-aware automatic circuit discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2792–2817.
- Donna K Harman. 1994. *The second text retrieval conference (TREC-2)*, volume 31. US Department of Commerce, National Institute of Standards and Technology.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Zongye Hu, Weiqing Luo, Yanjie Fu, Yu Gan, Haofeng Zhang, and Ziyi Huang. 2026. Not all answers are contextually persuadable: Inference dynamics in large language models under contextual influence. In *Forty-third International Conference on Machine Learning*.
- Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under context-memory conflicts with real documents. *arXiv preprint arXiv:2404.16032*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025a. Taming knowledge conflicts in language models. *arXiv preprint arXiv:2503.10996*.
- Xueshen Li, Xinlong Hou, Nirupama Ravi, Ziyi Huang, and Yu Gan. 2025b. A two-stage proactive dialogue generator for efficient clinical information collection using large language model. *Expert systems with applications*, 287:127833.
- Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025c. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.
- Yantao Liu, Zijun Yao, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, and Juanzi Li. 2024b. Untangle the knot: Interweaving conflicting knowledge and reasoning skills in large language models. In *Proceedings of the 2024 Joint International Conference*

- on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 17186–17204.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. [Large language model instruction following: A survey of progresses and challenges](#). *Computational Linguistics*, 50(3):1053–1095.
- Weiqing Luo, Zongye Hu, Xiao Wang, Zhiyuan Yu, Haofeng Zhang, and Ziyi Huang. 2026. Utility-oriented visual evidence selection for multimodal retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35091–35124.
- Xuhua Ma, Richong Zhang, and Zhijie Nie. 2026. Corect: Context-aware logit contrast for hidden state rectification to resolve knowledge conflicts. *arXiv preprint arXiv:2602.08221*.
- Sara Vera Marjanović, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. Dynamicqa: Tracing internal knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14346–14360.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.
- Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. 2025. Controllable context sensitivity and the knob behind it. In *International Conference on Learning Representations*, volume 2025, pages 81073–81102.
- Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. [Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.
- Yi Sui, Chaozhuo Li, Chen Zhang, Dawei Song, and Qiuchi Li. 2025. Bridging external and parametric knowledge: Mitigating hallucination of llms with shared-private semantic synergy in dual-stream knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10845–10869.
- Kaiser Sun, Fan Bai, and Mark Dredze. 2025a. Task matters: Knowledge requirements shape llm responses to context-memory conflict. *arXiv preprint arXiv:2506.06485*.
- Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025b. Redeeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *International Conference on Learning Representations*, volume 2025, pages 50250–50279.
- Gemma Team. 2025a. [Gemma 3 technical report](#). Technical report, Google DeepMind.
- Qwen Team. 2025b. [Qwen3 technical report](#). Preprint, arXiv:2505.09388.
- Zineddine Tighidet, Jiali Mei, Benjamin Piwowarski, and Patrick Gallinari. 2024. Probing language models on their knowledge source. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 604–614.
- Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. 2025. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571.
- Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, and 1 others. 2024. Knowledge mechanisms in large language models: A survey and perspective. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7097–7135.
- Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, volume 2024, pages 35623–35646.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565.

- Ye Yang, Donghe Li, Zuchen Li, Fengyuan Li, Jingyi Liu, Li Sun, and Qingyu Yang. 2025. [MisinfoBench: A multi-dimensional benchmark for evaluating LLMs’ resilience to misinformation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10202–10229, Suzhou, China. Association for Computational Linguistics.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.
- Jingwei Yi, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2025. Benchmarking and defending against indirect prompt injection attacks on large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1809–1820.
- Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Truth-aware context selection: Mitigating hallucinations of large language models being misled by untruthful contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10862–10884.
- Hao Zhang, Yuyang Zhang, Xiaoguang Li, Wenxuan Shi, Haonan Xu, Huanshuo Liu, Yasheng Wang, Lifeng Shang, Qun Liu, Yong Liu, and 1 others. 2024. Evaluating the external and parametric knowledge fusion of large language models. *arXiv preprint arXiv:2405.19010*.
- Wan Zhang and Jing Zhang. 2025. Hallucination mitigation for retrieval-augmented large language models: a review. *Mathematics*, 13(5):856.
- Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.

A Implementation Details

A.1 Benchmarks Pre-processing

MMLU. Experiments are performed on the *miscellaneous* subset of MMLU, aggregating the test, validation, and dev splits to yield 874 instances. We select the miscellaneous category as it spans diverse everyday topics without concentrating on any single specialized domain, ensuring that the internal knowledge probed is broadly representative and that cross-model comparisons are not confounded by domain-specific expertise gaps. For each instance, one answer option is randomly selected as a distractor, which is then used to construct the external evidence.

OpenBookQA. Experiments are performed on the train split of OpenBookQA. We retain only queries ending with “?” to ensure well-formed question structure compatible with our prompt template. For each instance, one answer option is randomly selected as a distractor, which is then used to construct the external evidence.

MINTAKA. Experiments are performed on the train split of the MINTAKA dataset. We first remove any examples whose language label is not English. Next, we exclude instances whose complexityType field is yesno or count. From the remaining pool, we use the Qwen3-4B tokenizer to select only those questions whose correct answer is tokenized as a single token. For each of these selected questions, we generate a distractor by prompting GPT-5.2 to produce a single-token answer candidate (The prompt is shown in Appendix A.4); the generated distractor candidates are then re-tokenized with the Qwen3-4B tokenizer and retained only if they consist of a single token. This procedure ensures both the original answer and the distractor are single-token under the Qwen3-4B tokenization scheme.

A.2 Model Prune

We apply two pruning configurations depending on the model.

Qwen3-1.7B, Qwen3-4B, Qwen3-8B, and Gemma3-12b-it. We rank all components within each layer by their mediated effect score I_c^D (Eq. 4) and set the quantile threshold $\alpha = 0.995$ in Eq. 5 to select external-evidence-mediating components from both attention heads and FFN neurons, with

the constraint that at least one attention head is pruned per layer.

Gemma3-1b-it and Gemma3-4b-it. We apply the same ranking and threshold ($\alpha = 0.995$) but restrict pruning to attention heads only, again ensuring that at least one attention head is pruned per layer.

A.3 Metrics

Prediction Stability. We restrict evaluation to the subset $\mathcal{Q}_{\text{override}}$ of examples where the model initially assigns correct preference ($p_{e_b} > p_{e_a}$) but external evidence shifts preference toward the wrong answer ($p_{e_b} < p_{e_a}$). Prediction stability is defined as

$$\text{PS} = \frac{|\{q \in \mathcal{Q}_{\text{override}} \mid \hat{p}_{e_b}(q) > \hat{p}_{e_a}(q)\}|}{|\mathcal{Q}_{\text{override}}|}, \quad (7)$$

where $\hat{p}$ denotes the pruned model’s output probabilities under the same external evidence.

Balanced Accuracy. We evaluate the pruned model on original QA inputs without external evidence. Balanced accuracy is defined as

$$\text{BA} = \frac{1}{2} \left(\frac{n_{\text{retain}}}{n_{\text{correct}}} + \frac{n_{\text{remain}}}{n_{\text{incorrect}}} \right), \quad (8)$$

where n_{correct} and $n_{\text{incorrect}}$ are the numbers of initially correct and incorrect predictions, n_{retain} is the number of initially correct predictions preserved after pruning, and n_{remain} is the number of initially incorrect predictions that remain incorrect. A value close to 50% indicates a knowledge-neutral intervention.

Recovery Rate. Recovery rate measures the percentage of initially incorrect predictions that shift to the correct preference after pruning:

$$\text{RR} = \frac{|\{q \in \mathcal{Q}_{\text{incorrect}} \mid \hat{p}_{e_b}(q) > \hat{p}_{e_a}(q)\}|}{|\mathcal{Q}_{\text{incorrect}}|}, \quad (9)$$

where $\mathcal{Q}_{\text{incorrect}} = \{q \mid p_{e_b}(q) < p_{e_a}(q)\}$ is the set of initially incorrect predictions and $\hat{p}$ denotes probabilities after pruning.

A.4 Contextual Interference

A.4.1 Repeated External Evidence Prompts

For multiple-choice questions, we construct a repeat-biased prompt that asks the model to respond with a single letter. Given a question Q and a chosen distractor option C , the template repeats the distractor assertion N times to construct $u^{(N)}$:

Close-end Template

Reply with only one letter: Q? A. <option A>, B. <option B>, C. <option C>, D. <option D>. C. <option C> is true. C. <option C> is true. ... Answer:

Example:

Reply with only one letter: Poison causes harm to which of the following? A. a Tree, B. a robot, C. a house, D. a car. A. a Tree is true. A. a Tree is true. A. a Tree is true. Answer:

For open-ended questions, we construct a prompt that directly presents the question Q and appends a repeated assertion of the distractor answer:

Open-end Template

Answer in one word.
<question>? "<distractor>" is true.
"<distractor>" is true. ...
Answer:

Example:

Answer in one word.
Did Avatar come out before or after 2010?
"After" is true. "After" is true. "After" is true. ...
Answer:

A.4.2 Distractor Generation for MINTAKA

For the MINTAKA benchmark, distractors are not available from the original dataset. We generate single-token distractor candidates by prompting GPT-5.2 with the following instruction:

Distractor Generation Prompt

I will give you several questions based on following structure: "idx, id, lang, question, answer, category, complexityType, distractor". You need to adjust the distractors to make it wrong but confusing, the distractors should be limited in one token by mainstream LLM tokenizer. I will send you several questions at a time, generate distractor one by one and return them based on the structure (best in csv).

The generated candidates are then tokenized with the Qwen3-4B tokenizer and retained only if they consist of a single token, as described in Appendix A.1.

A.4.3 Question-Specific Persuasive Contexts

For the generalization experiment in Section 4.6, we construct distractor-targeted contextual interference tailored to individual questions. We randomly sample 100 examples from each benchmark and prompt Claude Opus 4.6 to generate a question-specific rationale supporting the distractor:

Persuasive Context Generation Prompt

You are a very stubborn student. I will give you a csv file containing 100 questions. Each question contains a correct answer and a distractor. You need to read the questions and generate a reason to persuade LLM one by one. You can't generate a reason for a class of questions and reuse it, your chain of thought need to be: read a question, correct answer and distractor -> write a reason -> move to next question. Process all the questions and return a csv file.

This procedure produces semantically varied, per-question persuasive evidence, enabling evaluation of whether the identified components generalize beyond the repeated-evidence setting to broader forms of contextual interference.

B Additional Results

B.1 Prune Parameters

Since our intervention zeroes out selected components exclusively at the prediction position via dynamic masking on W_O pre-hooks, we quantify the intervention size by the fraction of parameters whose output is affected at that position.

For each pruned attention head h at layer ℓ , the affected parameters correspond to the h -th column block of $W_O^{(\ell)}$, comprising $d_{\text{head}} \times d_{\text{model}}$ parameters. For each pruned FFN neuron i at layer ℓ , the affected parameters correspond to the i -th column of $W_{\text{down}}^{(\ell)}$, comprising d_{model} parameters. The prune fraction at the prediction position is then defined as:

$$\text{param@pos} = \frac{N_h d_{\text{head}} d_{\text{model}} + N_n d_{\text{model}}}{L (H d_{\text{head}} d_{\text{model}} + d_{\text{ff}} d_{\text{model}})}, \quad (10)$$

where N_h and N_n are the total numbers of pruned heads and neurons across all layers, L is the number of layers, H is the number of heads per layer, and d_{ff} is the feed-forward intermediate dimension. The denominator counts all parameters in W_O and W_{down} across layers, representing the full set of residual-write parameters. Note that since the in-

Table 6: $\text{param@pos}/T$ across different benchmarks and models. OBQA denotes Openbook QA, and MTK denotes MINTAKA.

Model	MMLU	OBQA	MTK	Mean
Qwen3-1.7B	1.94	2.32	1.94	2.07
Qwen3-4B	1.28	1.64	1.28	1.40
Qwen3-8B	1.16	1.16	1.16	1.16
Gemma3-1b-it	3.23	4.11	3.23	3.52
Gemma3-4b-it	2.08	2.92	2.08	2.36
Gemma3-12b-it	1.71	1.71	1.71	1.71
Mean	1.90	2.31	1.90	2.04

Table 7: Recovery rate under random pruning. OBQA denotes Openbook QA.

Model	MMLU	OBQA	MINTAKA
Qwen3-1.7B	11.2	15.9	14.6
Qwen3-4B	4.3	5.6	5.5
Qwen3-8B	6.8	4.8	2.6
Gemma3-1b-it	38.1	34.6	26.7
Gemma3-4b-it	19.7	19.6	6.2
Gemma3-12b-it	9.0	8.8	10.9
Mean	14.9	14.9	11.1

intervention is applied at a single token position, the effective influence over the full sequence is approximately $\text{param@pos}/T$ where T is the sequence length.

We report $\text{param}\%@\text{pos}/T$ as it directly reflects the activation-channel coverage at the prediction position. Table 6 reports the prune fractions across all models and benchmarks. Across all configurations, the intervention affects fewer than 4% of residual-write parameters at the prediction position, with a global average of 2.04%, confirming that the identified external-evidence pathways are structurally sparse.

B.2 Ablation Study Results of Recovery Rate

We also report the recovery rate under random pruning as an ablation baseline, shown in Table 7. While random pruning yields non-zero recovery rates, the mean recovery is substantially lower than the one achieved by our mediated effect-based pruning.

B.3 Score Distribution

We report the complete head-level mediated effect score matrices (Figure 6) and layer-wise aggregated distributions (Figure 7) for all models and benchmarks, extending the representative results shown in Section 4.5. The full results consistently exhibit

the same structural pattern: mediated effect scores concentrate in a small number of heads within specific layers, confirming that the sparsity of external-context-sensitive components is a robust property across model families and scales.

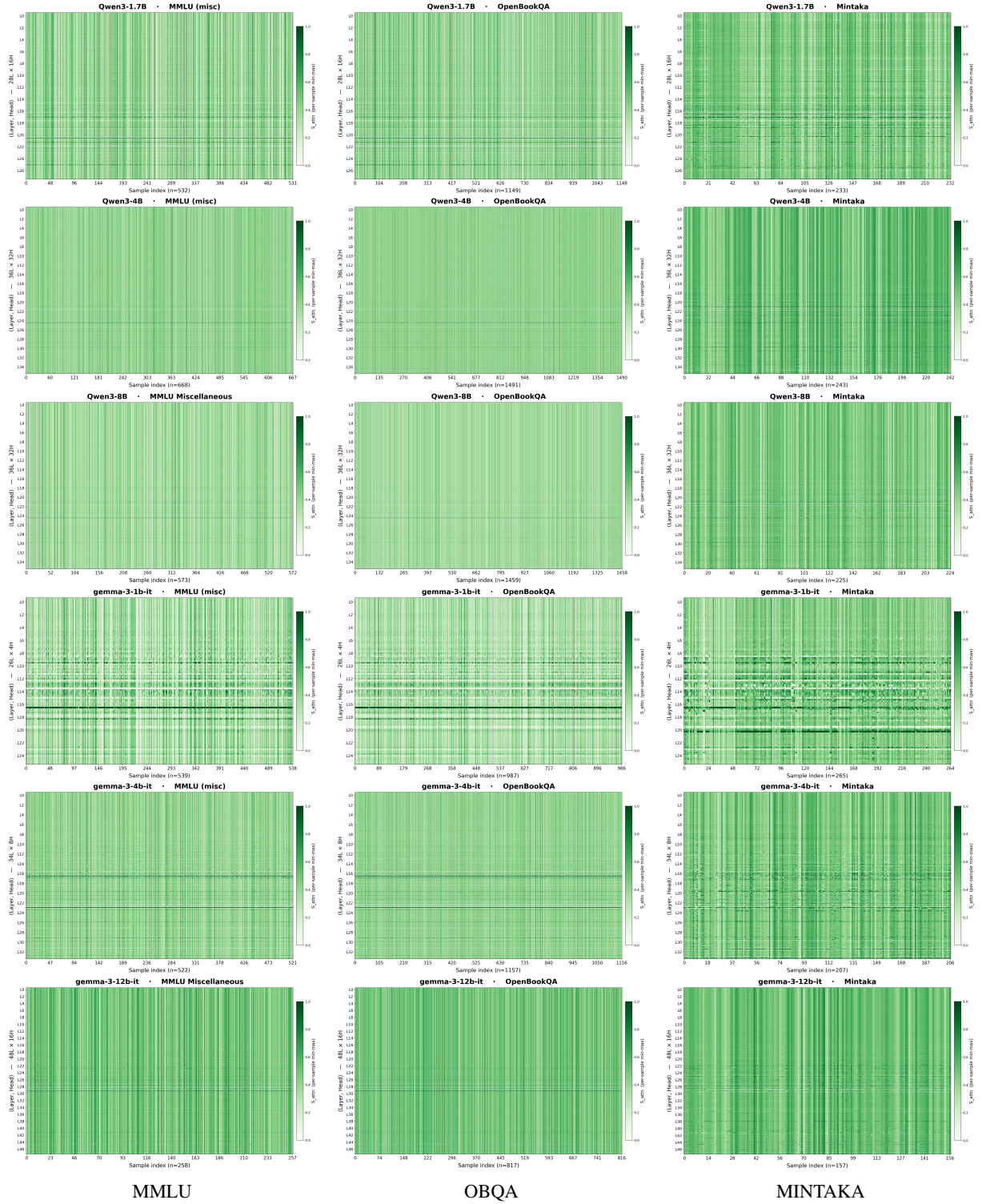


Figure 6: Head-level mediated effect score matrices across all models and benchmarks. Each heatmap visualizes mediated effect scores across attention heads (rows) and evaluation questions (columns) within representative high-mediated-effect layers. Rows from top to bottom: Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Gemma3-1B-it, Gemma3-4B-it, Gemma3-12B-it. Columns: MMLU, OpenBookQA, MINTAKA.

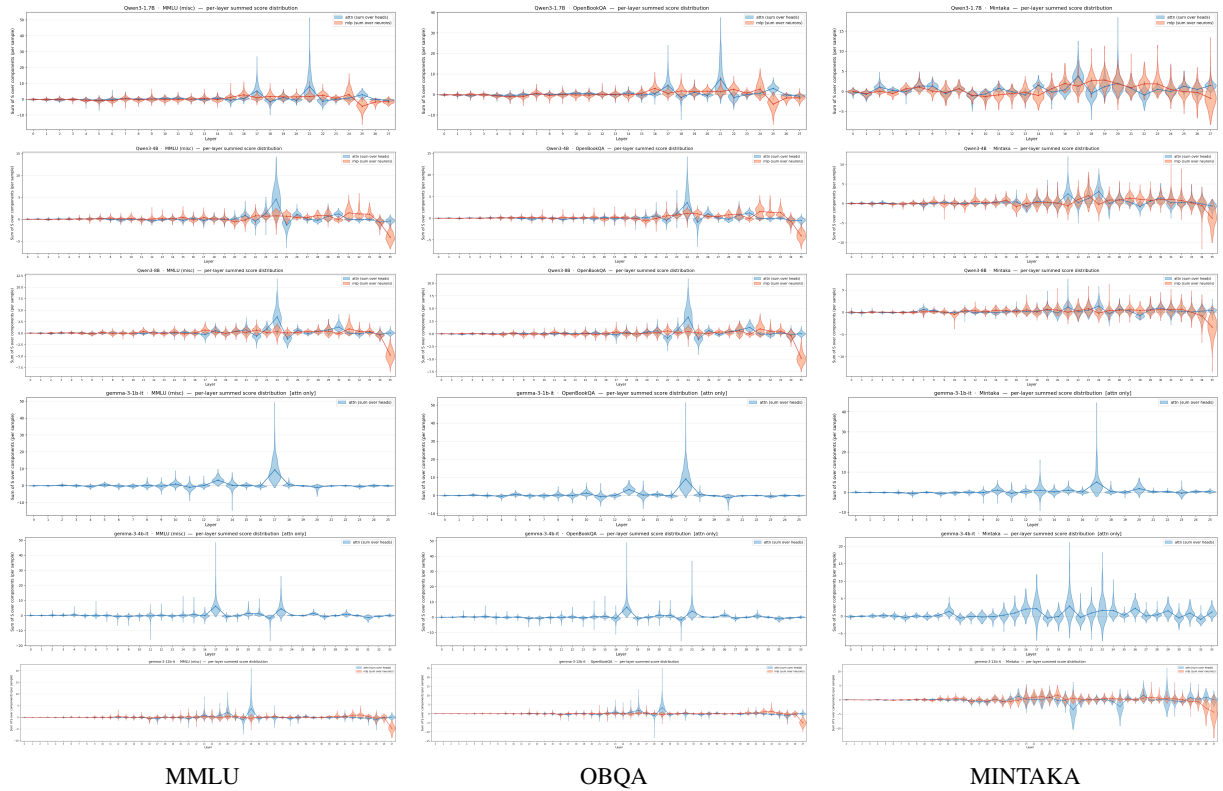


Figure 7: Layer-wise mediated effect score distributions across all models and benchmarks. Rows from top to bottom: Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Gemma3-1B-it, Gemma3-4B-it, Gemma3-12B-it. Columns: MMLU, OpenBookQA, MINTAKA.

C Theorem Framework

C.1 Detailed Notations

C.1.1 Continuous Domain Notations

Let $\lambda \geq 0$ denote the external evidence strength, treated as a continuous real variable. Let $e_a \in \mathcal{V}$ denote the wrong answer supported by the external evidence and $e_b \in \mathcal{V}$ the correct answer, where $\mathcal{V}$ is the token vocabulary.

Answer-Shift Metric. Let $W_U \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$ denote the unembedding matrix and define the logit direction as

$$\Delta_{e_a, e_b} = W_U[e_a] - W_U[e_b] \in \mathbb{R}^{d_{\text{model}}}. \quad (11)$$

Let $E(\lambda) \in \mathbb{R}^{d_{\text{model}}}$ denote the final-layer residual-stream representation at the prediction position. The answer-shift metric is defined as

$$g_{a,b}(\lambda) = \langle \Delta_{e_a, e_b}, E(\lambda) \rangle, \quad (12)$$

which equals the logit difference between e_a and e_b . A larger $g_{a,b}$ indicates greater model preference for the wrong answer e_a over the correct answer e_b .

Component Outputs. We consider two types of components c at the prediction position:

Attention heads. For head h at layer ℓ , the component output is the residual-space head contribution:

$$O_{\ell,h}^A(\lambda) = W_{O,\ell}^{(h)} a_{\ell,h}(\lambda), \quad (13)$$

where $a_{\ell,h}(\lambda) \in \mathbb{R}^{d_h}$ is the pre-projection attention output and $W_{O,\ell}^{(h)} \in \mathbb{R}^{d_{\text{model}} \times d_h}$ is the corresponding slice of the output projection matrix.

FFN neurons. For neuron r at layer ℓ , the component output is:

$$O_{\ell,r}^F(\lambda) = w_{\ell,r}^{\text{down}} m_{\ell,r}(\lambda), \quad (14)$$

where $m_{\ell,r}(\lambda) \in \mathbb{R}$ is the scalar neuron activation and $w_{\ell,r}^{\text{down}} \in \mathbb{R}^{d_{\text{model}}}$ is the corresponding column of the down-projection matrix. Note that $O_{\ell,r}^F(\lambda)$ lies in the one-dimensional subspace $\text{span}(w_{\ell,r}^{\text{down}})$, making its trajectory inherently one-dimensional.

We use $O_c(\lambda)$ to denote either $O_{\ell,h}^A(\lambda)$ or $O_{\ell,r}^F(\lambda)$ when the component type is clear from context.

Downstream Gradient. The downstream gradient of the answer-shift metric with respect to component c is

$$\nabla_{O_c} g_{a,b}(\lambda) = \frac{\partial g_{a,b}(\lambda)}{\partial O_c(\lambda)}, \quad (15)$$

which captures the sensitivity of the logit difference to perturbations of O_c at influence strength λ . For an attention head, $\nabla_{O_{\ell,h}^A} g_{a,b} \in \mathbb{R}^{d_{\text{model}}}$; for an FFN neuron in scalar form, $\partial g_{a,b} / \partial m_{\ell,r} = \langle \Delta_{e_a, e_b}, w_{\ell,r}^{\text{down}} \rangle \in \mathbb{R}$.

C.1.2 Discrete Domain Notations

In the discrete setting, the external evidence strength is instantiated by the integer repetition count $N \geq 0$. The input sequence is constructed as $u^{(N)} = y \parallel x^N \parallel z$, where $y = (y_1, \dots, y_m)$ is the tokenized query, $x = (x_1, \dots, x_T)$ is the external evidence template asserting wrong answer e_a , and $z = (z_1, \dots, z_s)$ is the suffix containing the answer cue. We adopt the same notations of same definition for understanding.

Answer-Shift Metric. The discrete answer-shift metric is defined as

$$g_{a,b}(N) = \langle \Delta_{e_a, e_b}, E(N) \rangle, \quad (16)$$

where $E(N)$ denotes the final-layer residual-stream representation at the prediction position when processing $u^{(N)}$, and Δ_{e_a, e_b} is defined identically to the continuous case.

Component Outputs. For each component c , the discrete output $O_c(N)$ corresponds to $O_c(\lambda)$ evaluated at integer points. Specifically:

Attention heads. $O_{\ell,h}^A(N) = W_{O,\ell}^{(h)} a_{\ell,h}(N)$, where $a_{\ell,h}(N)$ is the attention output of head h at layer ℓ at the prediction position when processing $u^{(N)}$.

FFN neurons. $O_{\ell,r}^F(N) = w_{\ell,r}^{\text{down}} m_{\ell,r}(N)$, where $m_{\ell,r}(N)$ is the neuron activation when processing $u^{(N)}$.

C.2 Proof of Proposition 3.3

Define the total displacement of component c over $[\lambda_0, \lambda_1]$:

$$V_c = \int_{\lambda_0}^{\lambda_1} \dot{O}_c(\lambda) d\lambda = O_c(\lambda_1) - O_c(\lambda_0).$$

Assume $V_c \neq 0$ and let $d_c = V_c / |V_c|$. Decompose $\dot{O}_c(\lambda)$ into an endpoint-aligned component and a transverse residual:

$$\dot{O}_c(\lambda) = \alpha_c(\lambda) d_c + r_c(\lambda), \quad (17)$$

where

$$\alpha_c(\lambda) = \langle \dot{O}_c(\lambda), d_c \rangle, \quad \langle r_c(\lambda), d_c \rangle = 0. \quad (18)$$

By construction, the endpoint-aligned component integrates to the full displacement magnitude:

$$\begin{aligned} \int_{\lambda_0}^{\lambda_1} \alpha_c(\lambda) d\lambda &= \int_{\lambda_0}^{\lambda_1} \langle \dot{O}_c(\lambda), d_c \rangle d\lambda \\ &= \left\langle \int_{\lambda_0}^{\lambda_1} \dot{O}_c(\lambda) d\lambda, d_c \right\rangle = \langle V_c, d_c \rangle = |V_c|; \end{aligned} \quad (19)$$

and the transverse residual integrates to zero:

$$\begin{aligned} \int_{\lambda_0}^{\lambda_1} r_c(\lambda) d\lambda &= \int_{\lambda_0}^{\lambda_1} \dot{O}_c(\lambda) d\lambda \\ - \left(\int_{\lambda_0}^{\lambda_1} \alpha_c(\lambda) d\lambda \right) d_c &= V_c - |V_c| d_c = 0. \end{aligned} \quad (20)$$

Substituting into Definition 3.1:

$$\begin{aligned} I_c(\lambda_0, \lambda_1) &= \int_{\lambda_0}^{\lambda_1} \langle \nabla_{O_c} g_{a,b}(\lambda), \alpha_c(\lambda) d_c + r_c(\lambda) \rangle d\lambda \\ &= \int_{\lambda_0}^{\lambda_1} \alpha_c(\lambda) \langle \nabla_{O_c} g_{a,b}(\lambda), d_c \rangle d\lambda \\ &\quad + \int_{\lambda_0}^{\lambda_1} \langle \nabla_{O_c} g_{a,b}(\lambda), r_c(\lambda) \rangle d\lambda \\ &= \left\langle \frac{1}{|V_c|} \int_{\lambda_0}^{\lambda_1} \alpha_c(\lambda) \nabla_{O_c} g_{a,b}(\lambda) d\lambda, |V_c| d_c \right\rangle \\ &\quad + E_c^\perp \\ &= \langle \bar{G}_c, V_c \rangle + E_c^\perp, \end{aligned} \quad (21)$$

where

$$\bar{G}_c = \frac{1}{|V_c|} \int_{\lambda_0}^{\lambda_1} \alpha_c(\lambda) \nabla_{O_c} g_{a,b}(\lambda) d\lambda \quad (22)$$

is the movement-weighted average gradient and

$$E_c^\perp = \int_{\lambda_0}^{\lambda_1} \langle \nabla_{O_c} g_{a,b}(\lambda), r_c(\lambda) \rangle d\lambda \quad (23)$$

is the transverse error. This identity is exact.

Now applying Assumption 3.2:

(i) *Directional Stability*. By Assumption 3.2(1), $\dot{O}_c(\lambda)$ is aligned with V_c for a.e. $\lambda \in [\lambda_0, \lambda_1]$, so $|r_c(\lambda)|$ is negligible. Since $\int_{\lambda_0}^{\lambda_1} r_c(\lambda) d\lambda = 0$, the transverse error satisfies

$$|E_c^\perp| \leq \sup_{\lambda \in [\lambda_0, \lambda_1]} |\nabla_{O_c} g_{a,b}(\lambda)| \cdot \int_{\lambda_0}^{\lambda_1} |r_c(\lambda)| d\lambda \rightarrow 0, \quad (24)$$

reducing the exact identity to

$$I_c(\lambda_0, \lambda_1) = \langle \bar{G}_c, V_c \rangle. \quad (25)$$

(ii) *Gradient Stability*. By Assumption 3.2(2), there exists $\varepsilon_G > 0$ such that for any anchor point $\lambda_m \in [\lambda_0, \lambda_1]$,

$$\sup_{\lambda \in [\lambda_0, \lambda_1]} |\nabla_{O_c} g_{a,b}(\lambda) - \nabla_{O_c} g_{a,b}(\lambda_m)| \leq \varepsilon_G. \quad (26)$$

Letting $\hat{G}_c := \nabla_{O_c} g_{a,b}(\lambda_m)$, the deviation from the movement-weighted average is bounded by

$$|\bar{G}_c - \hat{G}_c| \leq \varepsilon_G, \quad (27)$$

yielding

$$\begin{aligned} |I_c(\lambda_0, \lambda_1) - \langle \hat{G}_c, O_c(\lambda_1) - O_c(\lambda_0) \rangle| \\ \leq \varepsilon_G |V_c|. \end{aligned} \quad (28)$$

As $\varepsilon_G \rightarrow 0$, we recover the stated result:

$$I_c(\lambda_0, \lambda_1) = \langle \hat{G}_c, O_c(\lambda_1) - O_c(\lambda_0) \rangle. \quad \square$$

C.3 Proof of Lemma 3.4

We prove that for every component c , the limit $O_c(\infty) := \lim_{N \rightarrow \infty} O_c(N)$ exists and is finite. The proof proceeds by establishing convergence of the full residual stream at the prediction position layer by layer via a series of intermediate results (Appendix C.3.5–C.3.9); component-level convergence then follows as a direct consequence (Appendix C.3.10). Throughout this subsection, we introduce a **independent** set of proof-internal notations for the layer-wise transformer computation; these are used exclusively within this proof and do not override the notations in Section 3 and Appendix C.1.

C.3.1 Proof-Internal Notations and Notation Bridge

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. Recall the input construction from Appendix C.1.2: $u^{(N)} = y \|x^N\|z$, where $y = (y_1, \dots, y_m)$ is the tokenized query, $x = (x_1, \dots, x_T)$ is the external evidence template, and $z = (z_1, \dots, z_s)$ is the suffix.

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

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

for a fixed suffix offset $c \in \{0, \dots, s-1\}$, so that τ_N always indexes the same suffix token z_{c+1} regardless of N . For copy $\kappa \in \{1, \dots, N\}$ and template index $j \in \{1, \dots, T\}$, 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:

$$E_{\ell+\frac{1}{2}}^{(N)}(t) = E_\ell^{(N)}(t) + \text{Attn}_\ell(\text{LN}_\ell^{\text{attn}}(E_\ell^{(N)}))(t), \quad (29)$$

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

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

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

. 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

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

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

C.3.2 Regularity Assumption

Assumption C.3.1 (Regularity of pre-norm Transformer blocks). Suppose that: (i) the residual streams are uniformly bounded: $\sup_{N,\ell,t} \|E_\ell^{(N)}(t)\| \leq C$ 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.

C.3.3 RoPE Rotation

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

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

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

C.3.4 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 C.3.5). We then exploit this structure at the base layer, where loop values are exactly periodic, to establish Cesàro convergence of attention head outputs (Appendix C.3.6). 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 C.3.7). 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 C.3.8). 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 C.3.9).

Notation. Recall from Appendix C.3.1 that $\tau_N := m + NT + c$ is the prediction position and $t_{\kappa,j} := m + (\kappa - 1)T + (j - 1)$ is 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 $\leq t_{\kappa,j}$ and is therefore independent of N for every $N \geq \kappa$.

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 $q_{\ell,h}(\tau) = W_q \text{LN}_\ell^{\text{attn}}(E_\ell^{(N)}(\tau))$ and $k_{\ell,h}(t) = W_k \text{LN}_\ell^{\text{attn}}(E_\ell^{(N)}(t))$. 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 $t \leq \tau$, the RoPE-rotated key is

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

The relative distance (for causal attention) is

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

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

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

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

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

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

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

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

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

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 C.3.2 (Finite exponential form and uniform approximability of RoPE logits). *Suppose that Assumption C.3.1 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

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

C.3.5 Proof of Proposition C.3.2

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

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

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

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

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

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

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

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

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

$$|s_{\ell,h}(\tau, t)| \leq M_s, \quad \forall \delta \in \mathbb{Z}_{\geq 0}.$$

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

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

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

$$\sup_{\delta \geq 0} |e^{s_{\ell,h}(\tau, t)} - \Phi_J(\delta)| \xrightarrow{J \rightarrow \infty} 0.$$

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 C.3.3 (Cesàro limits for base-layer attention head output). *Suppose that Assumption C.3.1 holds. $v_{0,h}(t)$ is periodic with period T at loop positions when $\ell = 0$. Then the limits*

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

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

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

C.3.6 Proof of Proposition C.3.3

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,

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

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

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

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

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

converges as $N \rightarrow \infty$. 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

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

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 C.3.2.

Non-loop contribution vanishes. Since $|s_{0,h}(\tau_N, t)| \leq M_s$ for all t (Proposition C.3.2),

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

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

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

We show $\{\mathcal{D}_{\tau_N}^{(N,j)}\}_{N \geq 1}$ 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 $J \geq J_0$,

$$\sup_{\delta \geq 0} |e^{s^{(j)}(\delta)} - \Phi_J^{(j)}(\delta)| < \varepsilon.$$

Then for any N ,

$$\begin{aligned} & \left| \mathcal{D}_{\tau_N}^{(N,j)} - \frac{1}{N} \sum_{\kappa=1}^N \Phi_J^{(j)}(\delta_\kappa) \right| \\ & \leq \sup_{\delta} |e^{s^{(j)}(\delta)} - \Phi_J^{(j)}(\delta)| < \varepsilon. \end{aligned}$$

We now show that $\frac{1}{N} \sum_{\kappa=1}^N \Phi_J^{(j)}(\delta_\kappa)$ converges as $N \rightarrow \infty$. 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,

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

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,

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

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

$$\begin{aligned} |\mathcal{D}_{\tau_N}^{(N,j)} - \mathcal{D}_{\tau_{N'}}^{(N',j)}| & \leq |\mathcal{D}_{\tau_N}^{(N,j)} - \frac{1}{N} \sum_{\kappa} \Phi_J^{(j)}(\delta_\kappa)| \\ & + \left| \frac{1}{N} \sum_{\kappa} \Phi_J^{(j)}(\delta_\kappa) - \zeta_0^{(J,j)} \right| \\ & + \left| \zeta_0^{(J,j)} - \frac{1}{N'} \sum_{\kappa} \Phi_J^{(j)}(\delta_\kappa) \right| \\ & + \left| \frac{1}{N'} \sum_{\kappa} \Phi_J^{(j)}(\delta_\kappa) - \mathcal{D}_{\tau_{N'}}^{(N',j)} \right| < 4\varepsilon. \end{aligned}$$

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

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

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

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

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

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

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

Therefore the limit

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

exists and is finite. Moreover, $e^{-M_s} \leq \mathcal{A}_{\tau_N}^{(N)} \leq e^{M_s}$ 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 $N \rightarrow \infty$. 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, $\mathcal{D}_{\tau_N}^{(N,j)} \rightarrow \mu_p^{(j)}$ and $N/(\tau_N + 1) \rightarrow 1/T$, so

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

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 ($|s_{0,h}(\tau_N, t)| \leq M_s$ and $\|v_{0,h}(t)\| \leq C_v$ for all t),

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

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

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

and hence the limit

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

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:

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

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

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

□

Theorem C.3.4 (Base-layer convergence at the prediction position). *Suppose that Assumption C.3.1 holds. At layer $\ell = 0$, the limits $A_0^{(\infty)} := \lim_{N \rightarrow \infty} A_0^{(N)}(\tau_N)$, $F_0^{(\infty)} := \lim_{N \rightarrow \infty} F_0^{(N)}(\tau_N)$, and $E_1^{(\infty)} := E_0^{(\infty)} + A_0^{(\infty)} + F_0^{(\infty)}$ all exist and are finite.*

C.3.7 Proof of Theorem C.3.4

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

Step 1: Multi-head attention output $A_0^{(N)}(\tau_N)$. By Proposition C.3.3, each attention head $h \in \{0, 1, \dots, H-1\}$ admits a finite limit

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

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

$$A_0^{(N)}(\tau_N) = W_O [a_{0,0}^{(N)}(\tau_N); \dots; a_{0,H-1}^{(N)}(\tau_N)].$$

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

$$A_0^{(\infty)}(\tau_N) := \lim_{N \rightarrow \infty} A_0^{(N)}(\tau_N) = W_O[a_{0,0}^{(\infty)}(\tau_N); \dots; a_{0,H-1}^{(\infty)}(\tau_N)]$$

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,

$$E_0^{(\infty)}(\tau_N) := \lim_{N \rightarrow \infty} E_0^{(N)}(\tau_N) = E_0^{(N)}(\tau_N)$$

is trivially well-defined.

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

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

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

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

The FFN output is defined as

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

Under Assumption C.3.1, LN_0^{ffn} is continuous on bounded sets, and FFN_0 is continuous on bounded sets. Since the residual streams are uniformly bounded (Assumption C.3.1 (i)), the convergence $E_{1/2}^{(N)}(\tau_N) \rightarrow E_{1/2}^{(\infty)}(\tau_N)$ propagates through both maps:

$$\begin{aligned} F_0^{(\infty)}(\tau_N) &:= \lim_{N \rightarrow \infty} F_0^{(N)}(\tau_N) = \\ &\text{FFN}_0(\text{LN}_0^{\text{ffn}}(E_{1/2}^{(\infty)}(\tau_N))) \end{aligned}$$

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 $N \rightarrow \infty$, so does their sum:

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

which is finite. This completes the proof. $\square$

Proposition C.3.5 (Base-layer asymptotic periodicity at loop positions). *Suppose that Assumption C.3.1 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 $j \in \{1, \dots, T\}$ 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 $N \rightarrow \infty$.*

C.3.8 Proof of Proposition C.3.5

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 $y\|x\|x\| \dots \|x\|z$ loop structure required for the Cesàro argument at layer $\ell = 1$.

Step 1: Loop-position notation and relation to τ_N . For copy $\kappa \in \{1, \dots, N\}$ and template index $j \in \{1, \dots, T\}$, 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

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

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

- *Loop positions from copies 1, ..., $\kappa-1$* : 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 C.3.3 (Appendix C.3.6, Steps 2–4). We verify the two key prerequisites explicitly:

- Fixed query vector.* At layer $\ell = 0$, the pre-RoPE query at any loop position $t_{\kappa,j}$ is $q_{0,h}(t_{\kappa,j}) = W_q \text{LN}_0^{\text{attn}}(E_0^{(N)}(t_{\kappa,j})) = W_q \text{LN}_0^{\text{attn}}(\text{embed}(x_j))$, 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 .
- 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 $\kappa \rightarrow \infty$ (within $N \geq \kappa$), 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

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

exists and depends on j (through the query $W_q \text{embed}(x_j)$ 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 $\kappa \geq \kappa_0(h, j)$,

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

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 :

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

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

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

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

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

Under Assumption C.3.1, LN_0^{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 $\Lambda := \Lambda_{\text{FFN}} \cdot \Lambda_{\text{LN}}$. Then

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

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

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

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

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

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 $N \rightarrow \infty$ to obtain

$$N - N_0 \xrightarrow{N \rightarrow \infty} \infty.$$

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 $y\|x\|x\| \dots \|x\|z$ 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 $N \rightarrow \infty$. This is precisely the structural condition required by the Cesàro convergence argument (Appendix C.3.6) to be reapplied at layer $\ell = 1$, enabling the layer-wise induction in Theorem C.3.6. $\square$

Theorem C.3.6 (Full Transformer convergence). *Suppose that Assumption C.3.1 holds. Then for every $\ell \in \{0, 1, \dots, L\}$, the limit $E_\ell^{(\infty)} := \lim_{N \rightarrow \infty} E_\ell^{(N)}(\tau_N)$ exists and is finite. In particular, $\lim_{N \rightarrow \infty} E_L^{(N)}(\tau_N) = E_L^{(\infty)}$.*

C.3.9 Proof of Theorem C.3.6

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

(C $_\ell$) *Convergence at the target:* $E_\ell^{(N)}(\tau_N) \xrightarrow{N \rightarrow \infty} E_\ell^{(\infty)}$ exists and is finite.

(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 $N \geq 2N_0$, 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), $E_0^{(N)}(\tau_N) = \text{embed}(z_{c+1})$ is independent of N , so (C $_0$) holds trivially. At loop positions, $E_0^{(N)}(t_{\kappa,j}) = \text{embed}(x_j)$ depends only on the template index j and not on the copy index κ , so (P $_0$) holds with $N_0 = 1$ and ε -periodicity replaced by exact periodicity.

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

Step 1: Attention head convergence at τ_N ((P $_\ell$) $\Rightarrow$ head limits). By (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

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

converges as $N \rightarrow \infty$. 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 (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 $j \in \{1, \dots, T\}$ satisfy

$$\|E_\ell^{(N)}(t_{\kappa,j}) - E_\ell^{(N)}(t_{\kappa',j})\| < \varepsilon'. \quad (30)$$

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

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

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 C.3.1), there exist constants $L_k, L_v > 0$ (depending only on the model weights and the uniform bound C) such that for all $\kappa > \kappa_0$,

$$\begin{aligned} \|k_{\ell,h}(t_{\kappa,j}) - k_{\ell,h}^{\text{ref}}(j)\| &\leq L_k \varepsilon', \\ \|v_{\ell,h}(t_{\kappa,j}) - v_{\ell,h}^{\text{ref}}(j)\| &\leq L_v \varepsilon'. \end{aligned} \quad (31)$$

Similarly, by (C_ℓ), $E_\ell^{(N)}(\tau_N)$ converges as $N \rightarrow \infty$. Since the map $E_\ell^{(N)}(\tau_N) \mapsto q_{\ell,h}(\tau_N) = W_q \text{LN}_\ell^{\text{attn}}(E_\ell^{(N)}(\tau_N))$ is continuous on bounded sets (as a composition of the Lipschitz map $\text{LN}_\ell^{\text{attn}}$ and the linear map W_q ; Assumption C.3.1), there exist an integer $N_q \geq 1$ and a vector $q_{\ell,h}^{\text{ref}} \in \mathbb{R}^{d_h}$ such that for all $N \geq N_q$,

$$\|q_{\ell,h}(\tau_N) - q_{\ell,h}^{\text{ref}}\| < \varepsilon'. \quad (32)$$

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

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

Since $R(\cdot)$ is orthogonal and all query and key vectors are uniformly bounded by Assumption C.3.1 (denote the bounds C_q, C_k), we obtain for every $\kappa > \kappa_0$ and $N \geq N_q$:

$$\begin{aligned} |s_{\ell,h}(\tau_N, t_{\kappa,j}) - s_{\ell,h}^{\text{ref}}(\tau_N, t_{\kappa,j})| &\leq \\ \frac{1}{\sqrt{d_h}} (C_q L_k + C_k) \varepsilon' &=: \eta \varepsilon'. \end{aligned} \quad (33)$$

Here the term $C_q L_k \varepsilon' / \sqrt{d_h}$ bounds the key perturbation (31) with query fixed, and $C_k \varepsilon' / \sqrt{d_h}$ bounds the query perturbation (32) with key fixed. Since all logits are bounded by M_s (Proposition C.3.2), the mean-value theorem gives

$$|e^{s_{\ell,h}(\tau_N, t_{\kappa,j})} - e^{s_{\ell,h}^{\text{ref}}(\tau_N, t_{\kappa,j})}| \leq e^{M_s} \eta \varepsilon', \quad (34)$$

for all $\kappa > \kappa_0$, $N \geq N_q$.

Key structural observation. For fixed j , the reference logit $s_{\ell,h}^{\text{ref}}(\tau_N, t_{\kappa,j})$ depends on κ only through the relative distance $\delta = \tau_N - t_{\kappa,j}$, since both $q_{\ell,h}^{\text{ref}}$ and $k_{\ell,h}^{\text{ref}}(j)$ are independent of κ . Therefore $s_{\ell,h}^{\text{ref}}(\tau_N, t_{\kappa,j})$ is a trigonometric polynomial in δ with fixed coefficients. Similarly, $v_{\ell,h}^{\text{ref}}(j)$ is constant over κ . Thus the pair $(s_{\ell,h}^{\text{ref}}, v_{\ell,h}^{\text{ref}})$ 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 $\mathcal{A}_{\tau_N}^{(N)}$). We show that

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

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

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

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

$$\begin{aligned} \frac{1}{\tau_N + 1} \sum_{t \in \mathcal{I}^{\text{non}} \cup \mathcal{I}^{\text{early}}} e^{s_{\ell,h}(\tau_N, t)} &\leq \\ \frac{(m + s + 1 + \kappa_0 T) e^{M_s}}{\tau_N + 1} &\xrightarrow{N \rightarrow \infty} 0. \end{aligned} \quad (35)$$

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

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

By (34), the per-position error is at most $e^{M_s} \eta \varepsilon'$, so

$$\begin{aligned} \left| \frac{1}{\tau_N + 1} \sum_{t \in \mathcal{I}^{\text{late}}} e^{s_{\ell,h}(\tau_N, t)} - \tilde{\mathcal{A}}_{\tau_N}^{(N)} \right| &= \\ &\leq \frac{|\mathcal{I}^{\text{late}}|}{\tau_N + 1} e^{M_s} \eta \varepsilon' \leq e^{M_s} \eta \varepsilon'. \end{aligned} \quad (36)$$

Reference normalizer converges. By the structural observation above, $\tilde{\mathcal{A}}_{\tau_N}^{(N)}$ is a Cesàro average of $e^{s_{\ell,h}^{\text{ref}}(\tau_N, \cdot)}$ 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 C.3.6, Steps 1–2) via an explicit index substitution. For each template index j , the late-loop positions are $t_{\kappa,j}$ with $\kappa \in \{\kappa_0 + 1, \dots, N\}$. Introduce the reversed index $\kappa' := N - \kappa + 1$, ranging over $\{1, \dots, N - \kappa_0\}$. 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 $q_{\ell,h}^{\text{ref}}$ and $k_{\ell,h}^{\text{ref}}(j)$ are both fixed, the reference logit $s_{\ell,h}^{\text{ref}}(\tau_N, t_{\kappa,j}) = s_{\ell,h}^{\text{ref},j}(\kappa' T + c_j)$ is a trigonometric polynomial in κ' with coefficients that are *independent of N* . Therefore, grouping the

sum by template index gives

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

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 $N \rightarrow \infty$, the prefactor satisfies $(N-\kappa_0)/(\tau_N+1) \rightarrow 1/T$ and the inner average converges by the same trigonometric Cesàro limit (Appendix C.3.6, Steps 1–2). Hence $\tilde{\mathcal{A}}_{\tau_N}^{(N)}$ converges as $N \rightarrow \infty$.

$\mathcal{A}_{\tau_N}^{(N)}$ **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 (35), there exists $N_1 \geq \max(2\kappa_0, N_q)$ such that for all $N \geq N_1$ the non-loop and early-loop contribution is less than $\varepsilon/5$. By the convergence of $\tilde{\mathcal{A}}_{\tau_N}^{(N)}$, there exists $N_2 \geq N_1$ such that for all $N, N' \geq N_2$, $|\tilde{\mathcal{A}}_{\tau_N}^{(N)} - \tilde{\mathcal{A}}_{\tau_{N'}}^{(N')}| < \varepsilon/5$.

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

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

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

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

Therefore $\{\mathcal{A}_{\tau_N}^{(N)}\}_{N \geq 1}$ is Cauchy and hence convergent. Moreover, $e^{-M_s} \leq \mathcal{A}_{\tau_N}^{(N)} \leq e^{M_s}$ for all N , so the limit

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

is finite and strictly positive.

Step 1d (Cesàro convergence of $\mathcal{B}_{\tau_N}^{(N)}$). An analogous argument establishes convergence of the value-weighted Cesàro average $\mathcal{B}_{\tau_N}^{(N)} :=$

$\frac{1}{\tau_N+1} \sum_{t=0}^{\tau_N} e^{s_{\ell,h}(\tau_N,t)} v_{\ell,h}(t)$. Define the *reference numerator*

$$\tilde{\mathcal{B}}_{\tau_N}^{(N)} := \frac{1}{\tau_N+1} \sum_{j=1}^T v_{\ell,h}^{\text{ref}}(j) \sum_{\kappa=\kappa_0+1}^N e^{s_{\ell,h}^{\text{ref}}(\tau_N, t_{\kappa,j})}.$$

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

$$\begin{aligned} &\|e^{s_{\ell,h}(\tau_N,t)} v_{\ell,h}(t) - e^{s_{\ell,h}^{\text{ref}}(\tau_N,i)} v_{\ell,h}^{\text{ref}}(j)\| \\ &\leq |e^{s_{\ell,h}(\tau_N,t)} - e^{s_{\ell,h}^{\text{ref}}(\tau_N,i)}| \|v_{\ell,h}(t)\| \\ &\quad + e^{s_{\ell,h}^{\text{ref}}(\tau_N,i)} \|v_{\ell,h}(t) - v_{\ell,h}^{\text{ref}}(j)\| \\ &\leq e^{M_s}\eta\varepsilon' \cdot C_v + e^{M_s} \cdot L_v \varepsilon' = \Gamma \varepsilon'. \end{aligned} \quad (37)$$

where $C_v := \sup_{N,i} \|v_{\ell,h}(t)\| < \infty$ (Assumption C.3.1) and $\Gamma := e^{M_s}(\eta C_v + L_v)$. Summing over $\mathcal{I}^{\text{late}}$ and dividing by $\tau_N + 1$:

$$\begin{aligned} &\left\| \frac{1}{\tau_N+1} \sum_{t \in \mathcal{I}^{\text{late}}} e^{s_{\ell,h}(\tau_N,t)} v_{\ell,h}(t) - \tilde{\mathcal{B}}_{\tau_N}^{(N)} \right\| \\ &\leq \Gamma \varepsilon'. \end{aligned} \quad (38)$$

The reference numerator $\tilde{\mathcal{B}}_{\tau_N}^{(N)}$ has exactly periodic values and trigonometric-polynomial logits with fixed coefficients, so it converges by the base-layer argument (Appendix C.3.6, Step 3). The non-loop and early-loop contributions vanish as in (35). The same Cauchy-sequence argument as in Step 1c (with $e^{M_s}\eta\varepsilon'$ replaced by $\Gamma\varepsilon'$) then yields

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

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

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

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

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

$$\begin{aligned} A_\ell^{(\infty)}(\tau_N) &:= \lim_{N \rightarrow \infty} A_\ell^{(N)}(\tau_N) = \\ &W_O[a_{\ell,0}^{(\infty)}(\tau_N); \dots; a_{\ell,H-1}^{(\infty)}(\tau_N)]. \end{aligned}$$

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

$$\begin{aligned} E_{\ell+1/2}^{(N)}(\tau_N) &= E_\ell^{(N)}(\tau_N) + A_\ell^{(N)}(\tau_N) \\ &\xrightarrow{N \rightarrow \infty} E_\ell^{(\infty)}(\tau_N) + A_\ell^{(\infty)}(\tau_N) \\ &=: E_{\ell+1/2}^{(\infty)}(\tau_N). \end{aligned}$$

Under Assumption C.3.1, $\text{LN}_\ell^{\text{ffn}}$ and FFN_ℓ are continuous on bounded sets, so

$$\begin{aligned} F_\ell^{(\infty)}(\tau_N) &:= \lim_{N \rightarrow \infty} F_\ell^{(N)}(\tau_N) \\ &= \text{FFN}_\ell(\text{LN}_\ell^{\text{ffn}}(E_{\ell+1/2}^{(\infty)}(\tau_N))) \end{aligned}$$

exists. Applying the residual connection:

$$\begin{aligned} E_{\ell+1}^{(\infty)}(\tau_N) &:= \lim_{N \rightarrow \infty} E_{\ell+1}^{(N)}(\tau_N) \\ &= E_{\ell+1/2}^{(\infty)}(\tau_N) + F_\ell^{(\infty)}(\tau_N), \end{aligned}$$

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 $E_{\ell+1}^{(N)}$ 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 C.3.8.

At $\ell = 0$, the proof of Proposition C.3.5 used two facts:

- (i) $E_0^{(N)}(t_{\kappa,j}) = E_0^{(N)}(t_{\kappa',j})$ 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 $\kappa, \kappa' > \kappa_0$,

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

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 $a_{\ell,h}^{(N)}(t_{\kappa,j})$ sums over all preceding positions $i \leq t_{\kappa,j}$, of which $(\kappa-1)T$ are loop tokens from copies $1, \dots, \kappa-1$ 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 $q_{\ell,h}(t_{\kappa,j}) = W_q \text{LN}_\ell^{\text{attn}}(E_\ell^{(N)}(t_{\kappa,j}))$. 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 $W_q \circ \text{LN}_\ell^{\text{attn}}$). Fix a reference query $q_{\ell,h}^{\text{ref},j} := q_{\ell,h}(t_{\kappa_{\text{ref}},j})$ for some $\kappa_{\text{ref}} > \kappa_0$.
- **Key/value perturbation and reference logit.** For the preceding loop positions, the same reference key/value vectors $k_{\ell,h}^{\text{ref}}(\cdot)$, $v_{\ell,h}^{\text{ref}}(\cdot)$ and the same perturbation bounds (31) apply. The reference logit $s_{\ell,h}^{\text{ref}}(t_{\kappa,j}, t_{\kappa',j'}) = \frac{1}{\sqrt{d_h}} \langle q_{\ell,h}^{\text{ref},j}, R(-(t_{\kappa,j} - t_{\kappa',j'})) k_{\ell,h}^{\text{ref}}(j') \rangle$ is a trigonometric polynomial in the relative distance with fixed coefficients (independent of κ and N), since both $q_{\ell,h}^{\text{ref},j}$ and $k_{\ell,h}^{\text{ref}}(j')$ 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 $\kappa \rightarrow \infty$ (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 $\kappa, \kappa' > \kappa_1(h, j)$,

$$\begin{aligned} &\|a_{\ell,h}^{(N)}(t_{\kappa,j}) - a_{\ell,h}^{(N)}(t_{\kappa',j})\| \\ &< \frac{\varepsilon}{4H \|W_O\| (1 + \Lambda_\ell)}, \end{aligned}$$

where $\Lambda_\ell := \Lambda_{\text{FFN},\ell} \cdot \Lambda_{\text{LN},\ell}$ 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 C.3.8, concatenation over heads, application of W_O , and propagation through the Lipschitz maps $\text{LN}_\ell^{\text{ffn}}$ and FFN_ℓ yield

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

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

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

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

exists and is finite, completing the proof. $\square$

C.3.10 Deriving Component-Level Convergence

We now derive the convergence of each component output $O_c(N)$ from Theorem C.3.6.

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

Attention heads. Fix a layer ℓ and a head h . By Theorem C.3.6, the head-wise limit $a_{\ell,h}^{(\infty)}(\tau_N) := \lim_{N \rightarrow \infty} a_{\ell,h}^{(N)}(\tau_N)$ exists and is finite. Since $a_{\ell,h}(N) = a_{\ell,h}^{(N)}(\tau_N)$ (notation bridge in Appendix C.3.1) and $W_{O,\ell}^{(h)}$ is a fixed linear map, we have

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

which exists and is finite.

FFN neurons. Fix a layer ℓ and a neuron index r . By Theorem C.3.6, the post-attention residual limit $E_{\ell+\frac{1}{2}}^{(\infty)} := \lim_{N \rightarrow \infty} E_{\ell+\frac{1}{2}}^{(N)}(\tau_N)$ exists and is finite. Under Assumption C.3.1, $\text{LN}_\ell^{\text{ffn}}$ is continuous on bounded sets. Since the residual streams are uniformly bounded (Assumption C.3.1(i)), the composition

$$\text{LN}_\ell^{\text{ffn}}(E_{\ell+\frac{1}{2}}^{(N)}(\tau_N)) \xrightarrow{N \rightarrow \infty} \text{LN}_\ell^{\text{ffn}}(E_{\ell+\frac{1}{2}}^{(\infty)})$$

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

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

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

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

which exists and is finite.

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

C.4 Convergence Across Different N^*

To empirically validate the convergence guaranteed by Lemma 3.4, we track the per-layer answer-shift contribution $g_{\ell,c}^{(N)}(e_a, e_b)$ across repetition counts $N \in \{0, 1, \dots, N^*\}$ with $N^* = 1000$, for each layer ℓ and component type $c \in \{\text{Attn}, \text{FFN}\}$. We declare a layer to be **converged at tolerance** α if its tail variation is small relative to its total displacement:

$$\frac{\max_{N \in [N_{\text{tail}}, N_{\text{max}}]} |g_{\ell,c}^{(N)} - g_{\ell,c}^{(N_{\text{max}})}|}{|g_{\ell,c}^{(N_{\text{max}})} - g_{\ell,c}^{(0)}| + \epsilon} \leq \alpha,$$

where $N_{\text{tail}} = 0.8 \cdot N_{\text{max}}$ and ϵ is a small numerical-stability constant. For each model–benchmark configuration, we report the percentage of layers satisfying this criterion at $\alpha \in \{0.1, 0.05, 0.01\}$.

Table 8: Layer-wise convergence of inference dynamics at $N^* = 1000$ on three benchmarks at various divergence tolerances (0.1, 0.05, 0.01). Values report the percentage of layers whose answer-shift trajectory has converged, computed separately for attention (Attn) and feed-forward (FFN) modules.

Model	0.1		0.05		0.01	
	Attn	FFN	Attn	FFN	Attn	FFN
<i>OpenBook QA</i>						
Falcon3-7B-Base	98.1	97.3	95.3	94.3	81.5	81.3
Mistral-7B-v0.1	97.5	97.5	93.9	96.0	78.9	90.9
Apollo-1-4B	99.0	99.3	95.5	96.3	79.7	78.5
Qwen3-4B	99.5	99.4	97.0	97.5	83.8	84.4
Qwen2.5-1.5B	93.0	87.0	85.3	76.5	47.6	33.9
Falcon3-3B-Base	99.0	98.5	97.0	97.0	86.7	88.1
<i>MINTAKA</i>						
Falcon3-7B-Base	98.7	99.0	96.6	97.4	78.6	80.7
Mistral-7B-v0.1	94.0	93.7	89.1	89.1	66.3	68.7
Apollo-1-4B	99.8	99.9	98.7	99.4	87.0	91.4
Qwen3-4B	99.5	99.7	98.6	99.0	86.8	90.4
Qwen2.5-1.5B	99.8	99.6	97.0	97.6	73.6	69.1
Falcon3-3B-Base	100	100	99.8	99.7	93.3	94.0
<i>Simple QA</i>						
Falcon3-7B-Base	100	99.9	99.9	99.4	84.7	86.3
Mistral-7B-v0.1	95.5	94.9	92.9	91.5	72.8	73.2
Apollo-1-4B	100	100	99.9	99.9	93.0	93.9
Qwen3-4B	100	100	99.9	99.9	92.7	94.1
Qwen2.5-1.5B	99.8	99.7	99.0	98.5	75.8	68.7
Falcon3-3B-Base	100	100	100	100	95.2	93.8

Across all six models and three benchmarks, the vast majority of layers satisfy the convergence criterion at the moderate tolerances $\alpha = 0.1$ and $\alpha = 0.05$, and most layers remain converged even under the strict tolerance $\alpha = 0.01$. Attention and FFN components exhibit closely matched convergence rates, consistent with Lemma 3.4. These

results provide direct empirical support for our theoretical claim that internal inference trajectories under repeated contextual assertions converge to stable, query-dependent limits rather than drifting unboundedly.