FORMATTING INSTRUCTIONS FOR ICLR 2027 CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoder-only Transformers consistently concentrate attention on the first token, a phenomenon known as *attention sink*. We trace its origin to a structural constraint we call the *relational invariance bottleneck*: causal attention has no neutral gear—row-stochasticity forces every head to move mass somewhere, while causality makes the first position the only universal sink. An attention matrix is a *universal relational invariance* if it preserves every input up to a token-shared shift; under this definition, we establish the exact intersection:

$$\mathcal{A}_{\text{causal}} \cap \mathcal{N}_{\text{univ}} = \{\mathbf{1}\boldsymbol{e}_1^\top\}.$$

Collapse onto token one is therefore not an incidental artifact but the only available invariance target. To quantify this mechanism, we introduce a *relational energy* measuring how much token-level variation a head's output carries, and prove that it is bounded by the head's distance to the invariance target:

$$\Gamma(\boldsymbol{P}, \boldsymbol{V}) \;\le\; d_{\text{inv}}(\boldsymbol{P}) \;\le\; 2\big(1 - \alpha_1(\boldsymbol{P})\big).$$

This separates cause from symptom: first-token concentration matters only insofar as it signals proximity to the invariance target that suppresses relational output. Motivated by this diagnosis, we propose *Adaptive Diagonal Target Editing* (ADTE), a lightweight per-head module that cancels the self-token component of each attention row and rescales the remaining cross-token signal, giving heads an alternative relational invariance route that no longer requires concentrating mass on the first token. Across two architectures, ADTE consistently lowers pretraining perplexity compared with all baselines.

1 INTRODUCTION

Large Language models(LLMs) has achieved great success in natural language processing (Yang et al., 2025)(Grattafiori et al., 2024), and their mathematical foundation is developing (Geshkovski et al., 2025). There are phenomenons in Transformers such that they occur no matter the scale of models and input, between which attention sink is still not theoretically understood in mathematics. attention sink refers that decoder-only Transformers consistently allocate a disproportionate share of attention mass to a small set of tokens — most often the first one — even when those tokens are semantically uninformative (Xiao et al.; Gu et al., 2024). This *attention sink* phenomenon is not confined to autoregressive LMs (Darcet et al., 2023; Rulli et al., 2025) and resists activation-based explanations: reducing outlier activations does not always remove it (Sun et al., 2026). Jacobian-based analyses (Barbero et al.) explain attention sinks functionally: they arise to mitigate over-mixing, but leave the structural characterization open—what feature of the causal attention constraint singles out the first token as the dominant sink?

Our answer: a relational invariance bottleneck. We show that this dominance is not incidental, but enforced by the intersection of two constraints: the causal row-stochastic simplex and value-agnostic relational invariance. Their intersection collapses to a singleton, yielding a bottleneck that uniquely selects the first token. Relational invariance refers to attention behavior whose net effect is to preserve the token-relational information in input features. For attention in a residual block, relational invariance need not mean a zero attention output: the residual path already preserves the input, so it suffices to preserve tokenwise *relations*. Adding a common vector to every token changes absolute

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

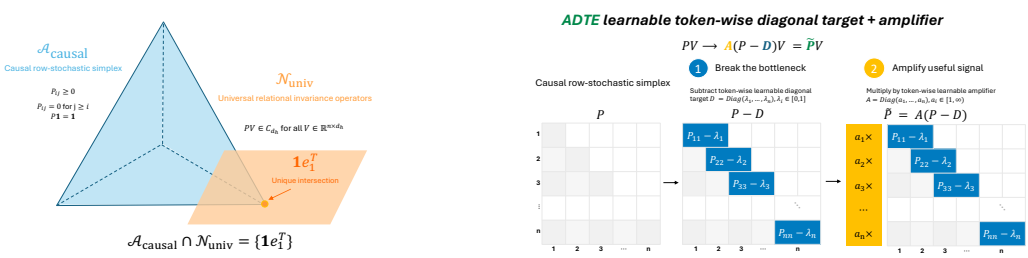


Figure 1: **Relational invariance bottleneck and ADTE.** (Left). Under causal softmax attention, collapsing mass onto the first token is the *only* value-agnostic way to achieve relational invariance—a structural bottleneck, not an artifact. (Right). ADTE breaks this bottleneck by subtracting a learnable diagonal target from the attention matrix and rescaling the remaining cross-token signal, giving each head an alternative relational invariance route without forcing mass onto token 1.

representations but preserves all pairwise differences. We prove that within the causal row-stochastic simplex the only *value-agnostic* relational-invariance operator is the first-token sink, which puts all attention mass on the first token. We call this singleton identity the *relational invariance bottleneck*.

A forward-pass measure. To make this mechanism measurable in a forward pass, we introduce relational energy, inspired by energy-based views of over-smoothing in graph neural networks (Guan & He, 2025; Guan & Shi, 2025). In our setting relational energy measures how much token-relational information remains in a head output after removing common-token shifts. We show that this relational output is controlled by the head's distance to the first-token invariance target. In turn, this distance is upper bounded by a simple observable quantity: the strength of the first-token sink. Thus the fundamental object is distance to the invariance target, while the first-token sink score provides a practical forward-pass proxy. We validate the bound on pretrained models, where heads cluster along the predicted sink–invariance frontier, and through a counterfactual probe that relocates the sink when the accessible invariance target is changed.

Adaptive Diagonal Target Editing. Building on this diagnosis, we propose *Adaptive Diagonal Target Editing* (ADTE), a lightweight attention module that supplies each head with a learnable diagonal self-cancellation path together with a tokenwise amplifier. ADTE provides an alternative relational invariance route at the self-token to step outside the causal row-stochastic simplex along controlled diagonal axes without suppressing non-self signal as scalar gates do (Bondarenko et al., 2023; Qiu et al.).

Perspective. We view attention sinks as more than an empirical pathology: they expose a structural design constraint of causal softmax attention. We turn this constraint into a principle for designing attention mechanisms beyond the row-stochastic simplex. We believe that this perspective would benefit the community in designing future attention mechanism.

2 Preliminaries and Related Work

2.1 Notation and Setup

We consider decoder-only Transformers with causal self-attention (Radford et al., 2018). Let $\boldsymbol{X}^\ell \in \mathbb{R}^{n \times d}$ denote the hidden representations at layer ℓ for a sequence of length n. For head h in layer ℓ, the causal attention matrix is

$$\boldsymbol{P}^{\ell,h} = \text{Softmax}\left(\frac{\boldsymbol{Q}^{\ell,h}(\boldsymbol{K}^{\ell,h})^\top}{\sqrt{d_h}} + \boldsymbol{M}\right) \in \mathbb{R}^{n \times n},$$

where $\boldsymbol{M}$ is the causal mask. Hence $\boldsymbol{P}^{\ell,h}$ is nonnegative, causal ($P_{ij}^{\ell,h} = 0$ for $j > i$), and row-stochastic ($\sum_j P_{ij}^{\ell,h} = 1$). These structural constraints are central to our analysis.

PreNorm Transformers. We focus on the PreNorm architecture adopted by most modern decoder-only LLMs (Xiong et al., 2020). A PreNorm block applies normalization such as Layernorm (Ba et al., 2016) or RMSnorm (Zhang & Sennrich, 2019) *before* each sub-layer:

$$\boldsymbol{Y}^{\ell+1} = \boldsymbol{X}^{\ell} + \mathrm{MHA}(\mathrm{Norm}(\boldsymbol{X}^{\ell})), \tag{1}$$

$$\boldsymbol{X}^{\ell+1} = \boldsymbol{Y}^{\ell+1} + \mathrm{FFN}(\mathrm{Norm}(\boldsymbol{Y}^{\ell+1})), \tag{2}$$

where Norm denotes normalization, MHA multi-head attention, and FFN the position-wise feed-forward network. For brevity we write $\boldsymbol{Z}^{\ell} := \mathrm{Norm}(\boldsymbol{X}^{\ell})$. The alternative Post-LN design, in which normalization is applied *after* the residual addition, is reviewed in Appendix A.1. Pre-LN enjoys better-behaved gradients (Wang et al., 2019) at initialization and is the default in models we study.

Attention sink and sink-rate metric. In decoder-only models, the first token often attracts disproportionate attention mass—a phenomenon termed *attention sink* (Xiao et al.; Gu et al., 2024; Barbero et al.). Token 1 is structurally privileged under causal masking because it is visible to every later position. We quantify this with the average first-token mass and the model-level sink rate:

$$\alpha_1^{\ell,h} := \frac{1}{n}\sum_{i=1}^{n} P_{i1}^{\ell,h}, \qquad \mathrm{Sink}_{\varepsilon} := \frac{1}{LN_h}\sum_{\ell=1}^{L}\sum_{h=1}^{N_h} \mathbb{I}\big(\alpha_1^{\ell,h} > \varepsilon\big),$$

where L is the number of layers, N_h heads per layer, and $\varepsilon > 0$ a threshold.

2.2 Related Work

Recent work on attention sink can be organized into three broad lines: utilization, mechanistic interpretation, and mitigation (Su et al., 2026). Since our paper is concerned with explanation and intervention, we focus on the latter two.

Mechanistic interpretations. Attention sink has been attributed to the softmax simplex constraint: heads that need to approximate no-op cannot produce exact zeros under row-stochastic normalization, driving outlier growth and sink patterns (Bondarenko et al., 2023; Huang et al., 2023). Also massive activation necessarily produce compression valley and hence attention sink and compression valley are two sides of one coin (Queipo-de Llano et al., 2025). Empirical studies trace when sink emerges during pre-training and link it to optimization dynamics, interpreting sink tokens as storing surplus attention akin to an implicit key bias (Gu et al., 2024). A Jacobian-based explanation posits that first-token sinks reduce cross-token sensitivity, helping deep Transformers avoid over-mixing (Barbero et al.). Another geometric explanation explains attention sink as the fundamental geometric principle that the establishment of reference frames (Ruscio et al., 2025). A perspective from backpropagation is also analyzed (Chen & Yao, 2026). Complementary views relate sinks to massive activations (Sun et al., 2026) and formalize their structural necessity under softmax normalization (Ran-Milo, 2026).

Mitigation strategies. Existing methods relax the coupling between row-stochastic simplex and relational invariance behavior. Gated attention adds an explicit gate so a head can suppress its update without collapsing mass onto sink tokens (Bondarenko et al., 2023; Qiu et al.). Modified-softmax and normalization-free variants reduce forced competition among tokens (Kaul et al.; Zuhri et al., 2025; Gu et al., 2024). Sink token method introduce a virtual token to absorb attention (Xiao et al.; Darcet et al., 2023; Zhang et al., 2026). Others make the implicit bias explicit via attention calibration or learnable key/value biases (Yu et al., 2024; Gu et al., 2024). Sink-aware pruning treats persistent BOS-dominant heads as redundancy markers (Sok et al., 2026).

3 The Relational Invariance Bottleneck of Causal Attention

3.1 The Relational Invariance Bottleneck

Prior work on over-mixing and rank collapse in deep Transformers (Barbero et al.) argues that some attention heads must suppress cross-token influence to remain useful at depth. In a residual block, however, the relevant notion of relational invariance is not a zero attention output: because the residual branch already preserves the input, a head may add the *same* vector to every token

without altering any tokenwise relation. We therefore define relational invariance at the level of token-relational information. We write d_h for the per-head value dimension, $\mathbf{1} \in \mathbb{R}^n$ for the all-ones vector, and $\boldsymbol{e}_i \in \mathbb{R}^n$ for the i-th standard basis vector.

Token-relational quotient. For any feature dimension m, define the common-token subspace

$$\mathcal{C}_m := \left\{ \mathbf{1}\boldsymbol{b}^\top : \boldsymbol{b} \in \mathbb{R}^m \right\} \subset \mathbb{R}^{n \times m}.$$

Any element of $\mathcal{C}_m$ has identical rows and therefore carries no token-specific relational information. Let

$$\boldsymbol{H} := \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

be the centering projector; then $\boldsymbol{H}\boldsymbol{X}$ is the canonical representative of the quotient space $\mathbb{R}^{n \times m}/\mathcal{C}_m$. This decomposition motivates the following formal notion of relational invariance.

Definition 1 (Exact relational invariance)**.** *A layer update $\boldsymbol{X}^\ell \mapsto \boldsymbol{X}^{\ell+1}$ is an* exact relational invariance *if $\boldsymbol{H}\boldsymbol{X}^{\ell+1} = \boldsymbol{H}\boldsymbol{X}^\ell$, or equivalently if there exists $\boldsymbol{b}^\ell \in \mathbb{R}^d$ such that*

$$\boldsymbol{X}^{\ell+1} = \boldsymbol{X}^\ell + \mathbf{1}(\boldsymbol{b}^\ell)^\top.$$

Definition 1 separates *identity in representation space* from *identity in token-relational space*: a common-token shift changes absolute hidden states but preserves all pairwise token differences, performing no token mixing.

Value-agnostic relational invariance under causal softmax. Consider a single-head residual update

$$\boldsymbol{X}^{\ell+1} = \boldsymbol{X}^\ell + \boldsymbol{P}\boldsymbol{V},$$

with causal attention matrix $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ and value matrix $\boldsymbol{V} \in \mathbb{R}^{n \times d_h}$. The update is relationally invariant precisely when $\boldsymbol{P}\boldsymbol{V} \in \mathcal{C}_{d_h}$. Although $\boldsymbol{P}$ and $\boldsymbol{V}$ are both computed from the same layer input $\boldsymbol{X}^\ell$, they do so through independently parameterized query–key and value projections. For relational invariance to persist across inputs rather than rely on coincidental cancellation between these two pathways, we require $\boldsymbol{P}\boldsymbol{V} \in \mathcal{C}_{d_h}$ for *every* $\boldsymbol{V}$; causality and row-stochasticity then pin this down to a single attention pattern.

Theorem 1 (Unique universal relational invariance under causal attention)**.** *Let $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ be causal and row-stochastic, then the following are equivalent:*

1. *$\boldsymbol{P}\boldsymbol{v} \in \operatorname{span}\{\mathbf{1}\}$ for every $\boldsymbol{v} \in \mathbb{R}^n$;*
2. *all rows of $\boldsymbol{P}$ are identical;*
3. *$\boldsymbol{P} = \mathbf{1}\boldsymbol{e}_1^\top$.*

Consequently, $\boldsymbol{P}\boldsymbol{V} \in \mathcal{C}_{d_h}$ for all $\boldsymbol{V} \in \mathbb{R}^{n \times d_h}$ if and only if $\boldsymbol{P} = \mathbf{1}\boldsymbol{e}_1^\top$.

The proof is given in Appendix B.1.[1]

The relational invariance bottleneck. To phrase Theorem 1 geometrically, let

$$\mathcal{A}_{\text{causal}} := \left\{ \boldsymbol{P} \in \mathbb{R}^{n \times n} : P_{ij} \geq 0,\ P_{ij} = 0 \text{ for } j > i,\ \boldsymbol{P}\mathbf{1} = \mathbf{1} \right\}$$

be the causal row-stochastic simplex and

$$\mathcal{N}_{\text{univ}} := \left\{ \boldsymbol{P} \in \mathbb{R}^{n \times n} : \boldsymbol{P}\boldsymbol{V} \in \mathcal{C}_{d_h} \text{ for all } \boldsymbol{V} \in \mathbb{R}^{n \times d_h} \right\}$$

be the set of universal relational invariance operators. Theorem 1 is equivalent to the singleton identity

$$\boxed{\mathcal{A}_{\text{causal}} \cap \mathcal{N}_{\text{univ}} = \left\{ \mathbf{1}\boldsymbol{e}_1^\top \right\}.}$$

[1]The universal quantifier "for all $\boldsymbol{V}$" is weaker than it may appear: as shown in Appendix B.2, if $\boldsymbol{P}\boldsymbol{V} \in \mathcal{C}_{d_h}$ holds merely on an open neighborhood of value matrices, then by linearity it already holds for every $\boldsymbol{V}$. Thus any relational invariance behavior that is stable under small perturbations of the values is automatically universal, and Theorem 1 is really a statement about robust, reusable relational invariance mechanisms rather than about contrived values.

Definition 2 (Relational invariance bottleneck)**.** *We call this singleton identity the* relational invariance bottleneck *of causal softmax attention: the only value-agnostic way for such a head to realize an exact relational invariance is to collapse all probability mass onto token* 1.

The relational invariance bottleneck identifies the empirically observed first-token attention sink (Xiao et al.; Gu et al., 2024) with a purely geometric object—the unique value-agnostic relational invariance point of the causal softmax simplex. Section 3.2 turns this binary characterization into a quantitative bound relating proximity to $\mathbf{1}\boldsymbol{e}_1^\top$ to a head's relational output.

3.2 Relational Energy and the Distance Bound

The relational invariance bottleneck is a binary statement: either $\boldsymbol{P} = \mathbf{1}\boldsymbol{e}_1^\top$ or not. To connect it to empirical measurements, we now develop a *forward-pass* quantity that scores how close a head is to the bottleneck point, and bound a head's relational output by it.

Relational energy and invariance target. Recall the centering projector $\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. The *relational energy* of a token matrix is the squared quotient norm

$$\mathcal{E}_{\mathrm{rel}}(\boldsymbol{X}) := \|\boldsymbol{H}\boldsymbol{X}\|_F^2,$$

which vanishes if and only if $\boldsymbol{X} \in \mathcal{C}_m$, i.e. all rows coincide. It therefore quantifies how far $\boldsymbol{X}$ lies from the relational-invariance subspace. In parallel, define the *invariance target* and its normalized squared distance from a causal row-stochastic $\boldsymbol{P}$:

$$\boldsymbol{P}_{\mathrm{inv}} := \mathbf{1}\boldsymbol{e}_1^\top, \qquad d_{\mathrm{inv}}(\boldsymbol{P}) := \frac{1}{n}\|\boldsymbol{P} - \boldsymbol{P}_{\mathrm{inv}}\|_F^2,$$

so that $d_{\mathrm{inv}}(\boldsymbol{P}) = 0$ exactly at the bottleneck point of Definition 2. The sink score $\alpha_1(\boldsymbol{P}) := \frac{1}{n}\sum_i P_{i1}$ is the quantity introduced in Section 2.1.

Theorem 2 (Distance to the invariance target controls relational output)**.** *For any causal row-stochastic $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ and any value matrix $\boldsymbol{V} \in \mathbb{R}^{n\times d_h}$, the* normalized relational-output score $\Gamma(\boldsymbol{P}, \boldsymbol{V}) := \mathcal{E}_{\mathrm{rel}}(\boldsymbol{P}\boldsymbol{V})/(n\,\mathcal{E}_{\mathrm{rel}}(\boldsymbol{V}) + \tau)$ *with $\tau > 0$ satisfies*

$$\Gamma(\boldsymbol{P}, \boldsymbol{V}) \;\le\; d_{\mathrm{inv}}(\boldsymbol{P}) \;\le\; 2\big(1 - \alpha_1(\boldsymbol{P})\big).$$

The proof is in Appendix B.3. Thus $d_{\mathrm{inv}}(\boldsymbol{P})$ is the primary theoretical object and $\alpha_1(\boldsymbol{P})$ an easily observable proxy.

Design implication. Theorem 2 shows that, within the causal row-stochastic simplex $\mathcal{A}_{\mathrm{causal}}$, value-agnostic suppression of relational output forces $\boldsymbol{P}$ toward $\mathbf{1}\boldsymbol{e}_1^\top$. To break this coupling the effective operator must step outside $\mathcal{A}_{\mathrm{causal}}$ while remaining close to $\boldsymbol{P}$ for stability.

4 Empirical Evidence for the Relational Invariance Bottleneck

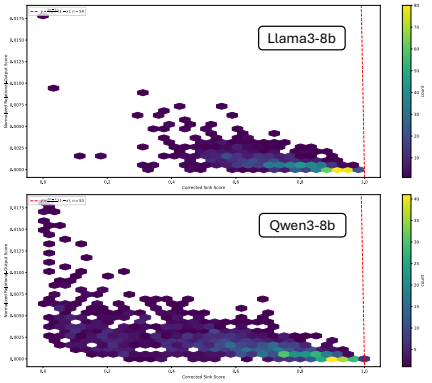
Theorems 1 and 2 yield two falsifiable predictions: (i) pretrained heads should lie below the sink–invariance envelope, and (ii) replacing $\mathbf{1}\boldsymbol{e}_1^\top$ with a different accessible invariance target should relocate the sink. We test both predictions in turn.

Throughout this section we use the normalized relational-output score $\Gamma^{\ell,h}$ from Theorem 2 together with the *corrected* sink score $\tilde{\alpha}_1^{\ell,h} := \frac{1}{n-1}\sum_{i=2}^{n} P_{i1}^{\ell,h} \in [0,1]$, which removes the trivial $P_{11} = 1$ under causal masking. Theorem 2 then gives the linear envelope $\Gamma^{\ell,h} \le \frac{2(n-1)}{n}(1 - \tilde{\alpha}_1^{\ell,h})$.

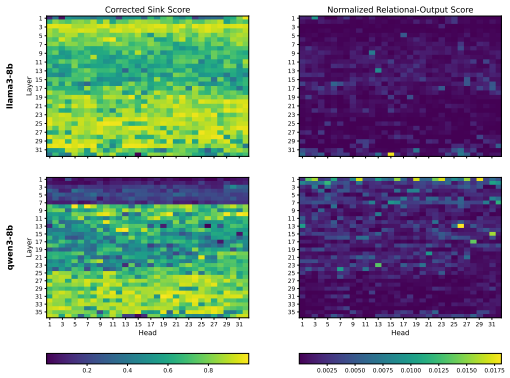
4.1 Heads Respect the Sink–invariance Frontier

We evaluate LLaMA3-8B (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) on C4 validation sequences and compute $(\tilde{\alpha}_1^{\ell,h}, \Gamma^{\ell,h})$ for every attention head. The head population should lie under the envelope above and, since Theorem 2 ties Γ to d_{inv} rather than to sink *semantics*, heads near the upper-right corner (high sink, high Γ) would falsify the mechanism. Figure 2(a) confirms the

envelope and shows density concentrating in the high-sink / low-Γ area; Figure 2(b) locates the trade-off at the head level rather than in pooled averages, with strong-sink blocks spatially coinciding with weak-Γ blocks.



(a) Hexbin density in the $(\tilde{\alpha}_1^{\ell,h}, \Gamma^{\ell,h})$ plane with the envelope $\Gamma \leq \frac{2(n-1)}{n}(1-\tilde{\alpha}_1)$.



(b) Aligned layer–head maps of $\tilde{\alpha}_1^{\ell,h}$ (left) and $\Gamma^{\ell,h}$ (right) per model.

Figure 2: **Pretrained heads occupy the predicted sink–invariance frontier.** (a) Over all heads of LLaMA3-8B and Qwen3-8B, the joint density respects the envelope of Theorem 2 and concentrates in the high-sink / low-output area: sink tokens pull the operator toward the unique value-agnostic target and suppress token-relational output. (b) Strong-sink blocks in the $\tilde{\alpha}_1$ map co-localize with weak-Γ blocks in the relational-output map.

4.2 A Counterfactual Invariance Target Relocates the Sink

If $\mathbf{1}\boldsymbol{e}_1^\top$ is the unique value-agnostic invariance point, replacing it with a different accessible target should relocate the sink. We train a matched pair of LLaMA-style models on C4 with

$$\boldsymbol{U}_{\text{base}}^{\ell,h} = \boldsymbol{P}^{\ell,h}\boldsymbol{V}^{\ell,h} \qquad \text{versus} \qquad \boldsymbol{U}_{\text{fix}}^{\ell,h} = \big(\boldsymbol{P}^{\ell,h} - \boldsymbol{I}_n\big)\boldsymbol{V}^{\ell,h},$$

so that the fixed variant achieves exact relational invariance at $\boldsymbol{P} = \boldsymbol{I}_n$ (diagonal concentration) rather than at $\boldsymbol{P} = \mathbf{1}\boldsymbol{e}_1^\top$. This rigid edit serves purely as a mechanism probe. We measure redirection via the diagonal concentration score $\alpha_{\text{diag}}^{\ell,h} := \frac{1}{n}\sum_i P_{ii}^{\ell,h}$.

Figure 3(a) shows that the fixed-diagonal model concentrates in regions with lower $\tilde{\alpha}_1$ and higher α_{diag}, while Figure 3(b) confirms that strong first-token sink blocks in the baseline fade and are replaced by diagonal concentration at the same positions. This provides causal evidence that the sink follows whichever value-agnostic invariance target the architecture makes accessible, rather than a token-1-specific artifact. The rigidity of this edit motivates the learnable relaxation developed next.

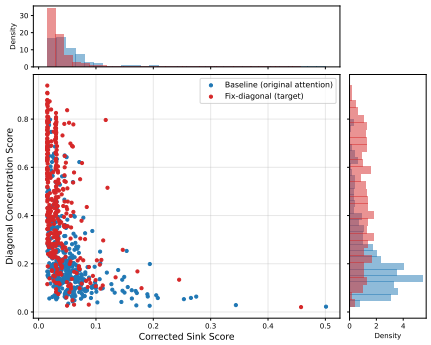
5 Adaptive Diagonal Target Editing

From theory to design. Theorem 1 shows that under causal row-stochastic softmax attention, the only value-agnostic invariance target is the first-token operator $\mathbf{1}\boldsymbol{e}_1^\top$; Theorem 2 further shows that moving toward this target suppresses relational output. Section 4.2 confirms the diagnosis by replacing $\boldsymbol{PV}$ with $(\boldsymbol{P}-\boldsymbol{I}_n)\boldsymbol{V}$: once the invariance target is redirected from the first token to the diagonal, first-token sink disappears.

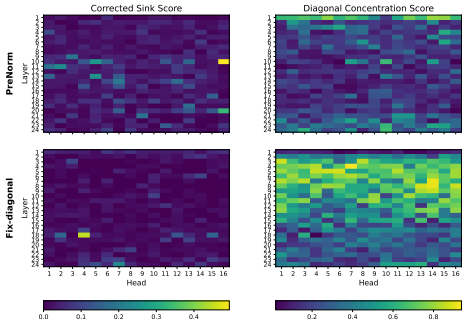
This motivates a learnable version of diagonal editing. Rather than modifying the attention distribution itself, *Adaptive Diagonal Target Editing (ADTE)* edits the residual operator acting on values:

$$\boldsymbol{PV} \longrightarrow \boldsymbol{A}(\boldsymbol{P} - \boldsymbol{D})\boldsymbol{V}$$

Here $\boldsymbol{D}$ applies tokenwise diagonal suppression and $\boldsymbol{A}$ rescales the remaining signal. The resulting operator stays close to $\boldsymbol{P}$ at initialization but can leave the causal row-stochastic simplex, opening relational invariance routes that are blocked under standard softmax attention.



(a) Distributions of attention heads of PreNorm model and Fix-diagonal model in the $(\tilde{\alpha}_1, \alpha_{\text{diag}})$ plane.



(b) Baseline (top) vs. fixed-diagonal (bottom) layer–head maps of $\tilde{\alpha}_1$ (left) and α_{diag} (right).

Figure 3: **A fixed diagonal target relocates the sink.** (a) Replacing the invariance target from $\mathbf{1}\boldsymbol{e}_1^\top$ to $\boldsymbol{I}_n$ shifts heads toward lower $\tilde{\alpha}_1$ and higher α_{diag}, confirming that the sink follows the accessible target. (b) Layer–head maps show that first-token sink blocks fade while diagonal concentration increases at the same positions.

5.1 Learnable Diagonal Target

We adopt the PreNorm notation of Section 2.1, with $\boldsymbol{Z}^\ell = \mathrm{Norm}(\boldsymbol{X}^\ell)$ and $\boldsymbol{V}^{\ell,h} = \boldsymbol{Z}^\ell \boldsymbol{W}_V^{\ell,h}$. We write σ for the elementwise normalization function, $\mathrm{diag}(\boldsymbol{M}) \in \mathbb{R}^n$ for the diagonal vector of $\boldsymbol{M}$, and $\mathrm{Diag}(\boldsymbol{v})$ for the diagonal matrix built from a vector $\boldsymbol{v}$.

With projection parameters $\boldsymbol{w}_D^{\ell,h} \in \mathbb{R}^d$, ADTE predicts a tokenwise diagonal target and applies it directly to the residual operator:

$$\boldsymbol{\lambda}^{\ell,h} = \sigma\Big(\boldsymbol{Z}^\ell \boldsymbol{w}_D^{\ell,h}\Big) \in [0,1]^n, \qquad \boldsymbol{D}^{\ell,h} = \mathrm{Diag}(\boldsymbol{\lambda}^{\ell,h}), \qquad \bar{\boldsymbol{U}}^{\ell,h} = \big(\boldsymbol{P}^{\ell,h} - \boldsymbol{D}^{\ell,h}\big)\boldsymbol{V}^{\ell,h}.$$

Because $\boldsymbol{D}^{\ell,h}$ is learned independently of $\mathrm{diag}(\boldsymbol{P}^{\ell,h})$, the edited diagonal coefficient $P_{ii}^{\ell,h} - \lambda_i^{\ell,h}$ can be positive, zero, or negative. Thus ADTE can interpolate between the baseline update ($\boldsymbol{\lambda}^{\ell,h} \equiv \mathbf{0}$), self-cancellation ($\boldsymbol{\lambda}^{\ell,h} > 0$) ; Appendix C.2 discusses the three resulting regimes in detail. Compared with the fixed counterfactual $(\boldsymbol{P} - \boldsymbol{I}_n)\boldsymbol{V}$ of Section 4.2, ADTE keeps the diagonal edit tokenwise and learnable rather than imposing unit subtraction on every row.

5.2 Tokenwise Amplifier

Subtracting $\boldsymbol{D}^{\ell,h}$ reduces the row sums by $D_{ii}^{\ell,h}$, which can attenuate useful cross-token signal. To decouple *what is cancelled* from *how strongly the remaining signal is expressed*, we introduce a tokenwise amplifier with parameters $\boldsymbol{w}_A^{\ell,h} \in \mathbb{R}^d$:

$$\boldsymbol{a}^{\ell,h} = 1 + \mathrm{softplus}\Big(\boldsymbol{Z}^\ell \boldsymbol{w}_A^{\ell,h}\Big) \in [1,\infty)^n, \quad \boldsymbol{A}^{\ell,h} = \mathrm{Diag}(\boldsymbol{a}^{\ell,h}), \quad \boldsymbol{U}^{\ell,h} = \boldsymbol{A}^{\ell,h}\big(\boldsymbol{P}^{\ell,h} - \boldsymbol{D}^{\ell,h}\big)\boldsymbol{V}^{\ell,h}.$$

We restrict $\boldsymbol{a}^{\ell,h} \geq 1$ so that the amplifier only compensates the row-sum loss $1 - D_{ii}^{\ell,h}$ rather than dampens the output; a scalar-gating-style attenuation pathway is explicitly avoided (Proposition 1). The multi-head attention block is

$$\boldsymbol{X}^{\ell+1} = \boldsymbol{X}^\ell + \Big(\boldsymbol{U}^{\ell,1} \,\|\, \cdots \,\|\, \boldsymbol{U}^{\ell,N_h}\Big)\boldsymbol{W}_O^\ell,$$

and causality is preserved because $\boldsymbol{A}^{\ell,h}$ and $\boldsymbol{D}^{\ell,h}$ are diagonal. The effective operator $\tilde{\boldsymbol{P}}^{\ell,h} := \boldsymbol{A}^{\ell,h}(\boldsymbol{P}^{\ell,h} - \boldsymbol{D}^{\ell,h})$ generally has row sums $\neq 1$; this is intentional. Once the residual operator is allowed to leave the causal row-stochastic simplex, ADTE can realize the the signed diagonal relational invariance route discussed in Appendix C.4.

Table 1: **Main pretraining results.** Validation perplexity and mechanistic profile. ADTE lowers PPL while redirecting the first-token sink to diagonal cancellation.

Standard PreNorm backbone

Scale	Method	PPL↓	$\mathrm{Sink}_\varepsilon \downarrow$	$\bar{\alpha}_1 \downarrow$	$\bar{\alpha}_{\mathrm{diag}} \uparrow$
130M	PreNorm	25.19(0.10)	0.1261	0.0494	0.2226
	Gated	24.81(0.12)	0.0835	0.0352	0.1253
	ADTE	24.54(0.07)	0.0732	0.0277	0.2785
250M	PreNorm	20.89(0.12)	0.1130	0.0485	0.1973
	Gated	20.74(0.14)	0.0663	0.0305	0.1221
	ADTE	20.20(0.05)	0.0781	0.0300	0.2841
350M	PreNorm	19.07(0.16)	0.0916	0.0420	0.2174
	Gated	19.10(0.16)	0.0649	0.0348	0.1113
	ADTE	18.60(0.06)	0.0588	0.0238	0.2886

LayerNorm Scaling (LNS) backbone

Scale	Method	PPL↓	$\mathrm{Sink}_\varepsilon \downarrow$	$\bar{\alpha}_1 \downarrow$	$\bar{\alpha}_{\mathrm{diag}} \uparrow$
130M	LNS	24.95(0.10)	0.1935	0.0729	0.2029
	Gated	24.78(0.12)	0.1134	0.0464	0.1475
	ADTE	24.58(0.05)	0.0896	0.0323	0.3114
250M	LNS	20.48(0.10)	0.2147	0.0780	0.1656
	Gated	20.23(0.08)	0.1237	0.0483	0.1350
	ADTE	20.18(0.04)	0.1203	0.0446	0.2594
350M	LNS	18.30(0.12)	0.2339	0.0910	0.1701
	Gated	18.17(0.10)	0.1119	0.0466	0.1525
	ADTE	17.94(0.05)	0.1048	0.0396	0.2854

6 Experimental Results

6.1 Experimental Setup

Models and Data. We train LLaMA-style decoder-only Transformers from scratch at scales $\{130\mathrm{M}, 250\mathrm{M}, 350\mathrm{M}\}$ on C4 (Radford et al., 2019). All models share the same optimizer, learning-rate schedule, and per-scale token budget; architectural and training details are in Appendix D.1. For larger models, see Appendix D.5. We defer the downstream evaluation to Appendix D.7.

Baseline Selection. At each scale, we compare three variants under two normalization schemes. Under RMSNorm, we consider: (i) **PreNorm**, the standard Transformer baseline; (ii) **Gated Attention** (Bondarenko et al., 2023), which applies a token-wise scalar gate to the attention output; and (iii) **ADTE**, the full method $\boldsymbol{A}(\boldsymbol{P}-\boldsymbol{D})\boldsymbol{V}$ defined in Section 5. We additionally evaluate the same variants under LayerNorm Scaling (LNS) (Sun et al.).

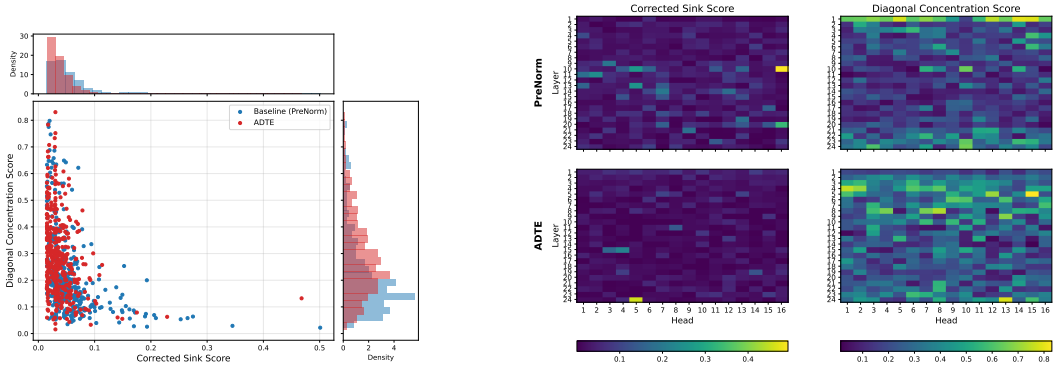
Evaluation Protocols. We run each experiment with three random seeds and report the mean and standard deviation of validation perplexity (PPL) on held-out C4 as the primary quality metric. For mechanistic analysis, we use the corrected sink score $\tilde{\alpha}_1^{\ell,h}$ (Section 4), the diagonal concentration $\alpha_{\mathrm{diag}}^{\ell,h} = \frac{1}{n}\sum_i P_{ii}^{\ell,h}$, and the sink-head fraction $\mathrm{Sink}_\varepsilon$ with $\varepsilon = 0.1$.

6.2 Main Results

Table 1 reports validation perplexity together with the mechanistic profile of each method at the three model scales. A successful intervention should reduce $\mathrm{Sink}_\varepsilon$ and $\bar{\alpha}_1$, *not* by uniformly suppressing attention, which would also suppress useful attention, but by substituting another relational invariance route, visible as larger $\bar{\alpha}_{\mathrm{diag}}$ and (for ADTE) larger $\bar{c}_{\mathrm{diag}}$. As shown in Table 1, ADTE consistently outperforms the baselines, while mitigating attention sink and increasing $\bar{\alpha}_{\mathrm{diag}}$.

6.3 Component Ablation

We isolate the roles of the two ADTE components at the 250M scale, holding all other training settings fixed. Table 2 compares five operators: the baseline $\boldsymbol{PV}$; the fixed-diagonal $(\boldsymbol{P}-\boldsymbol{I}_n)\boldsymbol{V}$ of Section 4.2, which caps cancellation at full identity; **ADTE-A** ($\boldsymbol{APV}$), amplification without a invariance target; **ADTE-D** ($(\boldsymbol{P}-\boldsymbol{D})\boldsymbol{V}$), learned diagonal cancellation without the amplifier; and the full **ADTE** ($\boldsymbol{A}(\boldsymbol{P}-\boldsymbol{D})\boldsymbol{V}$). The four-way contrast separates three effects: whether a relational invariance route is offered at all, whether the route is rigid or learned, and whether signal scale is recovered. For all models, we additionally report the average cancelled diagonal mass, defined as $c_{\mathrm{diag}} = \frac{1}{LN_h}\sum_{\ell,h} c_{\mathrm{diag}}^{\ell,h}$, where $c_{\mathrm{diag}}^{\ell,h} = \frac{1}{n}\sum_i (P_{ii}^{\ell,h} - \lambda_i^{\ell,h})$. For the fixed-diagonal, we set $\lambda_i^{\ell,h} = 1$, while for models without $\boldsymbol{D}$ we use $\lambda_i^{\ell,h} = 0$. As shown in Table 2, ablating both modules $\boldsymbol{A}$ and $\boldsymbol{D}$ leads to a noticeable degradation in performance. The metrics $\bar{\alpha}_1$ and $\mathrm{Sink}_\varepsilon$ indicate that each



(a) Distributions of attention heads of ADTE and PreNorm models in the $(\tilde{\alpha}_1, \alpha_{\text{diag}})$ plane.

(b) PreNorm (top) vs. ADTE (bottom) layer–head maps of $\tilde{\alpha}_1$ (left) and α_{diag} (right).

Figure 4: **ADTE redirects first-token sink to diagonal self-cancellation.** (a) Heads shift toward lower $\tilde{\alpha}_1$ and higher α_{diag} under ADTE. (b) Layer–head maps confirm that sink blocks (top-left) fade in ADTE (bottom-left) while diagonal concentration (bottom-right) increases at the same positions.

module individually contributes to mitigating attention sink. Furthermore, $\bar{c}_{\text{d}}$ shows that ADTE-D is more effective than ADTE-A at redirecting attention toward the diagonal.

A matrix-level visualization showing how $\boldsymbol{D}$, $\boldsymbol{P} - \boldsymbol{D}$, and $\boldsymbol{A}(\boldsymbol{P} - \boldsymbol{D})$ reshape a representative attention pattern is given as a case study in Appendix D.3.

Table 2: **Component ablation (250M).**

Method	PPL↓	Sink_ε↓	$\bar{\alpha}_1$↓	$\bar{\alpha}_{\text{d}}$↑	$\bar{c}_{\text{d}}$↑
PreNorm	20.89	.113	.049	.197	.197
Fixed-diagonal	21.62	.060	.024	.429	-.571
ADTE-A	20.36	.087	.035	.207	.207
ADTE-D	20.58	.082	.033	.297	-.235
ADTE	**20.20**	.078	.030	.284	-.231

6.4 Mechanistic Interpretation

Aggregate metrics summarize the redirection but cannot show whether it happens at the same heads that used first-token sink in the baseline. Figure 4 provides a head-resolved view at the end of training. Panel (a) plots attention heads of PreNorm and ADTE in the $(\tilde{\alpha}_1^{\ell,h}, \alpha_{\text{diag}}^{\ell,h})$ plane; the ADTE model concentrates in regions with higher diagonal concentration and lower corrected sink scores. Panel (b) gives the matched layer–head maps of $\tilde{\alpha}_1^{\ell,h}$ and $\alpha_{\text{diag}}^{\ell,h}$ for PreNorm and ADTE side-by-side; bright sink blocks in the baseline fade in ADTE while corresponding positions light up in the diagonal-concentration map.

Takeaway

ADTE lowers perplexity while redirecting relational invariance from first-token sink to diagonal self-cancellation at the same heads.

7 Conclusion

We reframed attention sink as a structural consequence of causal softmax attention: a residual head that needs to do nothing relationally has no explicit way to abstain, since every row must place unit mass somewhere and token 1 is the only position visible to every query under the causal mask. Formalizing relational invariance as identity on the token-relational quotient, we proved that $\boldsymbol{P}_{\text{inv}} = \mathbf{1}\boldsymbol{e}_1^\top$ is the unique value-agnostic relational-invariance operator — the *relational invariance bottleneck* — and that a head's relational output is controlled by its distance to this target. Pretrained heads cluster along the predicted sink–invariance frontier, sink emerges as relational output collapses, and a fixed diagonal edit relocates the sink from token 1 toward the diagonal as predicted. Building on this diagnosis, *Adaptive Diagonal Target Editing* supplies an alternative relational invariance route that lets a head stay still without silencing non-self signal.

AI USE STATEMENT

In this work, we used generative AI tools for polishing writing. We have not used generative AI tools for research ideation, retrieval, discovery, draft sections of the paper, generating synthetic datasets and proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work. We checked the LLM-polished writing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The complete anonymous experimental code is provided in the supplementary material. A complete proof of the theorems is included in the Appendix B and Appendix C.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M Bronstein. Why do llms attend to the first token? In *Second Conference on Language Modeling*.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems*, 36:75067–75096, 2023.

Chen Chen and Lai Wei. Post-layernorm is back: Stable, expressive, and deep. *arXiv preprint arXiv:2601.19895*, 2026.

Yihong Chen and Quanming Yao. Attention sinks induce gradient sinks. *arXiv preprint arXiv:2603.17771*, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025.

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

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024.

Weiqi Guan and Junlin He. Solving oversmoothing in gnns via nonlocal message passing: Algebraic smoothing and depth scalability. *arXiv preprint arXiv:2512.08475*, 2025.

Weiqi Guan and Zihao Shi. Measuring over-smoothing beyond dirichlet energy. *arXiv preprint arXiv:2512.06782*, 2025.

Zhongzhan Huang, Mingfu Liang, Jinghui Qin, Shanshan Zhong, and Liang Lin. Understanding self-attention mechanism via dynamical system perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1412–1422, 2023.

Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. From attention to activation: Unraveling the enigmas of large language models. In *The Thirteenth International Conference on Learning Representations*.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. *arXiv preprint arXiv:2510.06477*, 2025.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Yuval Ran-Milo. Attention sinks are provably necessary in softmax transformers: Evidence from trigger-conditional tasks. *arXiv preprint arXiv:2603.11487*, 2026.

Maximo Eduardo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane, and Alessio Devoto. Attention sinks in diffusion language models. *arXiv preprint arXiv:2510.15731*, 2025.

Valeria Ruscio, Umberto Nanni, Fabrizio Silvestri, et al. What are you sinking? a geometric approach on attention sink. In *NeurIPS 2025-The 39th Annual Conference on Neural Information Processing Systems*, 2025.

Jaewon Sok, Jewon Yeom, Seonghyeon Park, Jeongjae Park, and Taesup Kim. Garbage attention in large language models: Bos sink heads and sink-aware pruning. *arXiv preprint arXiv:2601.06787*, 2026.

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, Qingyao Yang, Yuxuan Sun, Rui Yang, Chao Zhang, et al. Attention sink in transformers: A survey on utilization, interpretation, and mitigation. *arXiv preprint arXiv:2604.10098*, 2026.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. *arXiv preprint arXiv:2603.05498*, 2026.

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1810–1822, 2019.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International conference on machine learning*, pp. 10524–10533. PMLR, 2020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Zihou Zhang, Zheyong Xie, Li Zhong, Haifeng Liu, Yao Hu, and Shaosheng Cao. One token is enough: Improving diffusion language models with a sink token. *arXiv preprint arXiv:2601.19657*, 2026.

Zayd MK Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. Softpick: No attention sink, no massive activations with rectified softmax. *arXiv preprint arXiv:2504.20966*, 2025.

A Additional Background

A.1 PostNorm Architecture

The PostNorm design normalizes *after* the residual addition rather than before. A PostNorm block reads as follows:

$$\boldsymbol{Y}^{\ell+1} = \mathrm{Norm}\big(\boldsymbol{X}^{\ell} + \mathrm{MHA}(\boldsymbol{X}^{\ell})\big), \tag{3}$$

$$\boldsymbol{X}^{\ell+1} = \mathrm{Norm}\big(\boldsymbol{Y}^{\ell+1} + \mathrm{FFN}(\boldsymbol{Y}^{\ell+1})\big). \tag{4}$$

PostNorm was used in the original Transformer (Vaswani et al., 2017) but tends to suffer from gradient instability at greater depths. Most modern decoder-only LLMs (including all models studied in this paper) adopt PreNorm; therefore, our formal analysis focuses on PreNorm blocks (Eqs. 1–2).

We consider two architectures. The first follows a PostNorm structure with RMSNorm used as the normalization module, which we still refer to as PostNorm. The second is Keel (Chen & Wei, 2026). The formulation of Keel is given by

$$\boldsymbol{Y}^{\ell+1} = \mathrm{Norm}\big(\sqrt{2L}\boldsymbol{X}^{\ell} + \mathrm{MHA}(\mathrm{Norm}(\boldsymbol{X}^{\ell}))\big), \tag{5}$$

$$\boldsymbol{X}^{\ell+1} = \mathrm{Norm}\big(\sqrt{2L}\boldsymbol{Y}^{\ell+1} + \mathrm{FFN}(\mathrm{Norm}(\boldsymbol{Y}^{\ell+1}))\big). \tag{6}$$

We also use RMSNorm to implement Keel. The results are reported in Table 6.

A.2 Our position

Our work is closest to the Jacobian-based anti-overmixing perspective (Barbero et al.), but differs in two key ways: (i) we formulate relational invariance directly in the representation space as identity on token-relational information, rather than through infinitesimal sensitivity; (ii) we show that under causal row-stochastic simplex, first token sink is the unique *value-agnostic* realization of such relational invariance; (iii) we propose Adaptive Diagonal Target Editing (ADTE), which uses a learnable diagonal self-cancellation path and a tokenwise amplifier to provide an alternative route to relational invariance beyond the causal row-stochastic simplex. These distinctions shift the explanation from a sensitivity-based diagnosis to an operator-level characterization, and from that characterization to a concrete attention modification.

B Proofs of Theorems

B.1 Proof of Theorem 1

Proof. (1) ⇒ (2). Let $\boldsymbol{p}_i^\top$ denote the i-th row of $\boldsymbol{P}$. If $\boldsymbol{P}\boldsymbol{v} \in \mathrm{span}\{\mathbf{1}\}$ for every $\boldsymbol{v}$, then for any i, j and any $\boldsymbol{v} \in \mathbb{R}^n$,

$$(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top \boldsymbol{v} = (\boldsymbol{P}\boldsymbol{v})_i - (\boldsymbol{P}\boldsymbol{v})_j = 0.$$

Hence $\boldsymbol{p}_i = \boldsymbol{p}_j$ for all i, j, so all rows are identical.

(2) ⇒ (3). If all rows are identical, there exists $\boldsymbol{p}^\top$ such that $\boldsymbol{P} = \mathbf{1}\boldsymbol{p}^\top$. Row-stochasticity gives $\boldsymbol{p}^\top\mathbf{1} = 1$. By causality, the first row can only attend to token 1, so the support of $\boldsymbol{p}$ must be contained in $\{1\}$. Therefore $\boldsymbol{p} = \boldsymbol{e}_1$ and $\boldsymbol{P} = \mathbf{1}\boldsymbol{e}_1^\top$.

(3) ⇒ (1). If $\boldsymbol{P} = \mathbf{1}\boldsymbol{e}_1^\top$, then for any $\boldsymbol{v} \in \mathbb{R}^n$,

$$\boldsymbol{P}\boldsymbol{v} = \mathbf{1}(\boldsymbol{e}_1^\top \boldsymbol{v}) = v_1\mathbf{1} \in \mathrm{span}\{\mathbf{1}\}.$$

□

B.2 Robust relational invariance implies universal relational invariance

The following lemma formalizes the footnote in Section 3.1: any relational invariance behavior that is stable on an open set of value matrices is automatically universal, and therefore (under causality and row-stochasticity) already forces $\boldsymbol{P} = \mathbf{1}\boldsymbol{e}_1^\top$.

Lemma 1 (Robust relational invariance implies universal relational invariance)**.** *Fix an attention matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$. Suppose there exist $\boldsymbol{V}_0 \in \mathbb{R}^{n\times m}$ and $\varepsilon > 0$ such that*

$$\boldsymbol{P}\boldsymbol{V} \in \mathcal{C}_m \qquad \text{for every } \boldsymbol{V} \text{ satisfying } \|\boldsymbol{V} - \boldsymbol{V}_0\|_F < \varepsilon.$$

Then

$$\boldsymbol{P}\boldsymbol{V} \in \mathcal{C}_m \qquad \text{for every } \boldsymbol{V} \in \mathbb{R}^{n\times m}.$$

Consequently, if $\boldsymbol{P}$ is causal and row-stochastic, then $\boldsymbol{P} = \mathbf{1}\boldsymbol{e}_1^\top$.

Proof. Let $\boldsymbol{\Delta} \in \mathbb{R}^{n\times m}$ be arbitrary. Choose $\eta > 0$ small enough so that $\eta\|\boldsymbol{\Delta}\|_F < \varepsilon$. Then both $\boldsymbol{V}_0$ and $\boldsymbol{V}_0 + \eta\boldsymbol{\Delta}$ lie in the assumed neighborhood, hence

$$\boldsymbol{P}\boldsymbol{V}_0 \in \mathcal{C}_m, \qquad \boldsymbol{P}(\boldsymbol{V}_0 + \eta\boldsymbol{\Delta}) \in \mathcal{C}_m.$$

Since $\mathcal{C}_m$ is a linear subspace,

$$\eta\boldsymbol{P}\boldsymbol{\Delta} = \boldsymbol{P}(\boldsymbol{V}_0 + \eta\boldsymbol{\Delta}) - \boldsymbol{P}\boldsymbol{V}_0 \in \mathcal{C}_m.$$

Thus $\boldsymbol{P}\boldsymbol{\Delta} \in \mathcal{C}_m$. Because $\boldsymbol{\Delta}$ was arbitrary, the relational invariance condition holds for all value matrices. The final claim follows from Theorem 1. □

B.3 Proof of Theorem 2

Proof. Let

$$\boldsymbol{P}_{\text{inv}} = \mathbf{1}\boldsymbol{e}_1^\top.$$

Since $\boldsymbol{H}\mathbf{1} = 0$, we have

$$\boldsymbol{H}\boldsymbol{P}_{\text{inv}} = 0.$$

Moreover, because both $\boldsymbol{P}$ and $\boldsymbol{P}_{\text{inv}}$ are row-stochastic,

$$(\boldsymbol{P} - \boldsymbol{P}_{\text{inv}})\mathbf{1} = 0.$$

For any value matrix $\boldsymbol{V}$, decompose

$$\boldsymbol{V} = \boldsymbol{H}\boldsymbol{V} + \mathbf{1}\bar{\boldsymbol{v}}^\top, \qquad \bar{\boldsymbol{v}} := \frac{1}{n}\boldsymbol{V}^\top\mathbf{1}.$$

Then

$$(\boldsymbol{P} - \boldsymbol{P}_{\text{inv}})\mathbf{1}\bar{\boldsymbol{v}}^\top = 0,$$

and hence

$$\boldsymbol{H}\boldsymbol{P}\boldsymbol{V} = \boldsymbol{H}(\boldsymbol{P} - \boldsymbol{P}_{\text{inv}})\boldsymbol{V} = \boldsymbol{H}(\boldsymbol{P} - \boldsymbol{P}_{\text{inv}})\boldsymbol{H}\boldsymbol{V}.$$

Therefore,

$$\mathcal{E}_{\text{rel}}(\boldsymbol{P}\boldsymbol{V}) = \|\boldsymbol{H}\boldsymbol{P}\boldsymbol{V}\|_F^2 \le \|\boldsymbol{P} - \boldsymbol{P}_{\text{inv}}\|_F^2\|\boldsymbol{H}\boldsymbol{V}\|_F^2 = \|\boldsymbol{P} - \boldsymbol{P}_{\text{inv}}\|_F^2\mathcal{E}_{\text{rel}}(\boldsymbol{V}).$$

Dividing by $n\mathcal{E}_{\text{rel}}(\boldsymbol{V}) + \tau$ gives

$$\Gamma(\boldsymbol{P}, \boldsymbol{V}) \le \frac{\|\boldsymbol{P} - \boldsymbol{P}_{\text{inv}}\|_F^2\mathcal{E}_{\text{rel}}(\boldsymbol{V})}{n\mathcal{E}_{\text{rel}}(\boldsymbol{V}) + \tau} \le \frac{1}{n}\|\boldsymbol{P} - \boldsymbol{P}_{\text{inv}}\|_F^2 = d_{\text{inv}}(\boldsymbol{P}).$$

It remains to relate $d_{\text{inv}}(\boldsymbol{P})$ to the sink score. Since $\boldsymbol{P}_{\text{inv}} = \mathbf{1}\boldsymbol{e}_1^\top$,

$$\|\boldsymbol{P} - \boldsymbol{P}_{\text{inv}}\|_F^2 = \sum_{i=1}^{n}\left[(1 - P_{i1})^2 + \sum_{j=2}^{n} P_{ij}^2\right].$$

Because $P_{ij} \ge 0$ and

$$\sum_{j=2}^{n} P_{ij} = 1 - P_{i1},$$

we have

$$\sum_{j=2}^{n} P_{ij}^2 \le (1 - P_{i1})^2.$$

Thus

$$\|\boldsymbol{P} - \boldsymbol{P}_{\text{sil}}\|_F^2 \le 2\sum_{i=1}^{n}(1 - P_{i1})^2 \le 2\sum_{i=1}^{n}(1 - P_{i1}) = 2n(1 - \alpha_1(\boldsymbol{P})).$$

Dividing by n yields

$$d_{\text{inv}}(\boldsymbol{P}) \le 2(1 - \alpha_1(\boldsymbol{P})).$$

□

C Discussion: Scalar-Gating Comparison and Practical Parameterization

C.1 From heads to the attention layer

The head-level bound of Theorem 2 aggregates to the full multi-head attention update. Writing

$$\Delta \boldsymbol{X}^{\ell}_{\mathrm{att}} = \Big(\boldsymbol{P}^{\ell,1}\boldsymbol{V}^{\ell,1} \,\|\, \cdots \,\|\, \boldsymbol{P}^{\ell,N_h}\boldsymbol{V}^{\ell,N_h}\Big)\boldsymbol{W}^{\ell}_{O},$$

the output projection $\boldsymbol{W}^{\ell}_{O}$ mixes feature coordinates but not token positions, so it commutes with $\boldsymbol{H}$.

Corollary 1 (Layer-level relational energy bound)**.** *For the multi-head attention update above,*

$$\mathcal{E}_{\mathrm{rel}}\big(\Delta \boldsymbol{X}^{\ell}_{\mathrm{att}}\big) \le \|\boldsymbol{W}^{\ell}_{O}\|_2^2 \sum_{h=1}^{N_h} \mathcal{E}_{\mathrm{rel}}\big(\boldsymbol{P}^{\ell,h}\boldsymbol{V}^{\ell,h}\big) \le \|\boldsymbol{W}^{\ell}_{O}\|_2^2 \sum_{h=1}^{N_h} \|\boldsymbol{P}^{\ell,h} - \boldsymbol{P}_{\mathrm{inv}}\|_F^2\, \mathcal{E}_{\mathrm{rel}}(\boldsymbol{V}^{\ell,h}).$$

Proof. Let $\boldsymbol{U}^{\ell} = \big(\boldsymbol{P}^{\ell,1}\boldsymbol{V}^{\ell,1} \,\|\, \cdots \,\|\, \boldsymbol{P}^{\ell,N_h}\boldsymbol{V}^{\ell,N_h}\big)$. Since $\boldsymbol{H}$ acts on the token dimension and $\boldsymbol{W}^{\ell}_{O}$ on the feature dimension, $\boldsymbol{H}\Delta \boldsymbol{X}^{\ell}_{\mathrm{att}} = \boldsymbol{H}\boldsymbol{U}^{\ell}\boldsymbol{W}^{\ell}_{O}$, so

$$\mathcal{E}_{\mathrm{rel}}(\Delta \boldsymbol{X}^{\ell}_{\mathrm{att}}) = \|\boldsymbol{H}\boldsymbol{U}^{\ell}\boldsymbol{W}^{\ell}_{O}\|_F^2 \le \|\boldsymbol{W}^{\ell}_{O}\|_2^2\, \|\boldsymbol{H}\boldsymbol{U}^{\ell}\|_F^2.$$

Concatenation along the feature dimension adds Frobenius energy, $\|\boldsymbol{H}\boldsymbol{U}^{\ell}\|_F^2 = \sum_h \mathcal{E}_{\mathrm{rel}}(\boldsymbol{P}^{\ell,h}\boldsymbol{V}^{\ell,h})$, and the second inequality follows from Theorem 2. □

Thus $\boldsymbol{W}^{\ell}_{O}$ can amplify but not create relational variation: layer-level relational output is controlled by the per-head distances $\|\boldsymbol{P}^{\ell,h} - \boldsymbol{P}_{\mathrm{inv}}\|_F^2$ and, through the second inequality of Theorem 2, by the per-head sink scores $\alpha_1^{\ell,h}$.

C.2 Three regimes of diagonal editing

For query token i, ADTE produces

$$\bar{u}_i^{\ell,h} = \sum_{j \ne i} P_{ij}^{\ell,h}\boldsymbol{v}_j^{\ell,h} + \big(P_{ii}^{\ell,h} - d_i^{\ell,h}\big)\boldsymbol{v}_i^{\ell,h}.$$

Thus ADTE edits only the self coefficient while preserving the geometry of the non-self interactions. This yields three regimes: baseline ($d_i^{\ell,h} \approx 0$), self-cancellation ($d_i^{\ell,h} > 0$). Hence self-concentration becomes a relational invariance route only after diagonal editing, when $(\boldsymbol{P} - \boldsymbol{D})\boldsymbol{V} \approx 0$.

C.3 Row-level expressivity vs. scalar gating

A natural baseline is scalar gated attention, $u_i^{\mathrm{gate}} = \alpha_i \sum_j p_{ij}\boldsymbol{v}_j$ with $\alpha_i \in [0,1]$, which couples self-target suppression with global signal suppression: to remove the self-target $p_{ii}\boldsymbol{v}_i$, the gate must set $\alpha_i = 0$, which also kills every cross-token interaction. The following proposition isolates the row-level difference between ADTE-style edits and scalar gating.

Proposition 1 (Selective cancellation with scale recovery)**.** *Fix one row $\boldsymbol{p}_i^{\top}$ of an attention matrix with $\boldsymbol{p}_i^{\top}\mathbf{1} = 1$ and $p_{ij} \ge 0$. Let*

$$\tilde{\boldsymbol{p}}_i^{\top} = a_i\big(\boldsymbol{p}_i^{\top} - d_i\boldsymbol{e}_i^{\top}\big), \qquad d_i \ge 0,\ a_i > 0,$$

be the effective row induced by a diagonal edit with scale. Then:

1. Ratio preservation. *For $j \ne i$, $\tilde{p}_{ij} = a_i p_{ij}$, so $\tilde{p}_{ij}/\tilde{p}_{ij'} = p_{ij}/p_{ij'}$ for $j, j' \ne i$.*
2. Row-sum control. $\sum_j \tilde{p}_{ij} = a_i(1 - d_i)$*; choosing $a_i = (1 - d_i)^{-1}$ restores unit row sum.*
3. Exact self-target removal. *Setting $d_i = p_{ii}$ and $a_i = (1 - p_{ii})^{-1}$ yields $\tilde{p}_{ii} = 0$ and $\tilde{p}_{ij} = p_{ij}/(1 - p_{ii})$ for $j \ne i$.*
4. Scalar gating cannot do this. *If $\hat{\boldsymbol{p}}_i^{\top} = \alpha_i\boldsymbol{p}_i^{\top}$ with $\alpha_i \in [0,1]$ and $p_{ii} > 0$, then $\hat{p}_{ii} = 0$ forces $\alpha_i = 0$ and hence $\hat{p}_{ij} = 0$ for all j.*

Proof. Items (1)–(2) follow directly from $\tilde{\boldsymbol{p}}_i^\top = a_i(\boldsymbol{p}_i^\top - d_i \boldsymbol{e}_i^\top)$. For item (3), substituting $d_i = p_{ii}$ and $a_i = (1-p_{ii})^{-1}$ gives $\tilde{\boldsymbol{p}}_i^\top = (1-p_{ii})^{-1}(\boldsymbol{p}_i^\top - p_{ii}\boldsymbol{e}_i^\top)$, hence $\tilde{p}_{ii} = 0$ and $\tilde{p}_{ij} = p_{ij}/(1-p_{ii})$ for $j \neq i$. For item (4), $\hat{p}_{ii} = \alpha_i p_{ii} = 0$ with $p_{ii} > 0$ forces $\alpha_i = 0$, whence $\hat{p}_{ij} = 0$ for all j. □

Proposition 1 describes what an ADTE-style row edit *can* realize. ADTE parameterizes $\boldsymbol{\lambda}^{\ell,h}$ and $\boldsymbol{a}^{\ell,h}$ independently, so it is not constrained to the row-sum-restoring choice $a_i = (1-d_i)^{-1}$; that choice is a special case and ADTE is strictly more general. Diagonal cancellation and tokenwise amplification therefore play orthogonal roles: the former removes a tokenwise self-target component, the latter preserves or amplifies the informative cross-token interactions that scalar gating would otherwise squash. Note that the proposition is stated at the level of a single row $\boldsymbol{p}_i^\top$; for the full ADTE operator $\boldsymbol{A}^{\ell,h}(\boldsymbol{P}^{\ell,h} - \boldsymbol{D}^{\ell,h})$, the same analysis applies row-by-row with $(d_i, a_i) = (\lambda_i^{\ell,h},\, a_i^{\ell,h})$.

C.4 Signed diagonal editing opens a relational invariance route

The following proposition gives a concise sufficient condition for ADTE to approach relational invariance.

Proposition 2 (Signed diagonal editing creates a relational invariance route)**.** *Let $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ be causal and row-stochastic, $\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{d})$ with $\boldsymbol{d} \in [0,1)^n$, $\boldsymbol{A} = \mathrm{Diag}(\boldsymbol{a})$ with $\boldsymbol{a} \in \mathbb{R}^n_{>0}$, and $\boldsymbol{M}_{\mathrm{ADTE}} := \boldsymbol{A}(\boldsymbol{P} - \boldsymbol{D})$. Then for any $\boldsymbol{V} \in \mathbb{R}^{n\times d_h}$,*

$$\mathcal{E}_{\mathrm{rel}}(\boldsymbol{M}_{\mathrm{ADTE}}\boldsymbol{V}) \le \|\boldsymbol{M}_{\mathrm{ADTE}}\|_F^2 \, \|\boldsymbol{V}\|_F^2, \qquad \|\boldsymbol{M}_{\mathrm{ADTE}}\|_F^2 = \sum_{i=1}^{n} a_i^2 \Big[(P_{ii} - d_i)^2 + \sum_{j\neq i} P_{ij}^2\Big].$$

Consequently, if $\boldsymbol{a}$ is bounded and $P_{ii} \to d_i$ and $P_{ij} \to 0$ for every $j \neq i$, then $\boldsymbol{M}_{\mathrm{ADTE}} \to 0$ and the residual update $\boldsymbol{X}^{\ell+1} = \boldsymbol{X}^{\ell} + \boldsymbol{M}_{\mathrm{ADTE}}\boldsymbol{V}$ approaches exact relational invariance for every $\boldsymbol{V}$. (In ADTE, $a_i = 1 + \mathrm{softplus}(\cdot)$ is always finite for finite inputs, so the boundedness condition is satisfied in practice.)

Proof. Since $\boldsymbol{H}$ is an orthogonal projector with $\|\boldsymbol{H}\|_2 \le 1$,

$$\mathcal{E}_{\mathrm{rel}}(\boldsymbol{M}_{\mathrm{ADTE}}\boldsymbol{V}) = \|\boldsymbol{H}\boldsymbol{M}_{\mathrm{ADTE}}\boldsymbol{V}\|_F^2 \le \|\boldsymbol{M}_{\mathrm{ADTE}}\boldsymbol{V}\|_F^2 \le \|\boldsymbol{M}_{\mathrm{ADTE}}\|_F^2 \|\boldsymbol{V}\|_F^2.$$

The i-th row of $\boldsymbol{M}_{\mathrm{ADTE}}$ is $a_i(\boldsymbol{p}_i^\top - d_i\boldsymbol{e}_i^\top)$, whose squared norm is $a_i^2[(P_{ii}-d_i)^2 + \sum_{j\neq i} P_{ij}^2]$. Summing over rows gives the stated expression, and the limiting claim follows because every row norm vanishes, so $\boldsymbol{M}_{\mathrm{ADTE}}\boldsymbol{V} \to 0 \in \mathcal{C}_{d_h}$. □

Parameter and compute overhead. Per layer, ADTE adds two tokenwise projections per head, $(\boldsymbol{w}_D^{\ell,h})$ and $(\boldsymbol{w}_A^{\ell,h})$. The additional parameter count is linear in the model width, $O(N_h\, d)$ per layer, rather than quadratic in d as in the main attention projections. Aggregated over L layers the overhead is $O(L\, N_h\, d)$, which in the models used in this paper amounts to a small fraction of the total parameter budget (below 1% in every configuration; the exact ratios are reported in Table 3). The extra compute is also linear in sequence length and width, $O(n\, N_h\, d)$, negligible compared with the $O(n^2 d)$ attention computation.

Table 3: **Ratio of additional parameters of ADTE**

	130M	250M	350M	1B
D	0.0823%	0.1189%	0.1066%	0.0918%
A	0.0823%	0.1189%	0.1066%	0.0918%
total	0.1647%	0.2379%	0.2133%	0.1836%

Relation to Theorem 2. The bound $\Gamma(\boldsymbol{P}, \boldsymbol{V}) \le d_{\mathrm{inv}}(\boldsymbol{P}) \le 2(1 - \alpha_1(\boldsymbol{P}))$ is stated for causal row-stochastic $\boldsymbol{P}$. After applying ADTE, the effective operator $\boldsymbol{A}^{\ell,h}(\boldsymbol{P}^{\ell,h} - \boldsymbol{D}^{\ell,h})$ is not row-stochastic, so neither inequality applies directly to the post-edit operator. This is precisely the feature the method exploits: shrinking the head-level relational energy $\mathcal{E}_{\mathrm{rel}}(\boldsymbol{U}^{\ell,h})$ no longer requires approaching $\mathbf{1}\boldsymbol{e}_1^\top$, because the geometric constraint of Theorem 1 no longer applies outside the simplex.

D Additional Experiments

D.1 Detailed Experimental Setup

Table 4 summarizes the architectural hyper-parameters for the three model scales used throughout the paper.

Table 4: **Model configurations.**

	130M	250M	350M	1B
Layers (L)	12	24	24	24
Hidden dim (d)	768	768	1024	2048
Heads (N_h)	12	16	16	32
Head dim (d_h)	64	48	64	64
FFN dim	2048	2560	2736	5461
Vocab size	32000	32000	32000	32000
Context length	1024	1024	1024	1024
Token budget	2.2B	3.9B	6.0B	5.5B

Optimizer and schedule. All models are trained with Adam under a cosine learning-rate schedule with a linear warm-up over the first $2000, 4000, 6000$ steps for $130M, 250M, 350M$ models respectively, and a peak learning rate of 1×10^{-3}. The batch size is $64, 128, 128$ sequences of length 1024 for $130M, 250M, 350M$ models respectively. Gradient clipping is set to 0.0.

Data. Training data comes from the English split of C4. We use the T5 tokenizer with a vocabulary of 32000 tokens. Validation perplexity is reported on a held-out shard of C4 that is disjoint from training.

Activation function for the diagonal target. In all experiments, the element-wise activation σ used to produce the diagonal cancellation coefficients $\boldsymbol{\lambda}^{\ell,h}$ is implemented as

$$\sigma(x) = \frac{\mathrm{ReLU}(x)}{1 + \mathrm{ReLU}(x)},$$

which maps $\mathbb{R} \to [0, 1)$ while being exactly zero for $x \leq 0$, ensuring a clean identity initialization when the projection weights start at zero.

Computer resources. We conduct experiments on 1 NVIDIA B200 GPU(s) (179.1 GiB per GPU)

D.2 Temporal emergence of sink during pretraining

Section 4 reports the sink–invariance frontier cross-sectionally on pretrained checkpoints. Here we provide a temporal view: if first-token sink is the default relational invariance route rather than a token-1-specific preference, then its formation should be coupled to the collapse of relational output across training, not independent of it.

We pretrain a LLaMA-style 350M model on C4 and log $\big(\tilde{\alpha}_1^{\ell,h}(t), \Gamma^{\ell,h}(t)\big)$ at checkpoints spanning the full training budget. Rather than plotting the two quantities on separate axes, we render each head as a trajectory in the phase plane $(\tilde{\alpha}_1, \Gamma)$ with color encoding the training step (Figure 6). Heads that eventually become strong sinks are not born as sinks: they drift from the low-$\tilde{\alpha}_1$ / moderate-Γ region toward the high-$\tilde{\alpha}_1$ / low-Γ corner, and remain there. Heads that retain nontrivial cross-token aggregation do not make this excursion. The coupling between sink growth and relational-output decay is therefore a dynamical signature, consistent with Theorem 2 but not implied by any purely semantic account of token 1.

D.3 Matrix-level case study: $\boldsymbol{P}$, $\boldsymbol{D}$, $\boldsymbol{P} - \boldsymbol{D}$, $\boldsymbol{A}(\boldsymbol{P} - \boldsymbol{D})$

Figure 7 visualizes how the two ADTE components reshape a representative attention matrix from a strong-sink head of a trained model. We show (i) the attention $\boldsymbol{P}^{\ell,h}$, (ii) the learned diagonal target

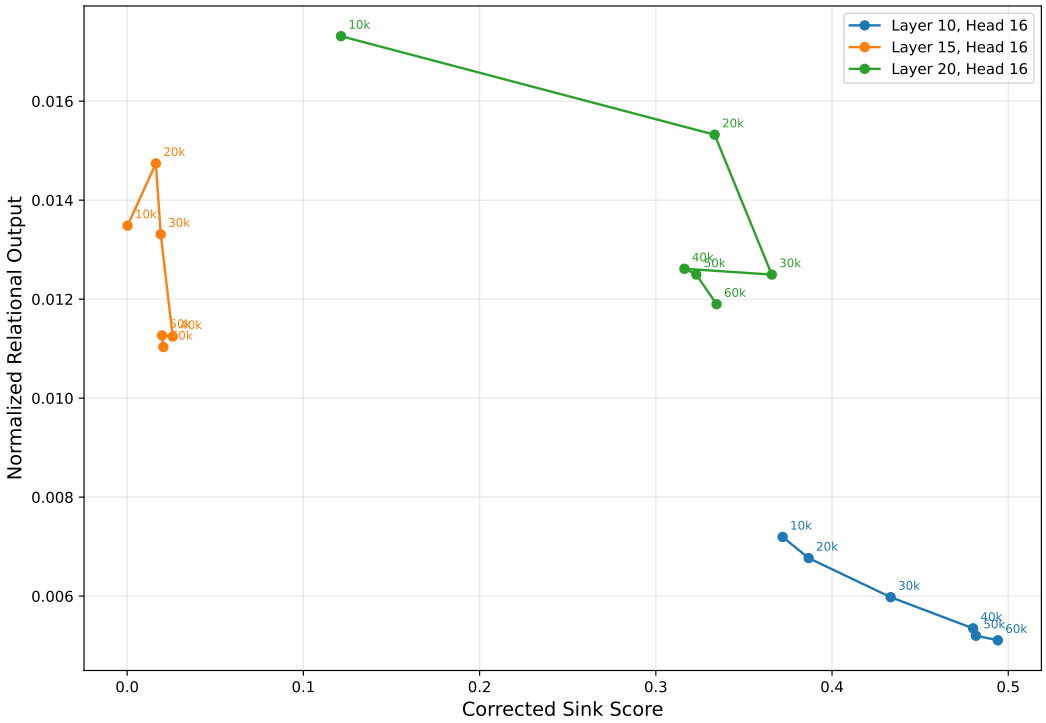


Figure 5: Trajectory of attention heads in the PreNorm model in the $(\tilde{\alpha}_1, \Gamma)$ plane.

Figure 6: **Sink emergence during pretraining.** Trajectories in the $(\tilde{\alpha}_1, \Gamma)$ plane for a LLaMA-style 350M model on C4. The sink head of PreNorm model later become strong sinks traverse the plane along paths of increasing sink and decreasing relational output, rather than starting there, while non-sink head stay in the regions with low sink score consistently. The systematic coupling between sink formation and relational-output collapse is consistent with the mechanistic picture of Section 3: first-token sink is acquired as a route toward approximate relational invariance.

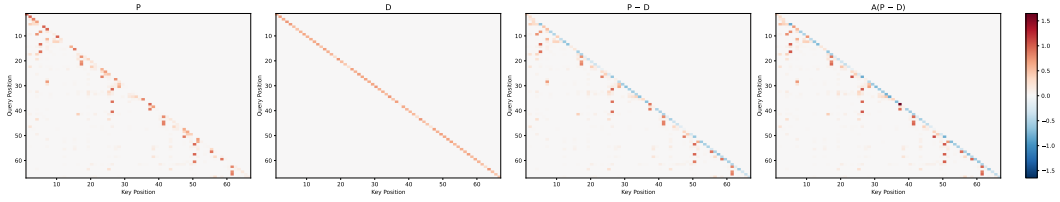


Figure 7: **How ADTE reshapes an attention matrix.** Left to right: $\boldsymbol{P}$ with a bright first column (sink) and a visible diagonal; learned diagonal target $\boldsymbol{D}$ localizing the cancellation on the self-token only; $\boldsymbol{P} - \boldsymbol{D}$ with the self-target selectively removed while the sink column and other interactions are preserved; and $\boldsymbol{A}(\boldsymbol{P} - \boldsymbol{D})$ after tokenwise amplification, which restores row-sum mass to the remaining entries. The operator is no longer row-stochastic, in line with Theorem 1: stepping outside the simplex is what lets ADTE realize near-relational-invariance behavior without collapsing onto $\mathbf{1}\boldsymbol{e}_1^\top$.

$\boldsymbol{D}^{\ell,h} = \mathrm{Diag}(\boldsymbol{\lambda}^{\ell,h})$, (iii) the post-cancellation operator $\boldsymbol{P}^{\ell,h} - \boldsymbol{D}^{\ell,h}$ (causal, row-substochastic, with possibly negative diagonal entries when $\lambda_i^{\ell,h} > P_{ii}^{\ell,h}$), and (iv) the final effective operator $\boldsymbol{A}^{\ell,h}(\boldsymbol{P}^{\ell,h} - \boldsymbol{D}^{\ell,h})$ after tokenwise amplification. The visual signature of the substitution is concentrated: the first column of $\boldsymbol{P}$ (sink tokens) remains unchanged; only the diagonal is selectively thinned in $\boldsymbol{P} - \boldsymbol{D}$; and $\boldsymbol{A}(\boldsymbol{P} - \boldsymbol{D})$ restores overall mass on off-diagonal entries while keeping the diagonal small.

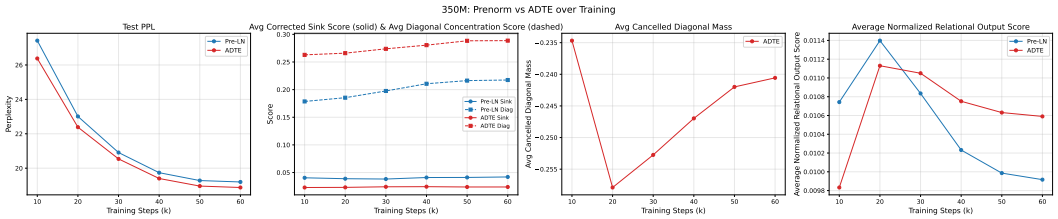


Figure 8: **Training dynamics: mechanistic substitution rather than global damping.** Pre-LN and ADTE follow comparable validation-loss trajectories (a), but the mechanistic profile diverges: under ADTE, the rise of first-token sink is arrested and replaced by growing diagonal cancellation ($\bar{\alpha}_{\text{diag}}$, $\bar{c}_{\text{diag}}$ in (b,c)), while relational output $\bar{\Gamma}$ stays close to baseline (d) instead of collapsing as it does under simple gating.

D.4 Training dynamics of ADTE

ADTE is designed to act as a smooth deformation of the standard attention block rather than as a disruptive replacement. Figure 8 overlays training-time curves for PreNorm and ADTE: validation loss, the aggregate sink and diagonal-concentration scores $\bar{\alpha}_1$ and $\bar{\alpha}_{\text{diag}}$, the cancelled diagonal mass $\bar{c}_{\text{diag}}$, and the normalized relational output $\bar{\Gamma}$. The signature of ADTE is *mechanistic substitution*: $\bar{\alpha}_1$ stops climbing while $\bar{\alpha}_{\text{diag}}$ and $\bar{c}_{\text{diag}}$ rise in step, $\bar{\Gamma}$ remains comparable to the baseline rather than collapsing, and the validation-loss curve is consistently lower than PreNorm model.

D.5 Performance of Models at 1B Scale

To extend the evidence beyond the original 350M limit, we additionally trained 1B-parameter models for 5.5B tokens under the same evaluation protocol:

Table 5: Performance of models at the 1B scale.

	PreNorm	Gated	ADTE
PPL	17.01	17.33	**16.82**

At 1B parameters, ADTE reduces validation perplexity by 0.19 relative to PreNorm (1.12% relative) and by 0.51 relative to Gated Attention (2.94% relative). This extends the controlled ADTE training evidence from 350M to 1B parameters and indicates that its benefit does not disappear at the next model scale.

D.6 Performance on PostNorm Models

Although PostNorm is no longer the dominant architecture for modern LLMs due to gradient instability at depth, it remains conceptually important as the original Transformer design and is still adopted in certain settings. We therefore evaluate whether ADTE generalizes beyond the PreNorm family. Table 6 reports results on two PostNorm-style architectures: a vanilla PostNorm with RMSNorm, and Keel (Chen & Wei, 2026), which adds a $\sqrt{2L}$ residual scaling factor. On the vanilla PostNorm baseline, training at the 250M and 350M scales diverges for all methods (PPL > 1400), confirming the known instability of PostNorm at depth—this is not specific to ADTE. At the 130M scale where PostNorm still converges, ADTE achieves competitive perplexity while substantially reducing the sink rate and boosting diagonal concentration ($\bar{\alpha}_{\text{diag}}$: 0.35 vs. 0.12 for the baseline). On the more stable Keel architecture, ADTE consistently outperforms both the Keel baseline and Keel+Gated across all three scales, with PPL improvements of 1.3, 1.2, and 0.7 points at 130M, 250M, and 350M respectively, while maintaining the expected mechanistic signature of lower sink and higher diagonal concentration. These results confirm that ADTE's self-cancellation mechanism is architecture-agnostic: it benefits any causal softmax attention block regardless of the normalization placement.

Table 6: **Extended pretraining results** across architectures (PostNorm, Keel) and model scales.

		PPL	Sink_ε	$\bar{\alpha}_1$	$\bar{\alpha}_{\text{diag}}$
130M	Post-LN	25.50	0.1855	0.0755	0.1241
	Gated	25.07	0.0819	0.0372	0.1287
	ADTE	25.16	0.0730	0.0258	0.3509
250M	Post-LN	1435.14	–	–	–
	Gated	1435.92	–	–	–
	ADTE	1431.23	–	–	–
350M	Post-LN	1432.79	–	–	–
	Gated	1438.61	–	–	–
	ADTE	1434.47	–	–	–
130M	Keel	27.03	0.0861	0.0391	0.2031
	Keel+Gated	26.69	0.0653	0.0278	0.1730
	Keel+ADTE	25.72	0.0657	0.0279	0.2385
250M	Keel	23.26	0.0694	0.0306	0.2304
	Keel+Gated	23.35	0.0667	0.0333	0.1889
	Keel+ADTE	22.03	0.0508	0.0215	0.3037
350M	Keel	20.28	0.0773	0.0338	0.2298
	Keel+Gated	20.34	0.0712	0.0358	0.1875
	Keel+ADTE	19.63	0.0563	0.0255	0.2788

D.7 Downstream performance

To test whether ADTE transfers beyond validation perplexity, we fine-tuned the 350M ADTE and 350M PreNorm checkpoints with LoRA on commonsense reasoning data, using the same fine-tuning and evaluation protocol for both models. The resulting accuracies are: ADTE improves accuracy on five of the six benchmarks, with particularly clear gains on ARC-Easy, ARC-Challenge, PIQA, and SocialIQA. BoolQ is effectively unchanged (-0.12 points). The macro-average accuracy increases from 11.68 to 22.26, a gain of 10.58 percentage points. These results provide initial evidence that the architectural change can improve task-level adaptation rather than merely changing attention statistics or C4 perplexity.

Table 7: Downstream Performance of models at the 350M scale on commonsense reasoning data.

	ARC-Easy	ARC-Challenge	HellaSwag	PIQA	BoolQ	SocialIQA	Macro avg.
PreNorm	5.20	4.10	0.57	2.20	**55.44**	2.56	11.68
ADTE	**12.29**	**13.51**	**4.99**	**27.75**	55.32	**19.70**	**22.26**
Δ(**ADTE-PreNorm**)	+7.09	+9.41	+4.42	+25.55	-0.12	+17.14	+10.58

E Broader Impact, Limitations, and Future Work

E.1 Broader Impact

Our contribution is primarily mechanistic: we identify first-token attention sink as the unique value-agnostic relational-invariance operator of causal softmax attention and show that a head's relational output is controlled by its distance to this target. Practically, this reframing suggests that attention sink is not only a defect to be removed but a trace of a missing primitive. Because the proposed changes act at the level of the attention operator and do not alter training data or objectives, we do not foresee distinctive negative societal impacts beyond those already associated with large language models in general.

E.2 Limitations

Our formal analysis targets residual PreNorm attention, which covers the decoder-only LLMs studied here; extending the exact relational invariance characterization to PostNorm blocks requires a local-

linearized treatment of the post-residual normalization and is not done here. The empirical study is restricted to moderate-scale open LLaMA and Qwen checkpoints and a controlled continued-pretraining setting; scaling laws of the sink–invariance geometry at frontier scale are not directly verified.

E.3 Future Work

Several directions follow naturally from the present framework. First, it would be interesting to analyze relational invariance geometry under alternative normalizations (Post-LN, RMSNorm variants, QK-norm) and under sigmoid or softmax-free attention, where the row-stochastic constraint is relaxed and different relational invariance operators become admissible. Second, the distance $d_{\mathrm{inv}}(\boldsymbol{P})$ gives a principled training-time regularizer; investigating whether explicitly penalizing d_{inv} or providing learned self-variance routes from initialization reduces sink formation without harming language-modeling performance is a concrete next step. Third, the counterfactual diagonal-target probe suggests a broader program: systematically designing alternative self-variance routes (diagonal, register-token (Darcet et al., 2023), or learned-subspace variants) and characterizing which ones Transformers preferentially adopt during training.