Quantified Robustness for Structured Rotary Attention

Internal note extending RoFormer and STRING analyses

November 9, 2025

Abstract

RoFormer [1] and STRING [2] show that commuting skew generators give rise to relative attention logits when paired with structured post-rotations. This note isolates the genuinely novel contribution of our work: explicit robustness guarantees when the commutator and post-rotation assumptions are only approximately satisfied. We keep the notation of the original analyses and prove quantitative bounds for the BCH series, structured rotations, Cayley post-rotations, and, finally, the attention logits.

1 Setup and baseline results

We adopt the shared notation used by RoFormer and STRING. Let $\{L_k\}_{k=1}^{d_c} \subset \mathfrak{so}(d_h)$ be skew-symmetric generators that commute pairwise in the idealized model. For a displacement $r \in \mathbb{R}^{d_c}$ define

$$A(r) \coloneqq \sum_{k=1}^{d_c} r_k L_k, \qquad R_{\mathrm{STR}}(r) \coloneqq \exp\big(A(r)\big) \in SO(d_h). \tag{1}$$

The joint-block structure [2] provides an orthogonal matrix U that block-diagonalizes the generators and exposes an active subspace of dimension $d_{\mathrm{act}} = 2m$. The corresponding projector is denoted by $\Pi_{\mathrm{act}} = U \begin{bmatrix} I_{2m} & 0 \\ 0 & 0 \end{bmatrix} U^\top$ and $\Pi_{\mathrm{null}} = I - \Pi_{\mathrm{act}}$. Following STRING, we allow for post-rotations $P_{\mathrm{sp}} \in SO(d_h)$ obtained from a Cayley transform

$$P_{\mathrm{sp}} = (I - S)(I + S)^{-1}, \qquad S = S_- + E, \tag{2}$$

where $S_-^\top = -S_-$ captures the exact skew component induced by the algebra and $E^\top = -E$ represents a small mixing perturbation.

Assumption 1 (Baseline relative property)**.** *In the unperturbed setting ($E = 0$ and $[L_a, L_b] = 0$) the active rotation obeys $R_{\mathrm{STR}}(r)^\top R_{\mathrm{STR}}(s) = R_{\mathrm{STR}}(s - r)$ and $\Pi_{\mathrm{act}} P_{\mathrm{sp}} \Pi_{\mathrm{act}} = \Pi_{\mathrm{act}}$.*

This captures the core theorems proved in RoFormer (Equations (11)–(13)) and the deterministic STRING analysis. The remainder of the note quantifies the deviation from this idealisation.

Error parameters

We measure the size of the commutator defect by

$$\varepsilon_{\mathrm{comm}} \coloneqq \max_{a,b} \|[L_a, L_b]\| . \tag{3}$$

The block structure induces weights $\lambda_{k,u}$ so that $A(r)$ becomes block-diagonal in the active planes. Define the structural constant

$$\Lambda \coloneqq \max_{u\in\{1,\dots,m\}} \sum_{k=1}^{d_c} |\lambda_{k,u}|\,, \qquad R \coloneqq \max_i \|r_i\|_1\,. \tag{4}$$

For the Cayley map we set $\rho \coloneqq \|S\|$ and $\eta \coloneqq \|E\|$, assuming $\rho < 1$ so that $(I+S)$ is invertible. When the post-rotation is supplied externally ("model-free" case) we instead work with

$$\eta_{\mathrm{mix}} \coloneqq \max\left\{\|\Pi_{\mathrm{null}} P_{\mathrm{sp}} \Pi_{\mathrm{act}}\|\,,\ \|\Pi_{\mathrm{act}} P_{\mathrm{sp}} \Pi_{\mathrm{null}}\|\right\}. \tag{5}$$

2 Quantitative BCH control

The first technical ingredient bounds the size of the generator and provides an explicit constant for the Baker–Campbell–Hausdorff (BCH) remainder. This portion recapitulates the deterministic algebra from STRING but makes the constants explicit for later use.

Lemma 1 (Generator norm)**.** *For every displacement $r \in \mathbb{R}^{d_c}$ we have $\|A(r)\| \le \Lambda \|r\|_1$. Define $C_{\mathrm{BCH}}(r,s) \coloneqq C_\star \exp\big(2\Lambda(\|r\|_1 + \|s\|_1)\big)$ for an absolute $C_\star$ that bounds higher-order BCH coefficients.*

Lemma 2 (Quantitative BCH)**.** *Let $r, s \in \mathbb{R}^{d_c}$. Then*

$$\big\|\log\big(\exp A(r)\,\exp A(s)\big) - \big(A(r)+A(s)\big)\big\| \le \tfrac{1}{2}\varepsilon_{\mathrm{comm}} \|r\|_1 \|s\|_1 + C_{\mathrm{BCH}}(r,s)\varepsilon_{\mathrm{comm}}^2 \|r\|_1^2 \|s\|_1^2\,. \tag{6}$$

Proposition 1 (Structured rotation stability)**.** *With the same hypotheses we obtain*

$$\Big\|R_{\mathrm{STR}}(r)^\top R_{\mathrm{STR}}(s) - R_{\mathrm{STR}}(s-r)\Big\| \le C_1 \varepsilon_{\mathrm{comm}} \|r\|_1 \|s\|_1 + C_2 \varepsilon_{\mathrm{comm}}^2 \|r\|_1^2 \|s\|_1^2\,, \tag{7}$$

where C_1, C_2 are universal (arising from the matrix exponential Lipschitz bound applied to Lemma 2).

3 Cayley post-rotations under perturbations

We now deviate from the exact STRING setup and allow S to acquire an additive perturbation E. All statements below are new: neither RoFormer nor STRING provide quantitative stability estimates in this regime.

Lemma 3 (Cayley Lipschitzness)**.** *If $\|S\| \le \rho < 1$ and $\|E\| = \eta$, then*

$$\|P_{\mathrm{sp}}(S) - P_{\mathrm{sp}}(S_-)\| \le \frac{2}{(1-\rho)^2}\,\eta. \tag{8}$$

Lemma 4 (Quadratic active mixing)**.** *Let $\Pi_{\mathrm{act}} S \Pi_{\mathrm{act}} = 0$ and $\Pi_{\mathrm{null}} S \Pi_{\mathrm{null}} = S_{\mathrm{null}}$, as in the STRING block-structured Cayley factorisation. Then*

$$\|\Pi_{\mathrm{act}} P_{\mathrm{sp}}(S) \Pi_{\mathrm{act}} - \Pi_{\mathrm{act}}\| \le \frac{2}{(1-\rho)^3}\,\eta^2. \tag{9}$$

Lemma 5 (Model-free bound)**.** *For an arbitrary $P_{\mathrm{sp}} \in SO(d_h)$ written in the STRING basis as $U \begin{bmatrix} A & B \\ C & D \end{bmatrix} U^\top$ we have*

$$\|\Pi_{\mathrm{act}} P_{\mathrm{sp}} \Pi_{\mathrm{act}} - \Pi_{\mathrm{act}}\| \le 2\,\eta_{\mathrm{mix}}. \tag{10}$$

Combining Lemmas 4 and 5 yields a unified parameter

$$\delta_{\mathrm{mix}} \coloneqq \begin{cases} \dfrac{2}{(1-\rho)^3}\,\eta^2, & \text{structured Cayley / ESPR case,} \\ 2\,\eta_{\mathrm{mix}}, & \text{model-free case.} \end{cases} \tag{11}$$

Corollary 1 (Active block approximation)**.** *For every displacement r,*

$$\|\Pi_{\mathrm{act}} R_{\mathrm{sp}}(r)\Pi_{\mathrm{act}} - \Pi_{\mathrm{act}} R_{\mathrm{STR}}(r)\Pi_{\mathrm{act}}\| \le \delta_{\mathrm{mix}}, \qquad R_{\mathrm{sp}}(r) \coloneqq R_{\mathrm{STR}}(r)P_{\mathrm{sp}}. \tag{12}$$

4 Robust relative logits

Let $q_i = W_Q x_i$ and $k_j = W_K x_j$ denote the query and key projections used in RoFormer/STRING, and set $q_i^{(\mathrm{act})} = \Pi_{\mathrm{act}} q_i$, $k_j^{(\mathrm{act})} = \Pi_{\mathrm{act}} k_j$. Define rotated vectors $\tilde{q}_i = \Pi_{\mathrm{act}} R_{\mathrm{sp}}(r_i) q_i^{(\mathrm{act})}$ and $\tilde{k}_j = \Pi_{\mathrm{act}} R_{\mathrm{sp}}(r_j) k_j^{(\mathrm{act})}$. The attention logit at head dimension $d_{\mathrm{act}} = 2m$ is

$$\alpha_{ij} = \frac{1}{\sqrt{2m}}\,\tilde{q}_i^{\top}\tilde{k}_j. \tag{13}$$

Let

$$\alpha_{ij}^{\star} \coloneqq \frac{1}{\sqrt{2m}}\big(q_i^{(\mathrm{act})}\big)^{\top}\big(\Pi_{\mathrm{act}} R_{\mathrm{STR}}(r_j - r_i)\Pi_{\mathrm{act}}\big)k_j^{(\mathrm{act})}, \tag{14}$$

the ideal relative logit from the original proofs.

Lemma 6 (Logit perturbation)**.** *For all (i,j),*

$$\big|\alpha_{ij} - \alpha_{ij}^{\star}\big| \le \Big\|q_i^{(\mathrm{act})}\Big\|\,\Big\|k_j^{(\mathrm{act})}\Big\|\,\Big(C_1\varepsilon_{\mathrm{comm}}\,\|r_i\|_1\,\|r_j\|_1 + C_2\varepsilon_{\mathrm{comm}}^2\,\|r_i\|_1^2\,\|r_j\|_1^2 + \delta_{\mathrm{mix}}\Big). \tag{15}$$

Theorem 1 (Quantified robust relative logits)**.** *Let $C_{\mathrm{mix}} = 1$ and $\delta_{\mathrm{mix}} = \eta_{\mathrm{mix}}$ in the model-free case, and $C_{\mathrm{mix}} = \frac{2}{(1-\rho)^3}$ together with $\delta_{\mathrm{mix}} = \eta^2$ in the structured Cayley case. Then*

$$\begin{aligned}\big|\alpha_{ij} - \alpha_{ij}^{\star}\big| \le{}& \Big(\tfrac{1}{2} + C_{\mathrm{BCH}}(r_i, r_j)\Big)\varepsilon_{\mathrm{comm}}\,\|r_i\|_1\,\|r_j\|_1\,\Big\|q_i^{(\mathrm{act})}\Big\|\,\Big\|k_j^{(\mathrm{act})}\Big\| \\ &+ C_{\mathrm{mix}}\,\delta_{\mathrm{mix}}\,\Big\|q_i^{(\mathrm{act})}\Big\|\,\Big\|k_j^{(\mathrm{act})}\Big\|.\end{aligned} \tag{16}$$

Proof. Combine Lemma 6 with Proposition 1 and Corollary 1. □

5 Discussion: overlap versus novelty

- Definitions of $A(r)$, R_{STR}, and the use of joint block diagonalisation mirror the core statements of RoFormer and STRING; they are reintroduced here solely to fix notation.
- Lemma 1 matches the deterministic bounds implicit in STRING, but we expose explicit constants needed for downstream stability.
- Lemmas 3–5, Corollary 1, Lemma 6, and Theorem 1 are new: neither RoFormer nor STRING quantify how attention logits behave when generators only approximately commute and when the Cayley post-rotation leaks outside the active subspace.

References

[1] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[2] Krzysztof Choromanski and Daniel Hardesty Lewis. STRING: Generalization of RoPE. Internal manuscript, 2023.