Spectral-Sphere-Constrained Hyper-Connections

Zhaoyi Liu[1] **Haichuan Zhang**[2] **Ang Li**[1]
[1]University of Maryland, College Park
[2]University of Utah
[1]{zhaoyil, angliece}@umd.edu
[2]{hc.zhang}@utah.edu

Abstract

Hyper-Connections (HC) extend residual connections into multiple streams, employing residual matrices for cross-stream mixing to enrich model expressivity. However, unconstrained mixing disrupts the identity mapping property intrinsic to the residual connection, causing unstable training. To address this, Manifold-Constrained Hyper-Connections (mHC) and its variants restrict these matrices to be doubly stochastic via Sinkhorn-Knopp (SK) algorithm or permutation-based parameterizations. We reveal three limitations of this doubly stochastic constraint: (1) identity degeneration, where learned matrices collapse around the identity initialization and diminish cross-stream interactions, (2) a expressivity bottleneck, where the doubly stochastic constraint restricts the freedom of the subdominant spectrum of the residual matrices, preventing the model from selectively preserving or attenuating cross-stream variations, and (3) parameterization inefficiencies, manifesting as unstable SK iterations or the factorial-scaling overhead of permutation-based parameterizations. To overcome these flaws, we propose Spectral-Sphere-Constrained Hyper-Connections (s^2HC[1]). By confining residual matrices to a spectral norm sphere, s^2HC restores free control over the subdominant spectrum, enabling the model to selectively preserve or attenuate cross-stream variations. This shift eliminates unstable SK iterations and factorial parameterization, enabling expressive, non-degenerate residual matrices while preserving training stability.

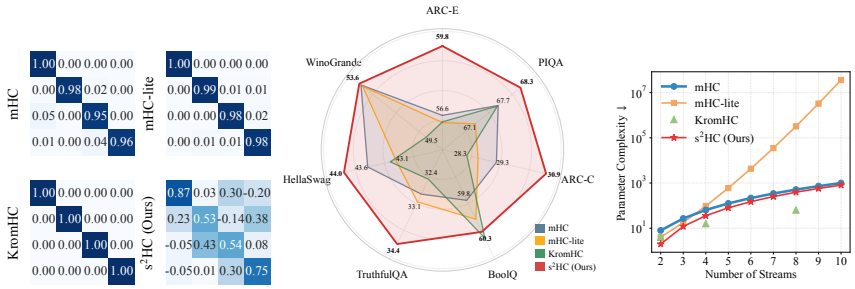

Figure 1: s^2HC addresses the identity degeneration, expressivity bottleneck, and parameterization inefficiencies in existing Manifold-Constrained Hyper-Connections (mHC [1]) and its permutation-based variants (mHC-lite [2] and KromHC [3]). **Left**: Learned residual matrices (4 streams). mHC, mHC-lite, and KromHC degenerate into the identity mapping, whereas s^2HC leverages expressive, non-degenerate residual matrices. **Middle**: Model performance, with each axis scaled to its own range and absolute scores annotated. s^2HC improves over baselines across benchmarks. **Right**: Parameterization overhead, measured by the parameter complexity of the residual parameterization as the number of streams grows. s^2HC scales compactly, avoiding both the factorial explosion of mHC-lite and the restriction of KromHC to specific stream counts.

[1]Code is available at https://github.com/6zHAOyi/s2HC.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Residual connections [4] have been a cornerstone of deep learning for over a decade, stabilizing gradient propagation through identity mappings and becoming a standard component of modern deep networks [5, 6, 7]. Hyper-Connections (HC) [8] have recently extended this single-stream residual connection into parallel residual streams by using a dynamic residual matrix at each layer to mix the features across the streams, increasing the model's topological complexity and capacity. However, unconstrained residual matrices compromise the identity preservation property intrinsic to the residual connection, which causes training instability [1].

To address this, Manifold-Constrained Hyper-Connections (mHC) [1] proposes constraining residual matrices to be doubly stochastic to preserve the identity mapping property. Doubly stochastic matrices are characterized by non-negative entries and unit row and column sums. This constraint bounds the spectral norm by one to mitigate gradient explosion, while preserving the mean component of the residual streams [1]. mHC enforces this constraint via Sinkhorn-Knopp (SK) [9] algorithm, which iteratively normalizes rows and columns of residual matrices. However, SK yields only an unstable constraint approximation. As reported in [2], the resulting constraint violations accumulate across depth, undermining stability.

A recent variant, mHC-lite [2], enforces exact doubly stochasticity by parameterizing residual matrices as convex combinations of permutation matrices. However, this introduces factorial growth in auxiliary parameters, since the number of permutation matrices scales factorially with the number of streams, limiting scalability [2]. KromHC [3] mitigates this overhead by constructing the residual matrix via Kronecker products of smaller doubly stochastic matrices, each parameterized similarly to mHC-lite. However, this results in an incomplete parameterization of the doubly stochastic manifold. Moreover, its efficiency relies on decomposing the number of residual streams into small factors, and degrades to intractable factorial complexity when large prime factors are present [3].

Beyond these parameterization inefficiencies, we identify two previously overlooked intrinsic limitations of the doubly stochastic constraint adopted in these methods. First, it is prone to ***identity degeneration***, where the learned residual matrices in each layer concentrate around the identity initialization, practically abandoning the intended cross-stream interactions. Second, it imposes a ***spectral expressivity bottleneck***: the doubly stochastic constraint restricts the freedom of the subdominant spectrum of the residual matrices, limiting the model's ability to selectively preserve or attenuate cross-stream variations while mixing. (Detailed analysis is in § 4.)

To overcome these limitations, we propose Spectral-Sphere-Constrained Hyper-Connections (s^2HC). We constrain the residual matrices to a spectral norm sphere rather than to a doubly stochastic manifold. As shown in Figure 1, this shift yields three key advantages: ❶ s^2HC restores subdominant spectrum control, yielding non-degenerate residual matrices. ❷ These non-degenerate matrices selectively preserve or attenuate cross-stream variations, supporting expressive feature interactions and improving the model performance. ❸ s^2HC provides an exact closed-form parameterization with stable spectral norm control, eliminating both the factorial overhead and the approximation errors of prior methods. In summary, our contributions are:

- We introduce Spectral-Sphere-Constrained Hyper-Connections (s^2HC), which reformulates hyper-connection constraints within a spectral norm sphere, mitigating identity degeneration, alleviating the expressivity bottleneck, and eliminating factorial parameterization overhead.
- Experiments demonstrate that s^2HC improves model capability over existing doubly stochastic constrained hyper-connection methods while preserving the training stability.
- The expressivity, stability, and parameterization efficiency of s^2HC offer the community a practical reference paradigm for deploying hyper-connection architectures.

2 Related Work

Residual connections [4] stabilize deep network training by introducing identity skip connections. Expanding on this, Hyper-Connections (HC) [8] introduce parallel residual streams mixed by dynamic matrices to enhance model capacity. Frac-Connections [10] fragment streams into chunks as an alternative extension for Hyper-Connections to reduce the memory access costs of parallel residual streams. To address the training instability inherent in HC, Manifold-Constrained Hyper-Connections

(mHC [1]) and its exact-stochasticity variants (mHC-lite [2], KromHC [3]) confine the residual matrices to be doubly stochastic. While effective for restoring stability, enforcing this constraint inevitably introduces parameterization inefficiencies and intrinsic expressivity bottlenecks. In this work, we align with the hyper-connection paradigm, focusing on resolving these expressivity and parameterization bottlenecks.

3 Preliminary

Hyper-Connections (HC). To enrich the expressivity of the model, Hyper-Connections [8] extend the single-stream paradigm to n parallel residual streams. Let $X_l = (\boldsymbol{x}_{l,1}^\top, \boldsymbol{x}_{l,2}^\top, ..., \boldsymbol{x}_{l,n}^\top)^\top \in \mathbb{R}^{n\times C}$ represent the expanded features of n streams at the l-th layer. HC introduces a dynamic mixing mechanism:

$$X_{l+1} = \mathcal{H}_l^{\text{res}} X_l + (\mathcal{H}_l^{\text{post}})^\top \mathcal{F}(\mathcal{H}_l^{\text{pre}} X_l, \mathcal{W}_l) \tag{1}$$

Here, $\mathcal{F}(\cdot, \mathcal{W}_l)$ denotes the learnable layer transformation with parameter $\mathcal{W}_l$ in the branch (e.g., Attention or MLP block). $\mathcal{H}_l^{\text{res}} \in \mathbb{R}^{n\times n}$ represents a learnable residual matrix that mixes features within the residual streams. Similarly, $\mathcal{H}_l^{\text{pre}} \in \mathbb{R}^{1\times n}$ aggregates features from the $(n \times C)$-dim stream into a $(1 \times C)$-dim layer branch input, and conversely, $\mathcal{H}_l^{\text{post}} \in \mathbb{R}^{1\times n}$ maps the layer branch output back onto the streams. However, without constraints on these residual matrices, HC may suffer from severe training instability [1].

Manifold-Constrained Hyper-Connections (mHC). To ensure training stability, mHC [1] constrains $\mathcal{H}_l^{\text{res}}$ to lie in the Birkhoff polytope $\mathcal{B}_n$, i.e., the set of doubly stochastic matrices. This polytope is a subset of the affine subspace $\mathcal{A}_n = \{\mathcal{H} \in \mathbb{R}^{n\times n} \mid \mathcal{H}\mathbf{1}_n = \mathbf{1}_n,\ \mathbf{1}_n^\top \mathcal{H} = \mathbf{1}_n^\top\}$, which is obtained by adding the element-wise non-negativity constraint to $\mathcal{A}_n$:

$$\mathcal{B}_n = \{\mathcal{H} \in \mathcal{A}_n \mid \mathcal{H} \geq 0\}. \tag{2}$$

The affine constraint in $\mathcal{A}_n$ ensures that the uniform vector $\mathbf{1}_n$ remains invariant, thereby conserving the mean component across residual streams. Imposing the additional non-negativity condition bounds the spectral norm to 1 (i.e., $\|\mathcal{H}_l^{\text{res}}\|_2 = 1$), which prevents signal amplification and stabilizes deep propagation [1].

As such, mHC can be fully formulated as:

$$\begin{cases} \boldsymbol{x}'_l = \text{RMSNorm}(\boldsymbol{x}_l) \\ \mathcal{H}_l^{\text{pre}} = \text{sigmoid}\left(\alpha_l^{\text{pre}} \cdot (\boldsymbol{x}'_l W_l^{\text{pre}}) + \boldsymbol{b}_l^{\text{pre}}\right) \\ \mathcal{H}_l^{\text{post}} = 2 \cdot \text{sigmoid}\left(\alpha_l^{\text{post}} \cdot (\boldsymbol{x}'_l W_l^{\text{post}}) + \boldsymbol{b}_l^{\text{post}}\right) \\ \mathcal{H}_l^{\text{res}} = \text{SK}\big(\exp(\alpha_l^{\text{res}} \cdot \text{mat}(\boldsymbol{x}'_l W_l^{\text{res}}) + \boldsymbol{b}_l^{\text{res}})\big) \end{cases} \tag{3}$$

where $\boldsymbol{x}_l \in \mathbb{R}^{1\times nC}$ is the vector flattened from the expanded streams X_l. $W_l^{\text{pre}}, W_l^{\text{post}} \in \mathbb{R}^{nC\times n}$ and $W_l^{\text{res}} \in \mathbb{R}^{nC\times n^2}$ are linear projections for dynamic mappings. The terms $\boldsymbol{b}_l^{\text{pre}}, \boldsymbol{b}_l^{\text{post}} \in \mathbb{R}^{1\times n}$ and $\boldsymbol{b}_l^{\text{res}} \in \mathbb{R}^{n\times n}$ are learnable biases. $\alpha_l^{\text{pre}}, \alpha_l^{\text{post}}, \alpha_l^{\text{res}}$ are scalar gating factors. RMSNorm$(\cdot)$ refers to the RMSNorm [11]. sigmoid$(\cdot)$ denotes the Sigmoid function. mat$(\cdot)$ reshapes a matrix from $\mathbb{R}^{1\times n^2}$ to $\mathbb{R}^{n\times n}$ while SK$(\cdot)$ denotes the Sinkhorn–Knopp algorithm which iteratively alternates row-wise and column-wise normalization to enforce approximate doubly stochasticity. However, finite SK iterations cannot guarantee exact doubly stochasticity. The resulting approximation errors can accumulate across layers, potentially undermining the stability of deep networks.

4 Observation

To evaluate the practical expressivity of these doubly stochastic constrained hyper-connections (mHC, mHC-lite, and KromHC), we analyze their learned residual matrices $\mathcal{H}_l^{\text{res}}$ and pairwise cosine similarity among residual streams after being mixed by $\mathcal{H}_l^{\text{res}}$ across all layers of the model. The experiments are conducted on a 28-layer 0.75B Qwen3 model with untied word embedding [12].

4.1 Identity Degeneration

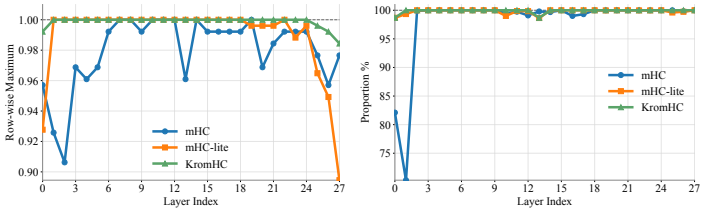

Figure 2: Depth-wise analysis of $\mathcal{H}_l^{\text{res}}$ for mHC, mHC-lite and KromHC after training. **Left**: the median row-wise maximum entries of $\mathcal{H}_l^{\text{res}}$. **Right**: the proportion of $\mathcal{H}_l^{\text{res}}$ where all row maximums are on the diagonal.

As shown in Figure 2, for all of these methods, the median row-wise maximum of the learned residual matrices $\mathcal{H}_l^{\text{res}}$ remains tightly concentrated around 1. All three methods exhibit minimal dispersion, with the widest range only spanning 0.98–1.0 for KromHC, 0.90-1.0 for mHC-lite and mHC. Moreover, the fraction of matrices whose row maxima lie strictly on the diagonal stays consistently high (above 99% for mHC-lite, KromHC). Since $\mathcal{H}_l^{\text{res}}$ is doubly stochastic (non-negative with unit row and column sums), the concentration of row-wise maxima on the diagonal with values approaching 1 implies that the learned matrices degenerate around the identity as shown in Figure 1, up to small off-diagonal tails. This indicates that the model actually abandons active cross-stream interactions at each layer. (More statistics and visualizations of $\mathcal{H}_l^{\text{res}}$ are provided in Appendix A.)

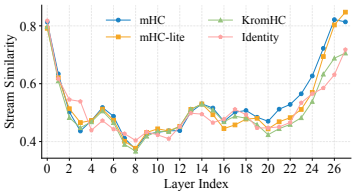

Figure 3: Mean pairwise cosine similarity among residual streams after being mixed by $\mathcal{H}_l^{\text{res}}$ across model layers. Identity denotes the baseline where $\mathcal{H}_l^{\text{res}}$ is fixed as an identity matrix while keeping all other settings identical to mHC, mHC-lite, and KromHC.

This degeneration is also reflected in the pairwise similarity within the residual streams. As illustrated in Figure 3, the similarity trajectories of mHC, mHC-lite, and KromHC closely follow those of the identity mapping baseline (where $\mathcal{H}_l^{\text{res}}$ is fixed as an identity matrix while keeping all other settings identical to these methods). This indicates that feature evolution is predominantly governed by the nonlinear branches rather than the residual feature mixing. This degeneration may explain the resulting training stability of these doubly stochastic constrained hyper-connections. However, prioritizing this form of apparent stability runs counter to the original design motivation of hyper-connections, which aims to enhance the model's topological expressivity through active cross-stream feature interactions [1, 8].

4.2 Spectral Expressivity Bottleneck

As discussed above, we observe that these doubly stochastic constrained hyper-connections degenerate into near-identity mappings, limiting cross-stream mixing. We suggest that this degeneration may be an inherent consequence of the spectral property of the doubly stochastic constraint.

As established by [13], under standard mixing conditions, doubly stochastic operators exhibit a strong inductive bias toward uniformity: while the mean component is preserved, the subdominant spectral modes are progressively attenuated. Consequently, repeated cross-stream mixing suppresses variations orthogonal to the mean, leading to feature homogenization.

Under this perspective, ***our key insight*** is that this suppression creates a spectral expressivity bottleneck for hyper-connections. When a doubly stochastic residual matrix departs from the identity to perform cross-stream mixing, its restricted subdominant singular values attenuate the off-mean directions being mixed. Thus, each layer faces a dilemma: the constraint biases mixing toward attenuation, limiting the model's ability to choose which cross-stream variations to preserve. The model may consequently remain near its identity initialization to avoid such attenuation, as observed in § 4.1. This renders cross-stream interaction effectively inert: hyper-connections retain their dedicated design but contribute little, quietly reducing to ordinary residual connections.

5 Methodology

Motivated by these observations, we propose Spectral-Sphere-Constrained Hyper-Connections (s^2HC). Instead of confining the residual matrix $\mathcal{H}_l^{\text{res}}$ to be doubly stochastic, we reformulate it into a spectral norm sphere restricted to the affine subspace $\mathcal{A}_n$.

5.1 Affine-Constrained Spectral Sphere

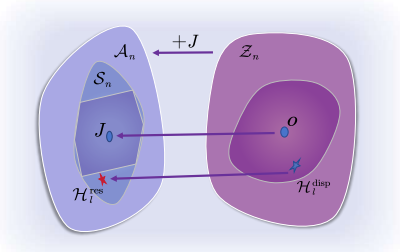

Figure 4: Overview of Spectral-Sphere-Constrained Hyper-Connections (s^2HC). The right plane depicts the zero-marginal subspace $\mathcal{Z}_n$, where the SVD parameterization generates $\mathcal{H}_l^{\text{disp}}$ under the bounded spectral norm $\|\mathcal{H}_l^{\text{disp}}\|_2 \leq 1$, centered at the origin O. The left plane depicts the target affine subspace $\mathcal{A}_n$, where the Birkhoff polytope $\mathcal{B}_n$ (inner polygon) is contained in the affine-constrained spectral norm sphere $\mathcal{S}_n$. The affine shift $+J$ maps O to J, and $\mathcal{H}_l^{\text{disp}}$ to the residual matrix $\mathcal{H}_l^{\text{res}}$.

We define an affine-constrained spectral norm sphere $\mathcal{S}_n = \{\mathcal{H} \in \mathcal{A}_n \mid \|\mathcal{H}\|_2 = 1\}$. Constraining the residual matrix $\mathcal{H}_l^{\text{res}}$ within $\mathcal{S}_n$ guarantees three critical properties:

1. **Mean Preservation**. $\mathcal{S}_n$ resides in $\mathcal{A}_n$, where the affine constraint ensures that the uniform vector $\mathbf{1}_n$ remains an invariant direction, thereby conserving the mean component across residual streams.
2. **Spectral Stability**. Any $\mathcal{H}_l^{\text{res}} \in \mathcal{S}_n$ satisfies $\|\mathcal{H}_l^{\text{res}}\|_2 = 1$. Moreover, $\mathcal{S}_n$ is closed under multiplication (*Proof in Appendix D.1*), preventing signal amplification and gradient explosion in deep networks.
3. **Enhanced Expressivity.** Since all doubly stochastic matrices possess a spectral norm of one, the Birkhoff polytope $\mathcal{B}_n$ (i.e. the set of doubly stochastic matrices) is contained within $\mathcal{S}_n$. By confining to $\mathcal{S}_n$, s^2HC eliminates the suppression of subdominant singular values inherent to the doubly stochastic constraint. This restored spectral freedom naturally enables the model's capacity for selectively preserving or alleviating the cross-stream variations.

Parameterization Equivalence via Spectral Decoupling. Directly parameterizing $\mathcal{H}_l^{\text{res}} \in \mathcal{A}_n$ with a strict unit spectral norm is challenging. To address this, we note that the affine subspace $\mathcal{A}_n$ is a translation of the zero-marginal subspace $\mathcal{Z}_n = \{\mathcal{H} \in \mathbb{R}^{n\times n} \mid \mathcal{H}\mathbf{1}_n = \mathbf{0}_n, \mathbf{1}_n^\top \mathcal{H} = \mathbf{0}_n^\top\}$ by the uniform matrix $J = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ (*Proof in Appendix D.2*). Thus, any target residual matrix admits a unique decomposition $\mathcal{H}_l^{\text{res}} = J + \mathcal{H}_l^{\text{disp}}$, where $\mathcal{H}_l^{\text{disp}} \in \mathcal{Z}_n$.

Then consider any column $\boldsymbol{x} \in \mathbb{R}^n$ of the expanded residual streams X_l, which can be uniquely decomposed as $\boldsymbol{x} = \boldsymbol{x}_\parallel + \boldsymbol{x}_\perp, \boldsymbol{x}_\parallel \in \text{span}\{\mathbf{1}_n\},\ \boldsymbol{x}_\perp \in \mathbf{1}_n^\perp$. The operators J and $\mathcal{H}_l^{\text{disp}}$ act orthogonally on these components: $J\boldsymbol{x}_\perp = \mathbf{0}_n, \mathcal{H}_l^{\text{disp}}\boldsymbol{x}_\parallel = \mathbf{0}_n$. This decouples their spectral contributions and leads to Proposition 1 (*Proof in Appendix D.3*).

Proposition 1 (Spectral Decoupling)**.** *Let $J = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. For any $\mathcal{H}_l^{\text{disp}} \in \mathcal{Z}_n$, the spectral norm of the corresponding residual matrix $\mathcal{H}_l^{\text{res}} = J + \mathcal{H}_l^{\text{disp}}$ satisfies:*

$$\|\mathcal{H}_l^{\text{res}}\|_2 = \max\left(\|J\|_2, \|\mathcal{H}_l^{\text{disp}}\|_2\right) \tag{4}$$

Remark 1. Since $\|J\|_2 = 1$, Proposition 1 establishes that enforcing $\|\mathcal{H}_l^{\text{res}}\|_2 = 1$ in the affine subspace $\mathcal{A}_n$ is equivalent to bounding $\|\mathcal{H}_l^{\text{disp}}\|_2 \leq 1$ in the zero-marginal subspace $\mathcal{Z}_n$.

5.2 Spectral-Sphere-Constrained Hyper-Connections

As established in Proposition 1, enforcing $\mathcal{H}_l^{\text{res}} \in \mathcal{S}_n$ reduces to constructing $\mathcal{H}_l^{\text{disp}} \in \mathcal{Z}_n$ satisfying $\|\mathcal{H}_l^{\text{disp}}\|_2 \leq 1$. Since $\mathcal{H}_l^{\text{disp}}$ lies in the zero-marginal subspace $\mathcal{Z}_n$, its rank is at most $n-1$. We therefore parameterize it via the compact singular value decomposition (SVD):

$$\mathcal{H}_l^{\text{disp}} = U_l \Sigma_l V_l^\top \tag{5}$$

where $U_l, V_l \in \mathbb{R}^{n\times(n-1)}$ satisfy $U_l^\top U_l = V_l^\top V_l = I_{n-1}$, and $\Sigma_l \in \mathbb{R}^{(n-1)\times(n-1)}$ is a diagonal matrix containing singular values.

To satisfy the zero-sum constraint (i.e. $\mathcal{H}_l^{\text{disp}} \in \mathcal{Z}_n$), we constrain the singular vectors to lie in $\mathbf{1}_n^\perp$ via factorizing:

$$U_l = U_{\mathcal{Z}} U_l^{\text{core}}, \qquad V_l = U_{\mathcal{Z}} V_l^{\text{core}} \tag{6}$$

where $U_{\mathcal{Z}} \in \mathbb{R}^{n\times(n-1)}$ denotes the truncated Helmert matrix, which forms an orthonormal basis of subspace $\mathbf{1}_n^\perp$, and $U_l^{\text{core}}, V_l^{\text{core}} \in \mathbb{R}^{(n-1)\times(n-1)}$ are orthogonal matrices. Substituting these factorizations into the the compact SVD in Eq. 5 yields our proposed parameterization, which provides a complete and spectral-controlled parameterization of the zero-marginal subspace $\mathcal{Z}_n$ as in Proposition 2 (*Proof in Appendix D.4*).

Proposition 2 (Complete and Spectral-Controlled Parameterization)**.** *Let $U_{\mathcal{Z}} \in \mathbb{R}^{n\times(n-1)}$ be an orthonormal basis of the subspace $\mathbf{1}_n^\perp$, and let $U_l^{\text{core}}, V_l^{\text{core}} \in \mathbb{R}^{(n-1)\times(n-1)}$ be orthogonal matrices. The parameterization,*

$$\mathcal{H}_l^{\text{disp}} = (U_{\mathcal{Z}} U_l^{\text{core}})\, \Sigma_l\, (U_{\mathcal{Z}} V_l^{\text{core}})^\top \tag{7}$$

provides a complete representation of the zero-marginal subspace $\mathcal{Z}_n$. Furthermore, bounding the singular values in Σ_l provides exact control over the spectral norm $\|\mathcal{H}_l^{\text{disp}}\|_2$.

Remark 2. As established in Proposition 2, parameterizing $\mathcal{H}_l^{\text{disp}}$ satisfying $\|\mathcal{H}_l^{\text{disp}}\|_2 \le 1$ reduces to generating the orthogonal matrices $U_l^{\text{core}}, V_l^{\text{core}}$ and bounding Σ_l.

Given the flattened vector $\boldsymbol{x}_l \in \mathbb{R}^{1\times nC}$ from the expanded streams X_l at l-th layer, we dynamically generate these components:

$$\begin{cases} \boldsymbol{x}'_l = \text{RMSNorm}(\boldsymbol{x}_l) \\ U_l^{\text{core}} = \text{Cayley}\left(\text{skew}\left(\gamma_l^U \tanh(\tau_l^U(\boldsymbol{x}'_l W_l^U) + \boldsymbol{b}_l^U)\right)\right) \\ V_l^{\text{core}} = \text{Cayley}\left(\text{skew}\left(\gamma_l^V \tanh(\tau_l^V(\boldsymbol{x}'_l W_l^V) + \boldsymbol{b}_l^V)\right)\right) \\ \Sigma_l = \text{diag}\left(\tanh\left(\tau_l^S(\boldsymbol{x}'_l W_l^S) + \boldsymbol{b}_l^S\right)\right) \end{cases} \tag{8}$$

Here, $W_l^U, W_l^V \in \mathbb{R}^{nC\times k}$ (with $k = \frac{1}{2}(n-1)(n-2)$) and $W_l^S \in \mathbb{R}^{nC\times(n-1)}$ are learnable projection weights, while $\boldsymbol{b}_l^U, \boldsymbol{b}_l^V \in \mathbb{R}^{1\times k}$ and $\boldsymbol{b}_l^S \in \mathbb{R}^{1\times(n-1)}$ are learnable biases. The parameters $\tau_l^U, \tau_l^V, \tau_l^S$ are trainable scalar factors, and γ_l^U, γ_l^V act as learnable rotation magnitude gates. The operator $\text{skew}(\cdot)$ constructs a skew-symmetric matrix in $\mathbb{R}^{(n-1)\times(n-1)}$ by populating its strictly upper-triangular entries with the k activated outputs and completing the lower triangle via anti-symmetry. $\text{Cayley}(\cdot)$ is the Cayley transform which maps the constructed skew-symmetric matrices into orthogonal matrices U_l^{core} and V_l^{core}. $\text{diag}(\cdot)$ maps the output $n-1$ singular values onto the main diagonal, where the $\tanh(\cdot)$ bounds the singular values $|(\Sigma_l)_{i,i}| \le 1$ for $i = 1 \dots n-1$, thus ensuring $\|\mathcal{H}_l^{\text{disp}}\|_2 \le 1$.

Substituting the parametrized matrices and adding the translation J yields the final residual matrix:

$$\mathcal{H}_l^{\text{res}} = J + (U_{\mathcal{Z}} U_l^{\text{core}})\Sigma_l (U_{\mathcal{Z}} V_l^{\text{core}})^\top \tag{9}$$

Figure 4 illustrates the geometric intuition of our parameterization and its associated constraint spaces. Finally, except for the above residual matrix $\mathcal{H}_l^{\text{res}}$, our s^{2}HC keeps other structures of mHC unchanged.

6 Experiment

Models and Datasets. To demonstrate the generalizability of our method, we evaluate s^{2}HC across diverse model architectures on configurations both with and without tied word embeddings. Accordingly, leveraging the LitGPT framework [14], we evaluate s^{2}HC on three distinct base models: an untied Qwen3-0.6B variant (∼0.75B parameters) [12], the default tied Gemma3-1B [15] and Qwen2.5-1.5B [16]. For all models, all standard architectural hyperparameters are strictly preserved. The sole architectural modification is the expansion of standard residual connections into corresponding hyper-connections. Models are trained from scratch on the FineWebEdu dataset [17].

Table 1: Evaluation results (accuracy %) on benchmarks. Higher values indicate better performance.

Method	WinoGrande	BoolQ	PIQA	HellaSwag	ARC-E	ARC-C	MMLU	TruthfulQA	Avg.
					Qwen3				
RC	50.6	58.4	67.1	42.0	58.8	29.7	25.3	**36.7**	46.1
mHC	53.5	59.8	67.7	43.6	56.6	29.3	25.4	32.9	46.1
mHC-lite	53.4	60.1	67.1	43.1	56.4	28.7	24.3	33.1	45.8
KromHC	49.5	**60.4**	67.7	43.2	56.3	28.3	24.7	32.4	45.3
s^{2}HC (Ours)	**53.6**	60.3	**68.3**	**44.0**	**59.8**	**30.9**	**25.6**	34.4	**47.1**
					Gemma3				
RC	52.5	**62.1**	69.3	47.0	65.3	35.0	24.3	35.7	48.9
mHC	51.7	60.7	69.6	**48.2**	67.0	33.9	25.2	**38.9**	49.4
mHC-lite	53.2	61.3	70.2	48.1	67.3	36.3	25.4	37.3	49.9
KromHC	53.3	59.4	70.6	48.0	**67.6**	**36.6**	**25.8**	38.4	50.0
s^{2}HC (Ours)	**55.3**	60.6	**70.7**	**48.2**	67.5	36.0	25.3	37.9	**50.2**
					Qwen2.5				
RC	51.4	60.9	70.9	49.9	64.2	34.5	25.4	34.9	49.0
mHC	54.5	62.1	**71.6**	**51.7**	61.2	35.2	24.5	35.3	49.5
mHC-lite	53.4	**62.4**	70.7	51.3	63.3	34.9	25.1	35.6	49.6
KromHC	55.1	62.3	71.3	50.4	64.3	35.3	**25.7**	31.5	49.5
s^{2}HC (Ours)	**55.2**	62.2	**71.6**	51.5	**66.5**	**37.0**	25.1	**35.7**	**50.7**

Adhering to compute-optimal scaling laws [18], we allocate a $20\times$ training token budget relative to their actual parameter counts (including embedding parameters). This yields approximately 15B tokens for Qwen3, 20B tokens for Gemma3 and 30B tokens for Qwen2.5.

Baselines. We evaluate different constrained hyper-connection paradigms including Manifold-Constrained Hyper-Connections using Sinkhorn-Knopp iterations (mHC [1]), and its permutation-based variants (mHC-lite [2] and KromHC [3]). Aligned with their settings, the number of residual streams for all hyper-connections including our s^{2}HC is set to $n = 4$. We further include the standard single-stream residual connection (RC [4]), a reference for assessing the magnitude of the observed improvements.

Initialization. Following mHC/mHC-lite/KromHC, we adopt their initialization schemes so that each variant reduces to a standard residual connection (identity mapping) at initialization. s^{2}HC is also set as identity mapping at initialization.

Evaluation. To assess the performance of these hyper-connections, we evaluate the trained models on eight diverse benchmarks, including PIQA [19], WinoGrande [20], HellaSwag [21], BoolQ [22], ARC-E and ARC-C [23], MMLU [24], and TruthfulQA [25], using the lm-eval harness [26]. These benchmarks cover a broad range of capabilities, including commonsense reasoning, reading comprehension, multi-domain knowledge reasoning, and factual truthfulness.

Detailed experiment setup is provided in Appendix E.

6.1 Performance and Expressivity Analysis

Performance. As shown in Table 1, s^{2}HC demonstrates cross-model consistency, achieving the highest average accuracy on all base models. Specifically on Qwen3, s^{2}HC ranks first on six out of eight benchmarks, with notable gains over the strongest hyper-connection baseline on reasoning (+3.2% on ARC-E, +1.6% on ARC-C). This trend also holds on Qwen2.5 (+2.2% on ARC-E, +1.7% on ARC-C), while the baselines themselves remain unstable across the three models. Against the standard residual connection (RC), the strongest doubly stochastic variant improves by +0.0%, +1.1% and +0.6% average points on Qwen3, Gemma3 and Qwen2.5, whereas s^{2}HC adds a further +1.0%, +0.2% and +1.1% on top of them. The gain s^{2}HC contributes over existing constrained hyper-connections is thus comparable to what those methods obtain over RC.

Expressivity. The right panel of Figure 5 traces the spectrum of $\mathcal{H}_l^{\text{res}}$ within the subspace orthogonal to the mean component (i.e., the singular values of $\mathcal{H}_l^{\text{disp}}$). The two leading subdominant singular values (σ_1, σ_2) remain near 1 across layers, while the smallest one (σ_3) fluctuates below 1, reaching as low as 0.77. This contrast reveals that, instead of rigidly preserving all directions, the model selectively retains the primary ones while adaptively attenuating others, exercising exactly the spectral freedom that s^{2}HC restores. The left panel shows how this manifests in the mean pairwise cosine

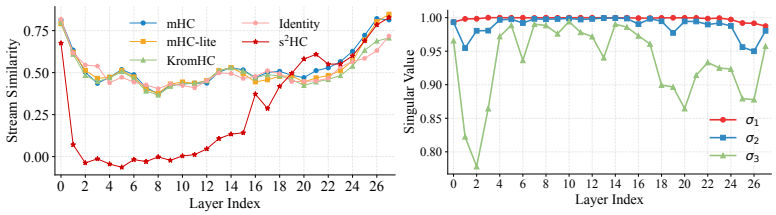

Figure 5: Expressivity analysis of s^{2}HC on Qwen3. **Left:** Mean pairwise cosine similarity among residual streams after being mixed by $\mathcal{H}_l^{\text{res}}$ across model layers. Identity denotes the baseline where $\mathcal{H}_l^{\text{res}}$ is fixed as an identity matrix. **Right:** the spectrum of $\mathcal{H}_l^{\text{res}}$ within the subspace orthogonal to the mean component. More results on Gemma3 and Qwen2.5 can be found in Appendix F.

similarity among streams. Unlike the baselines, whose trajectories stay close to the identity mapping above 0.4, the similarity under s^{2}HC departs markedly, dropping near 0 across early and middle layers and recovering in deeper ones. Cross-stream variations are thus introduced where they are useful and discarded where they are not. This aligns with our observation (Appendix A) that s^{2}HC leverages expressive, non-degenerate residual matrices for active stream feature mixing.

Ablation Study. We verify the effectiveness of s^{2}HC's dynamic cross-stream interactions, and further show that s^{2}HC outperforms an orthogonal-constrained baseline, confirming the benefit of s^{2}HC's flexible singular value control. Details are in Appendix B.

6.2 Stability Analysis

Training Stability. Figure 6 shows the gradient norms during training. The unconstrained Hyper-Connections (HC) [8] is unstable, and its gradient fluctuates severely during training. The constrained hyper-connections (s^{2}HC, mHC, mHC-lite, and KromHC) stabilize the optimization. As shown, all constrained variants successfully suppress the explosive gradient spikes of HC, rapidly converging to a consistently low and stable gradient norm. However, while mHC, mHC-lite, and KromHC achieve this gradient stability at the cost of collapsing into the identity mapping, s^{2}HC maintains optimization stability without passively sacrificing expressive, non-degenerate feature mixing.

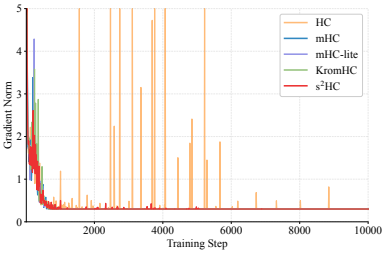

Figure 6: Gradient norm dynamics during training for the Qwen3 model on FineWebEdu. HC denotes the unconstrained Hyper-Connections [8]. The unconstrained HC exhibits exploding gradients, clamped at 5 for visualization. Constrained hyper-connection paradigms show stable gradient trajectories.

Hyper-Connection Stability. We evaluate the layer-wise and propagation stability of residual matrices for mHC, mHC-lite, KromHC, and s^{2}HC. Results are shown in Figure 7. As illustrated in the left panel, mHC fails to strictly satisfy the doubly stochastic constraint despite 20 Sinkhorn-Knopp iterations. The column sums of its layer-wise residual matrices $\mathcal{H}_l^{\text{res}}$ deviate from 1.0, with a peak reaching 2.0. These column-sum deviations accumulate across layers as illustrated in the middle panel: the composite mapping $\prod_{l=0}^{27} \mathcal{H}_{27-l}^{\text{res}}$ across 28 layers for mHC exhibits a pronounced shift of its column sums away from 1.0, with accumulated outliers spiking to 3.6. In contrast, s^{2}HC, mHC-lite and KromHC maintain unit column sums for both the layer-wise residual matrices and the global composite mapping, thereby preserving the mean component of residual streams without signal diminishment or amplification.

Furthermore, the right panel tracks the spectral norm of the cumulative composite mapping $\prod_{l=0}^{L-1} \mathcal{H}_{L-1-l}^{\text{res}}$ across the first L layers. It reveals that mHC fails to bound this cumulative norm, which increases as model depth grows. Conversely, s^{2}HC, mHC-lite, and KromHC constrain the cumulative spectral norm around 1.0, effectively preventing depth-induced gradient amplification.

6.3 Efficiency and Parameterization Overhead

We evaluate the end-to-end training throughput (tokens/s) and parameterization overhead (params) of mHC, mHC-lite, KromHC, and s^{2}HC based on the Qwen3 model. All experiments are conducted on

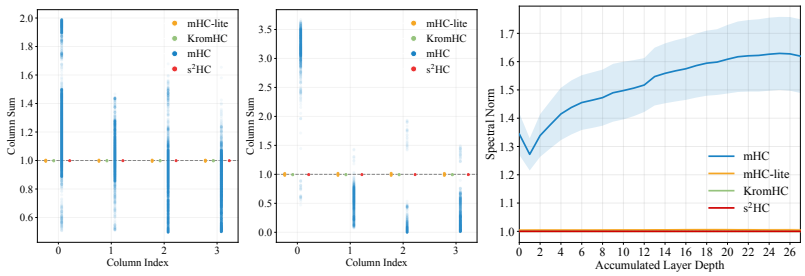

Figure 7: Empirical validation of the stability of constrained residual matrices for mHC, mHC-lite, and s^{2}HC. Statistics are computed from the trained Qwen3 model with 4 residual streams on the 1024 samples. **Left:** Column-sum distribution of layer-wise residual matrices $\mathcal{H}_l^{\text{res}}$ for each layer. **Middle:** Column-sum distribution of the full-model composite mapping $\prod_{l=0}^{27} \mathcal{H}_{27-l}^{\text{res}}$. **Right:** Spectral norm of the cumulative composite mapping $\| \prod_{l=0}^{L-1} \mathcal{H}_{L-1-l}^{\text{res}} \|_2$ across the first L layers. Lines denote means across samples while shaded regions indicate standard deviations. More results on Gemma3 and Qwen2.5 are provided in Appendix F.

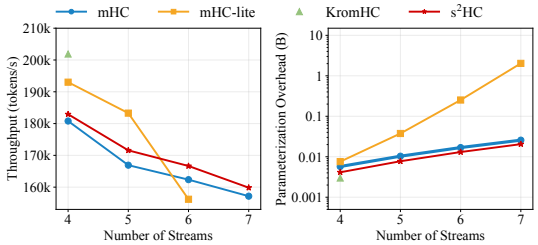

Figure 8: Training throughput and parameterization overhead across varying numbers of streams n. The mHC result is based on our PyTorch re-implementation and may underestimate the throughput of the specialized-kernel implementation in [1]. KromHC is plotted as discrete points at $n = 4$ due to its restriction to powers of two under the default setting. **Left:** Training throughput evaluated on a 4×H200 setup with a fixed global batch size. The data point for mHC-lite at $n = 7$ is omitted, as its factorial parameter explosion exceeds the GPU memory budget, resulting in an Out-Of-Memory failure. **Right:** Parameterization overhead (params) of the hyper-connections, excluding base model weights.

a 4×H200 GPU setup with a fixed sequence length of 2048 tokens and a constant 512 global batch size (micro-batch 8).

As shown in Figure 8, s^{2}HC scales stably with compact parameterizations. For s^{2}HC, the projection weights (W_l^U, W_l^V, W_l^S) have a combined parameter complexity corresponding to $\mathbb{R}^{nC\times(n-1)^2}$. Similar to mHC ($W_l^{\text{res}} \in \mathbb{R}^{nC\times n^2}$), they grow cubically with the number of streams n. In contrast, the auxiliary parameters of mHC-lite grow factorially with n. This difference in scaling explains why s^{2}HC and mHC maintain stable throughput, whereas the throughput of mHC-lite degrades rapidly beyond $n = 5$. At $n = 7$, the auxiliary parameters of mHC-lite reach 2B, approximately 3.3 times the size of the base model, exceeding our GPU memory and resulting in an Out-of-Memory failure while mHC and s^{2}HC introduce only 26M and 20M auxiliary parameters respectively.

While KromHC alleviates the factorial complexity via Kronecker products to achieve competitive throughput, its efficiency critically depends on the integer factorization of n. When n includes large prime factors and cannot be decomposed into small integers, the method effectively reverts to intractable factorial complexity [3]. Consequently, in its default settings, KromHC restricts the number of streams to powers of two (e.g., $n = 4, 8$) to preserve the efficiency of the parameterization. This limitation prevents it from efficiently supporting arbitrary stream widths. In contrast, s^{2}HC avoids both the factorial scaling of mHC-lite and the structural constraints of KromHC, providing a scalable and flexible solution for arbitrary residual stream counts.

7 Conclusion

In this paper, we introduce Spectral-Sphere-Constrained Hyper-Connections (s^{2}HC) to eliminate factorial parameterization overhead, and alleviate the identity degeneration and expressivity bottleneck

inherent to prior doubly stochastic constraint hyper-connections. By reformulating the constraint to an affine-constrained spectral norm sphere, s^2HC overcomes these limitations. We hope that the expressivity, stability, and parameterization efficiency of s^2HC can potentially illuminate new pathways to contribute to ongoing research efforts to diversify connectivity patterns beyond standard residual architectures. Further discussion regarding the broader impacts and limitations of s^2HC is provided in Appendix C.

References

[1] Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, et al. mhc: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025.

[2] Yongyi Yang and Jianyang Gao. mhc-lite: You don't need 20 sinkhorn-knopp iterations. *arXiv preprint arXiv:2601.05732*, 2026.

[3] Wuyang Zhou, Yuxuan Gu, Giorgos Iacovides, and Danilo Mandic. Kromhc: Manifold-constrained hyper-connections with kronecker-product residual matrices. *arXiv preprint arXiv:2601.21579*, 2026.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[6] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. Hyper-connections. *arXiv preprint arXiv:2409.19606*, 2024.

[9] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[10] Defa Zhu, Hongzhi Huang, Jundong Zhou, Zihao Huang, Yutao Zeng, Banggu Wu, Qiyang Min, and Xun Zhou. Frac-connections: Fractional extension of hyper-connections. *arXiv preprint arXiv:2503.14125*, 2025.

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

[12] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[13] Yizhi Liu. The homogeneity trap: Spectral collapse in doubly-stochastic deep networks. *arXiv preprint arXiv:2601.02080*, 2026.

[14] Lightning AI. Litgpt. https://github.com/Lightning-AI/litgpt, 2023.

[15] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[16] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[17] Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

[18] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[19] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*, 2020.

[20] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

[21] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800, 2019.

[22] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, 2019.

[23] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[24] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[25] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[26] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024.

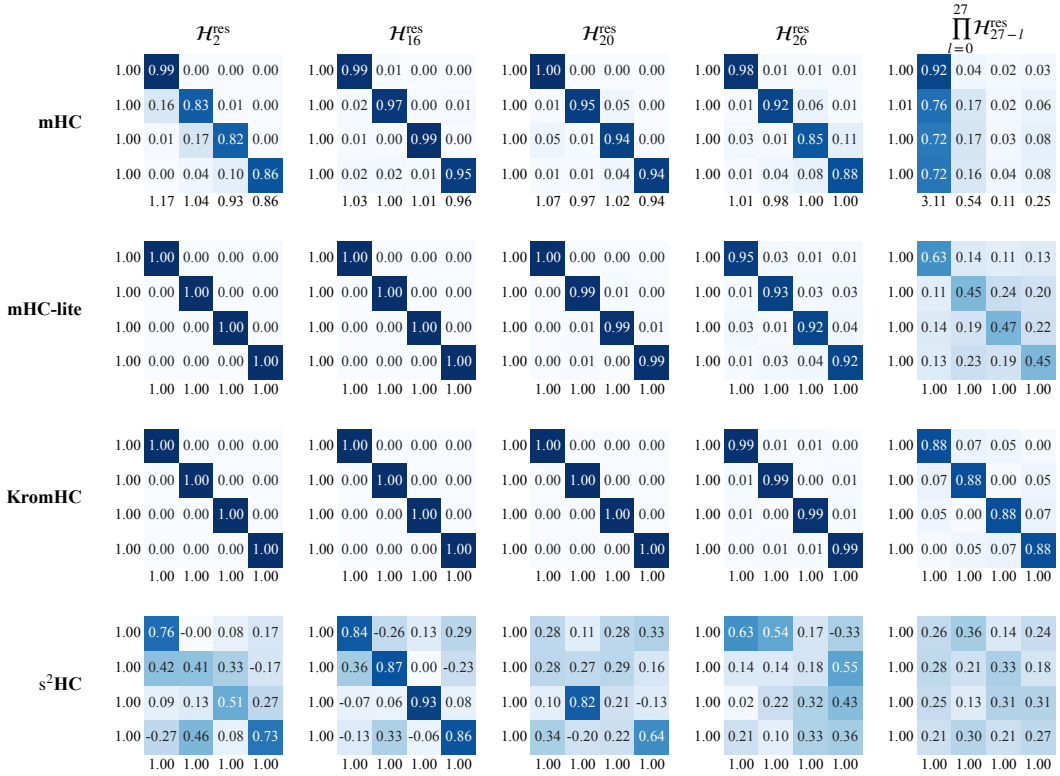

Figure 9: Visualization of layer-wise residual matrices and composite mappings on Qwen3. The figure displays single-layer residual matrices $\mathcal{H}_l^{\text{res}}$ at selected depths ($l \in \{2, 16, 20, 26\}$), alongside their end-to-end composite mappings ($\prod_{l=0}^{27} \mathcal{H}_{27-l}^{\text{res}}$) across mHC, mHC-lite, KromHC, and s^{2}HC (Ours). Each matrix is computed by averaging over all tokens within a sequence. The numerical annotations along the vertical and horizontal axes denote the row sums and column sums, respectively.

A Expressivity Analysis

Residual Matrix Dynamics. We provide the visualizations (Figure 9, 12) and entry distributions (Figure 10) of the learned layer-wise residual matrices $\mathcal{H}_l^{\text{res}}$ based on the trained Qwen3, Gemma3 and Qwen2.5 model. For all doubly stochastic constrained hyper-connections (mHC, mHC-lite, and KromHC), matrix entries sharply concentrate around 0.00 and 1.00, producing sparse and near-identity layer-wise residual matrices. These observations are consistent with the identity degeneration discussed in § 4. In contrast, s^{2}HC exhibits a continuous entry distribution, indicating that it leverages non-degenerated residual matrices for active feature interactions. Importantly, this increased expressivity does not compromise stability. As shown in Figure 9, s^{2}HC maintains exact row and column sums of 1.00 across all depths, whereas mHC accumulates normalization errors, resulting in composite column sums ranging from 0.11 to 3.11.

B Ablation Study

In this section, we conduct two ablation studies to verify the effectiveness of s^{2}HC. In each study, only the studied component is modified, while all other settings follow those of the main experiments in § 6.

Comparison with Identity Mapping Baseline. To verify the effectiveness of s^{2}HC's dynamic cross-stream interactions, we conduct an ablation study using an unmixed multi-stream baseline. In this setup, we remove the learnable residual matrix $\mathcal{H}_l^{\text{res}}$ (while retaining $\mathcal{H}_l^{\text{pre}}, \mathcal{H}_l^{\text{post}}$), which is equivalent to fixing the connection to a static multi-stream identity mapping. The results are presented in Table 2.

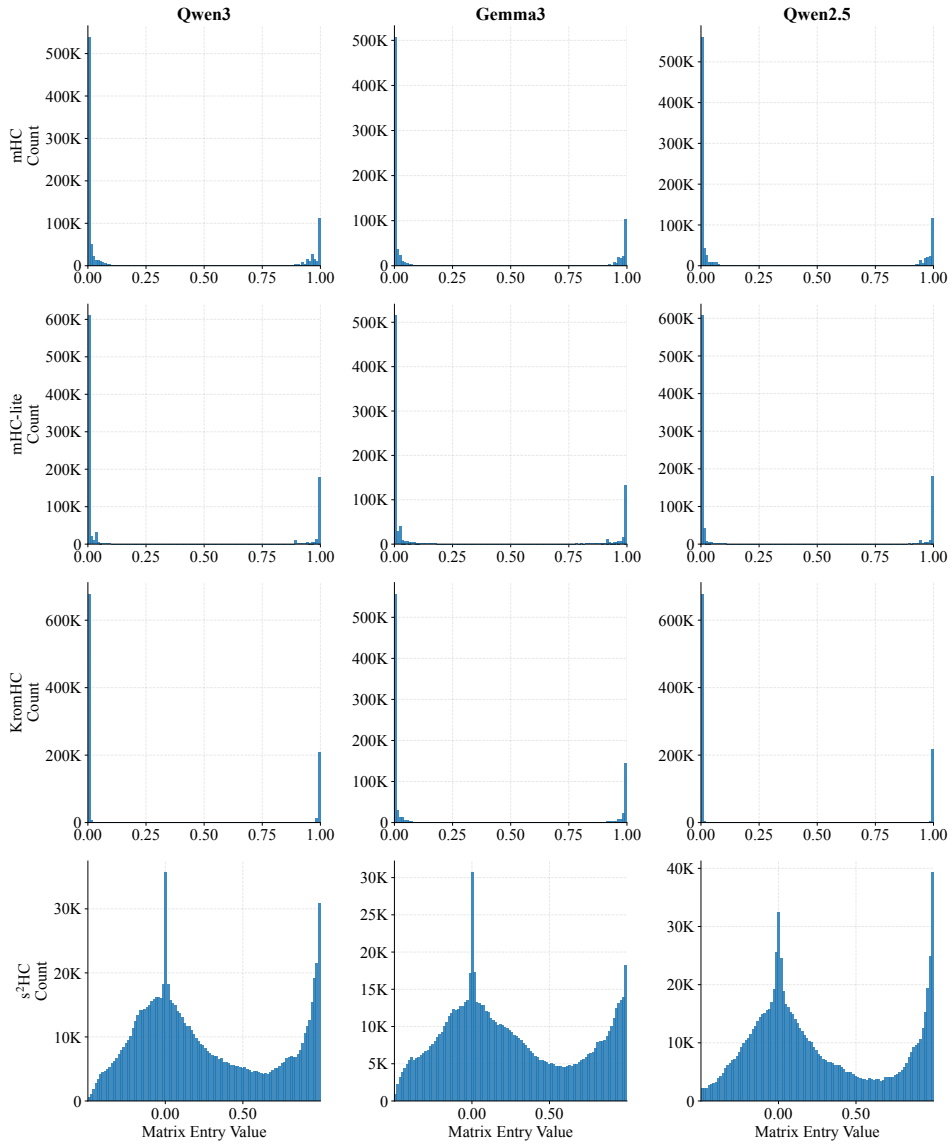

Figure 10: Distribution of layer-wise residual matrix entries in Qwen3 across mHC, mHC-lite, KromHC and our proposed s^2HC.

As shown, enabling dynamic stream mixing yields consistent improvements across all three models and most benchmarks, with the static baseline matching s^2HC on only a few individual tasks. This consistency indicates that s^2HC makes effective use of its learnable residual matrices under the spectral-sphere constraint, capturing beneficial cross-stream interactions beyond a static identity mapping.

Comparison with Orthogonal-Constrained Baseline. An intuitive way to prevent the suppression of the subdominant spectrum while retaining the idea of constrained hyper-connections is to restrict the residual matrix to be orthogonal. An orthogonal matrix fixes all singular values to one, which inherently bounds the spectral norm and avoids the collapse of subdominant singular values. To examine whether this design suffices, we dynamically generate orthogonal residual matrices from the normalized residual input $\boldsymbol{x}'_l \in \mathbb{R}^{1\times nC}$:

$$\mathcal{H}_l^{\text{res}} = \text{Cayley}\left(\text{skew}\left(\gamma_l^O \tanh(\tau_l^O(\boldsymbol{x}'_l W_l^O) + \boldsymbol{b}_l^O)\right)\right) \tag{10}$$

Table 2: Comparison with Identity Mapping Baseline.

$\mathcal{H}_l^{\text{res}}$	WinoGrande	BoolQ	PIQA	HellaSwag	ARC-E	ARC-C	MMLU	TruthfulQA
				Qwen3				
✗	53.3	**60.5**	67.4	43.2	58.9	**31.1**	24.0	34.2
✓	**53.6**	60.3	**68.3**	**44.0**	**59.8**	30.9	**25.6**	**34.4**
				Gemma3				
✗	52.9	59.6	70.3	48.0	67.0	35.4	24.3	36.8
✓	**55.3**	**60.6**	**70.7**	**48.2**	**67.5**	**36.0**	**25.3**	**37.9**
				Qwen2.5				
✗	**55.2**	60.3	**71.6**	50.2	65.3	35.1	**25.2**	33.1
✓	**55.2**	**62.2**	**71.6**	**51.5**	**66.5**	**37.0**	25.1	**35.7**

Table 3: Comparison with Orthogonal-Constrained Baseline.

	WinoGrande	BoolQ	PIQA	HellaSwag	ARC-E	ARC-C	MMLU	TruthfulQA
				Qwen3				
Orthogonal	52.8	**60.5**	67.9	43.9	**59.8**	30.4	24.3	**34.7**
s^2HC (Ours)	**53.6**	60.3	**68.3**	**44.0**	**59.8**	**30.9**	**25.6**	34.4
				Gemma3				
Orthogonal	53.1	59.5	**71.5**	**48.5**	65.7	33.8	25.2	**38.6**
s^2HC (Ours)	**55.3**	**60.6**	70.7	48.2	**67.5**	**36.0**	**25.3**	37.9
				Qwen2.5				
Orthogonal	52.9	60.1	70.2	50.1	**66.5**	33.7	**26.2**	33.9
s^2HC (Ours)	**55.2**	**62.2**	**71.6**	**51.5**	**66.5**	**37.0**	25.1	**35.7**

Here, $W_l^O \in \mathbb{R}^{nC \times q}$ (with $q = \frac{1}{2}n(n-1)$) is a learnable projection weight and $\boldsymbol{b}_l^O \in \mathbb{R}^{1 \times q}$ is a learnable bias, while τ_l^O is a trainable scalar factor and γ_l^O is a learnable rotation magnitude gate. The operator $\mathrm{skew}(\cdot)$ populates the strictly upper-triangular entries of an $n \times n$ skew-symmetric matrix with the q activated outputs, and $\mathrm{Cayley}(\cdot)$ maps it to an orthogonal matrix. Except for constraining $\mathcal{H}_l^{\text{res}}$ to be orthogonal, we keep all other components unchanged and identical to s^2HC.

The results are reported in Table 3. As shown, s^2HC consistently outperforms orthogonal-constrained baseline on most benchmarks across all three models. This indicates that beyond preventing spectral suppression, the ability of s^2HC to selectively attenuate singular values, rather than rigidly fixing them to one, contributes to its effectiveness. This substantiates our observation on the spectrum of the learned residual matrices of s^2HC as in Figure 5.

Moreover, the orthogonal constraint discards the mean-preservation property. This property underlies existing constrained hyper-connections [1, 2, 3] and is important for the training stability of deep models. Figure 11 shows the gradient norms of the orthogonal-constrained baseline and s^2HC during training on three models.

As shown, the orthogonal-constrained baseline exhibits persistently large and oscillating gradient norms throughout training. In contrast, s^2HC rapidly converges to a consistently low and stable gradient norm. This confirms that, without the mean-preservation guarantee, orthogonal residual matrices may induce unstable optimization, underscoring the necessity of the affine constraint in s^2HC.

C Discussion

Broader Impacts. In this paper, our analysis uncovers that while existing doubly stochastic constrained hyper-connections stabilize training by restricting residual matrices to a doubly stochastic manifold, this constraint causes the learned matrices to degenerate toward their identity initialization,

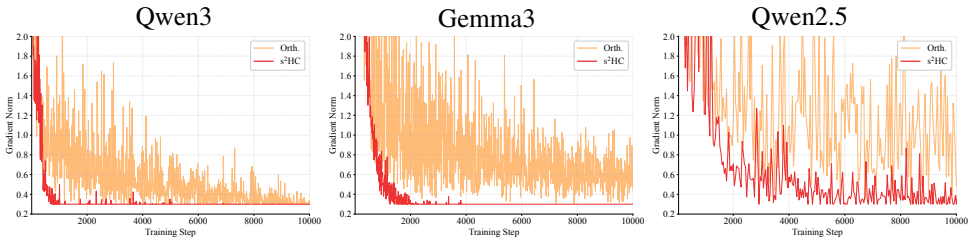

Figure 11: Gradient norm dynamics during training on three models. Orth. denotes the orthogonal-constrained baseline. The results are clamped at 2 for visualization.

running counter to the original intention of hyper-connections to facilitate active cross-stream feature interactions. Our proposed framework, Spectral-Sphere-Constrained Hyper-Connections (s^2HC), addresses these limitations, enabling active feature mixing, while ensuring training stability through an exact, closed-form spectral constraint. Furthermore, the proposed parameterization enforces this bound without introducing the factorial parameter overhead inherent in prior exact-constraint methods. Consequently, s^2HC renders hyper-connection designs generalizable, contributing to the ongoing research efforts to diversify connectivity patterns beyond standard residual architectures. By reconciling expressivity with parameterization efficiency, s^2HC offers the community a reference paradigm for deploying the hyper-connection architecture.

Limitations. While our s^2HC enhances model expressivity and eliminates the parameterization inefficiency of existing constrained hyper-connections, it inherits the memory access overhead intrinsic to the multi-stream residual architecture of existing hyper-connections [1, 2, 3, 8]. This increased demand for memory bandwidth can restrict actual training throughput at scale. However, this problem may be mitigated through system-level kernel optimizations, as demonstrated by mHC [1]. In addition, our experiments are conducted at the 0.6B–1.5B scale models under compute-optimal token budgets, reflecting the compute available to us. The validation of s^2HC at substantially larger model and token scales is still remaining as the future work.

D Proofs

D.1 Multiplicative Closure

Lemma 1 (Multiplicative Closure of $\mathcal{S}_n$)**.** *The affine-constrained spectral norm sphere* $\mathcal{S}_n = \{\mathcal{H} \in \mathcal{A}_n \mid \|\mathcal{H}\|_2 = 1\}$ *is closed under matrix multiplication.*

Proof. For any $\mathcal{H}_1, \mathcal{H}_2 \in \mathcal{S}_n$, the affine constraints are preserved under multiplication:

$$\begin{aligned}(\mathcal{H}_1\mathcal{H}_2)\mathbf{1}_n &= \mathcal{H}_1(\mathcal{H}_2\mathbf{1}_n) = \mathcal{H}_1\mathbf{1}_n = \mathbf{1}_n,\\ \mathbf{1}_n^\top(\mathcal{H}_1\mathcal{H}_2) &= (\mathbf{1}_n^\top\mathcal{H}_1)\mathcal{H}_2 = \mathbf{1}_n^\top\mathcal{H}_2 = \mathbf{1}_n^\top.\end{aligned} \tag{11}$$

Hence $\mathcal{H}_1\mathcal{H}_2 \in \mathcal{A}_n$.

For the spectral norm, submultiplicativity gives

$$\|\mathcal{H}_1\mathcal{H}_2\|_2 \le \|\mathcal{H}_1\|_2\|\mathcal{H}_2\|_2 = 1. \tag{12}$$

On the other hand, since $\mathcal{H}_1\mathcal{H}_2\mathbf{1}_n = \mathbf{1}_n$, we have

$$\|\mathcal{H}_1\mathcal{H}_2\|_2 = \sup_{\boldsymbol{x}\neq 0}\frac{\|\mathcal{H}_1\mathcal{H}_2\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \ge \frac{\|\mathcal{H}_1\mathcal{H}_2\mathbf{1}_n\|_2}{\|\mathbf{1}_n\|_2} = 1. \tag{13}$$

Therefore

$$1 \le \|\mathcal{H}_1\mathcal{H}_2\|_2 \le 1, \tag{14}$$

which implies $\|\mathcal{H}_1\mathcal{H}_2\|_2 = 1$ and hence $\mathcal{H}_1\mathcal{H}_2 \in \mathcal{S}_n$. □

D.2 Affine Subspace Translation

Lemma 2 (Affine Subspace Translation)**.** *Let $J = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ be the uniform matrix, and let $\mathcal{Z}_n = \{\mathcal{H} \in \mathbb{R}^{n\times n} \mid \mathcal{H}\mathbf{1}_n = \mathbf{0}_n, \mathbf{1}_n^\top\mathcal{H} = \mathbf{0}_n^\top\}$ be the zero-marginal subspace. The affine subspace $\mathcal{A}_n = \{\mathcal{H} \in \mathbb{R}^{n\times n} \mid \mathcal{H}\mathbf{1}_n = \mathbf{1}_n, \mathbf{1}_n^\top\mathcal{H} = \mathbf{1}_n^\top\}$ is a translation of $\mathcal{Z}_n$ by J, i.e., $\mathcal{A}_n = J + \mathcal{Z}_n$.*

Proof. Let $J = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$, which satisfies

$$J\mathbf{1}_n = \mathbf{1}_n, \qquad \mathbf{1}_n^\top J = \mathbf{1}_n^\top. \tag{15}$$

For any matrix $\mathcal{H}$,

$$\mathcal{H} \in \mathcal{A}_n \iff (\mathcal{H} - J)\mathbf{1}_n = \mathbf{0}_n,\ \mathbf{1}_n^\top(\mathcal{H} - J) = \mathbf{0}_n^\top \iff \mathcal{H} - J \in \mathcal{Z}_n. \tag{16}$$

Therefore,

$$\mathcal{A}_n = J + \mathcal{Z}_n. \tag{17}$$

□

D.3 Spectral Decoupling

Proof of Proposition 1. Let $\boldsymbol{x} \in \mathbb{R}^n$, which can be orthogonally decomposed $\boldsymbol{x} = \boldsymbol{x}_\parallel + \boldsymbol{x}_\perp$, where $\boldsymbol{x}_\parallel \in \operatorname{span}\{\mathbf{1}_n\}$ and $\boldsymbol{x}_\perp \in \mathbf{1}_n^\perp$. Thus, $\|\boldsymbol{x}\|_2^2 = \|\boldsymbol{x}_\parallel\|_2^2 + \|\boldsymbol{x}_\perp\|_2^2$.

Since $J = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ and $\mathcal{H}_l^{\text{disp}} \in \mathcal{Z}_n$, we have:

$$J\boldsymbol{x}_\perp = \mathbf{0}_n, \quad \mathcal{H}_l^{\text{disp}}\boldsymbol{x}_\parallel = \mathbf{0}_n, \quad \text{and} \quad \mathbf{1}_n^\top\mathcal{H}_l^{\text{disp}}\boldsymbol{x}_\perp = 0. \tag{18}$$

Applying $\mathcal{H}_l^{\text{res}} = J + \mathcal{H}_l^{\text{disp}}$ (Lemma 2) to $\boldsymbol{x}$, the action decouples:

$$\mathcal{H}_l^{\text{res}}\boldsymbol{x} = (J + \mathcal{H}_l^{\text{disp}})(\boldsymbol{x}_\parallel + \boldsymbol{x}_\perp) = J\boldsymbol{x}_\parallel + \mathcal{H}_l^{\text{disp}}\boldsymbol{x}_\perp. \tag{19}$$

Since $J\boldsymbol{x}_\parallel \in \operatorname{span}\{\mathbf{1}_n\}$ and $\mathcal{H}_l^{\text{disp}}\boldsymbol{x}_\perp \in \mathbf{1}_n^\perp$, the output terms are mutually orthogonal. Consequently:

$$\begin{aligned}
\|\mathcal{H}_l^{\text{res}}\boldsymbol{x}\|_2^2 &= \|J\boldsymbol{x}_\parallel\|_2^2 + \|\mathcal{H}_l^{\text{disp}}\boldsymbol{x}_\perp\|_2^2 \\
&\le \|J\|_2^2\|\boldsymbol{x}_\parallel\|_2^2 + \|\mathcal{H}_l^{\text{disp}}\|_2^2\|\boldsymbol{x}_\perp\|_2^2 \\
&\le \max\left(\|J\|_2^2, \|\mathcal{H}_l^{\text{disp}}\|_2^2\right)\left(\|\boldsymbol{x}_\parallel\|_2^2 + \|\boldsymbol{x}_\perp\|_2^2\right) \\
&= \max\left(\|J\|_2^2, \|\mathcal{H}_l^{\text{disp}}\|_2^2\right)\|\boldsymbol{x}\|_2^2.
\end{aligned} \tag{20}$$

Dividing by $\|\boldsymbol{x}\|_2^2$ and taking the supremum over $\boldsymbol{x} \neq \mathbf{0}_n$ explicitly establishes the upper bound:

$$\|\mathcal{H}_l^{\text{res}}\|_2 = \sup_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|\mathcal{H}_l^{\text{res}}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \le \max\left(\|J\|_2, \|\mathcal{H}_l^{\text{disp}}\|_2\right). \tag{21}$$

To establish the lower bound, we evaluate the quotient over specific subspaces.

For any non-zero vector $\boldsymbol{u} \in \operatorname{span}\{\mathbf{1}_n\}$, we have $\mathcal{H}_l^{\text{disp}}\boldsymbol{u} = \mathbf{0}_n$, which implies $\mathcal{H}_l^{\text{res}}\boldsymbol{u} = J\boldsymbol{u}$. Therefore,

$$\|\mathcal{H}_l^{\text{res}}\|_2 = \sup_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|\mathcal{H}_l^{\text{res}}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \ge \sup_{\boldsymbol{u}\in\operatorname{span}\{\mathbf{1}_n\},\,\boldsymbol{u}\neq\mathbf{0}} \frac{\|J\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} = \|J\|_2. \tag{22}$$

Similarly, for any non-zero vector $\boldsymbol{v} \in \mathbf{1}_n^\perp$, we have $J\boldsymbol{v} = \mathbf{0}$, which implies $\mathcal{H}_l^{\text{res}}\boldsymbol{v} = \mathcal{H}_l^{\text{disp}}\boldsymbol{v}$. Thus,

$$\|\mathcal{H}_l^{\text{res}}\|_2 = \sup_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|\mathcal{H}_l^{\text{res}}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \ge \sup_{\boldsymbol{v}\in\mathbf{1}_n^\perp,\,\boldsymbol{v}\neq\mathbf{0}} \frac{\|\mathcal{H}_l^{\text{disp}}\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2} = \|\mathcal{H}_l^{\text{disp}}\|_2. \tag{23}$$

Combining these conditions yields $\|\mathcal{H}_l^{\text{res}}\|_2 \ge \max\left(\|J\|_2, \|\mathcal{H}_l^{\text{disp}}\|_2\right)$. Since both the upper and lower bounds hold, exact equality is established:

$$\|\mathcal{H}_l^{\text{res}}\|_2 = \max\left(\|J\|_2, \|\mathcal{H}_l^{\text{disp}}\|_2\right). \tag{24}$$

□

D.4 Complete and Spectral-Controlled Parameterization

Proof of Proposition 2. Completeness. For any $\mathcal{H}_l^{\text{disp}} \in \mathcal{Z}_n$ with rank $r \leq n-1$, consider its compact SVD:

$$\mathcal{H}_l^{\text{disp}} = \tilde{U}_l \, \tilde{\Sigma}_l \, \tilde{V}_l^\top, \tag{25}$$

where $\tilde{U}_l, \tilde{V}_l \in \mathbb{R}^{n \times r}$ satisfy $\tilde{U}_l^\top \tilde{U}_l = \tilde{V}_l^\top \tilde{V}_l = I_r$, and $\tilde{\Sigma}_l \in \mathbb{R}^{r \times r}$ is a diagonal matrix containing non-zero singular values.

From $\mathcal{H}_l^{\text{disp}} \mathbf{1}_n = \mathbf{0}_n$ and $\mathbf{1}_n^\top \mathcal{H}_l^{\text{disp}} = \mathbf{0}_n^\top$, we obtain

$$\tilde{U}_l \, \tilde{\Sigma}_l \, (\tilde{V}_l^\top \mathbf{1}_n) = \mathbf{0}_n, \quad (\mathbf{1}_n^\top \tilde{U}_l) \, \tilde{\Sigma}_l \, \tilde{V}_l^\top = \mathbf{0}_n^\top. \tag{26}$$

Since $\tilde{U}_l$ and $\tilde{V}_l$ have full column rank and $\tilde{\Sigma}_l$ is invertible, this implies

$$\tilde{V}_l^\top \mathbf{1}_n = \mathbf{0}_r, \quad \mathbf{1}_n^\top \tilde{U}_l = \mathbf{0}_r^\top, \tag{27}$$

hence the column spaces of $\tilde{U}_l$ and $\tilde{V}_l$ lie in $\mathbf{1}_n^\perp$.

To align with the parameterization size $n-1$, since $\mathbf{1}_n^\perp$ is an $(n-1)$-dimensional subspace, we can find $(n-1-r)$ orthonormal vectors in $\mathbf{1}_n^\perp$ to expand $\tilde{U}_l$ and $\tilde{V}_l$ into $U_l, V_l \in \mathbb{R}^{n \times (n-1)}$, such that

$$\begin{aligned} &\text{col}(U_l), \ \text{col}(V_l) \subset \mathbf{1}_n^\perp, \\ &U_l^\top U_l = V_l^\top V_l = I_{n-1}. \end{aligned} \tag{28}$$

By padding $\tilde{\Sigma}_l$ with zeros to form an $\mathbb{R}^{(n-1)\times(n-1)}$ diagonal matrix Σ_l, we have equivalently:

$$\mathcal{H}_l^{\text{disp}} = U_l \, \Sigma_l \, V_l^\top. \tag{29}$$

Since $\text{col}(U_l), \text{col}(V_l) \subset \mathbf{1}_n^\perp$ and $U_{\mathcal{Z}}$ is an orthonormal basis of $\mathbf{1}_n^\perp$, we have

$$U_l = U_{\mathcal{Z}} \, U_{\mathcal{Z}}^\top U_l, \quad V_l = U_{\mathcal{Z}} \, U_{\mathcal{Z}}^\top V_l. \tag{30}$$

Define $U_l^{\text{core}} = U_{\mathcal{Z}}^\top U_l$ and $V_l^{\text{core}} = U_{\mathcal{Z}}^\top V_l$. Then

$$\begin{aligned} (U_l^{\text{core}})^\top U_l^{\text{core}} &= U_l^\top (U_{\mathcal{Z}} U_{\mathcal{Z}}^\top U_l) = U_l^\top U_l = I_{n-1}, \\ (V_l^{\text{core}})^\top V_l^{\text{core}} &= V_l^\top (U_{\mathcal{Z}} U_{\mathcal{Z}}^\top V_l) = V_l^\top V_l = I_{n-1}. \end{aligned} \tag{31}$$

Therefore any $\mathcal{H}_l^{\text{disp}} \in \mathcal{Z}_n$ can be written as

$$\mathcal{H}_l^{\text{disp}} = (U_{\mathcal{Z}} U_l^{\text{core}}) \Sigma_l (U_{\mathcal{Z}} V_l^{\text{core}})^\top. \tag{32}$$

Spectral norm preservation. Given the parameterization $U_l = U_{\mathcal{Z}} U_l^{\text{core}}$ and $V_l = U_{\mathcal{Z}} V_l^{\text{core}}$ with orthogonal matrices $U_l^{\text{core}}, V_l^{\text{core}}$, we have:

$$U_l^\top U_l = (U_l^{\text{core}})^\top (U_{\mathcal{Z}}^\top U_{\mathcal{Z}}) U_l^{\text{core}} = I_{n-1}, \tag{33}$$

$$V_l^\top V_l = (V_l^{\text{core}})^\top (U_{\mathcal{Z}}^\top U_{\mathcal{Z}}) V_l^{\text{core}} = I_{n-1}. \tag{34}$$

Thus both U_l and V_l have exact orthonormal columns. Then the spectral norm satisfies:

$$\|\mathcal{H}_l^{\text{disp}}\|_2 = \|U_l \Sigma_l V_l^\top\|_2 = \|\Sigma_l\|_2 = \max_i |(\Sigma_l)_{i,i}|. \tag{35}$$

Hence bounding Σ_l directly controls the spectral norm of $\mathcal{H}_l^{\text{disp}}$. □

E Experiment Setup

Initialization. In all cases, W_l^{res}, W_l^{pre}, and W_l^{post} are initialized to zero; α_l^{res}, α_l^{pre}, and α_l^{post} are set to 0.01; and $\boldsymbol{b}_l^{\text{pre}}$ and $\boldsymbol{b}_l^{\text{post}}$ are initialized to -1 except for a single entry set to 1. The residual branch differs across variants. In mHC, off-diagonal entries are set to -8 and diagonal entries to 0, yielding an identity-like matrix after Sinkhorn normalization. For mHC-lite and KromHC, $\boldsymbol{b}_l^{\text{res}}$ is set to -8 for all entries except the entry corresponding to the identity matrix, which is set to 0, so that after the softmax operation the weights concentrate on the identity matrix.

For s^{2}HC initialization, projection weights W_l^U, W_l^V, W_l^S are zero-initialized. We set the biases $\boldsymbol{b}_l^U, \boldsymbol{b}_l^V$ to zero, so that $U_l^{\text{core}}, V_l^{\text{core}}$ reduce to identity after the Cayley transform. Simultaneously, we initialize $\boldsymbol{b}_l^S$ to 4, saturating the $\tanh$ activation to yield an identity matrix for Σ_l. Finally, the rotation magnitude gates γ_l^U, γ_l^V are set to 1, and the scaling factors $\tau_l^U, \tau_l^V, \tau_l^S$ are set to 0.01.

Training Setup. Our implementation builds upon Qwen3-0.6B [12], Gemma3-1B [15] and Qwen2.5-1.5B [16]. We set the max sequence length to 2048 and expand the standard residual connections into corresponding hyper-connections with all other hyperparameters of the architecture at their default values. Models are trained from scratch using the AdamW optimizer with a cosine learning rate schedule and linear warmup. We employ mixed-precision training (bfloat16) and apply gradient clipping throughout training.

All experiments are conducted on 4 NVIDIA H200 GPUs using PyTorch DistributedDataParallel (DDP). We fix the global batch size across all methods. The hyperparameters of training are summarized in Table 4.

Table 4: Training hyperparameters.

Setting	**Qwen3**	**Gemma3**	**Qwen2.5**
Micro-batch size per GPU	8	8	4
Global batch size	512	512	1024
Sequence length	2048	2048	2048
Training tokens	15B	20B	30B
Warmup iterations	1000	1000	1000
Weight decay	0.1	0.1	0.1
β_1	0.9	0.9	0.9
β_2	0.95	0.95	0.95
Gradient clipping norm	1.0	1.0	1.0
Initial learning rate	6×10^{-4}	4×10^{-4}	4×10^{-4}
Minimum learning rate	6×10^{-5}	4×10^{-5}	4×10^{-5}
Tie word embeddings	False	True	True

Evaluation Setup. Aligned with the Hugging Face Open LLM Leaderboard v1 [2], we utilize 25-shot for ARC-E and ARC-C, 10-shot for HellaSwag, 5-shot for Winogrande, and 0-shot for TruthfulQA. For PIQA and BoolQ, we adopt the widely established 0-shot configuration in the foundational model literature [6, 7]. We also adapt MMLU to a 0-shot configuration to prevent unintended truncation caused by exceeding our maximum sequence length. The configurations are summarized in Table 5.

Table 5: Configurations of number of shots for each evaluation benchmark.

Benchmarks	PIQA	BoolQ	MMLU	TruthfulQA	WinoGrande	HellaSwag	ARC-E	ARC-C
# Shots	0	0	0	0	5	10	25	25

F More Results

[2] https://huggingface.co/docs/leaderboards/open_llm_leaderboard/archive

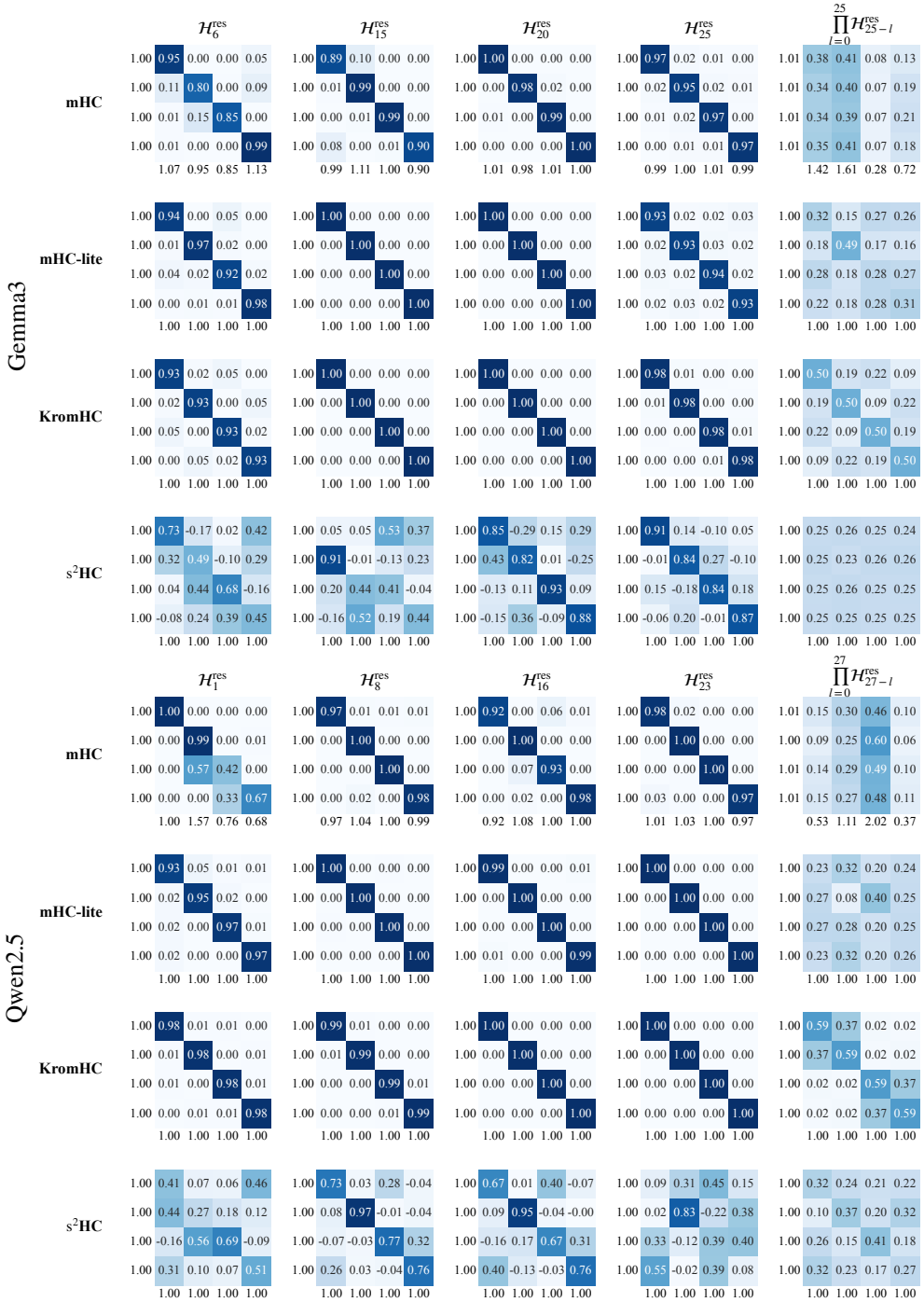

Figure 12: Visualization of layer-wise residual matrices and composite mappings. The figure displays single-layer residual matrices $\mathcal{H}^{\text{res}}_{l}$ at selected depths ($l \in \{6, 15, 20, 25\}$ for Gemma3 and $l \in \{1, 8, 16, 23\}$ for Qwen2.5), alongside their end-to-end composite mappings ($\prod_{l=0}^{25} \mathcal{H}^{\text{res}}_{25-l}$ for Gemma3 and $\prod_{l=0}^{27} \mathcal{H}^{\text{res}}_{27-l}$ for Qwen2.5) across mHC, mHC-lite, KromHC, and s^{2}HC (Ours). Each matrix is computed by averaging over all tokens within a sequence. The numerical annotations along the vertical and horizontal axes denote the row sums and column sums, respectively.

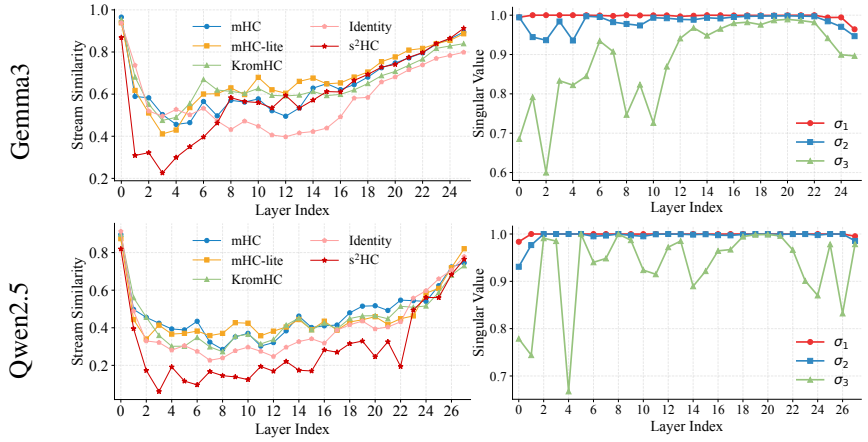

Figure 13: More expressivity analysis of s^{2}HC on Gemma3 and Qwen2.5. **Left:** Mean pairwise cosine similarity among residual streams after being mixed by $\mathcal{H}_l^{\text{res}}$ across model layers. Identity denotes the baseline where $\mathcal{H}_l^{\text{res}}$ is fixed as an identity matrix. **Right:** the spectrum of $\mathcal{H}_l^{\text{res}}$ within the subspace orthogonal to the mean component.

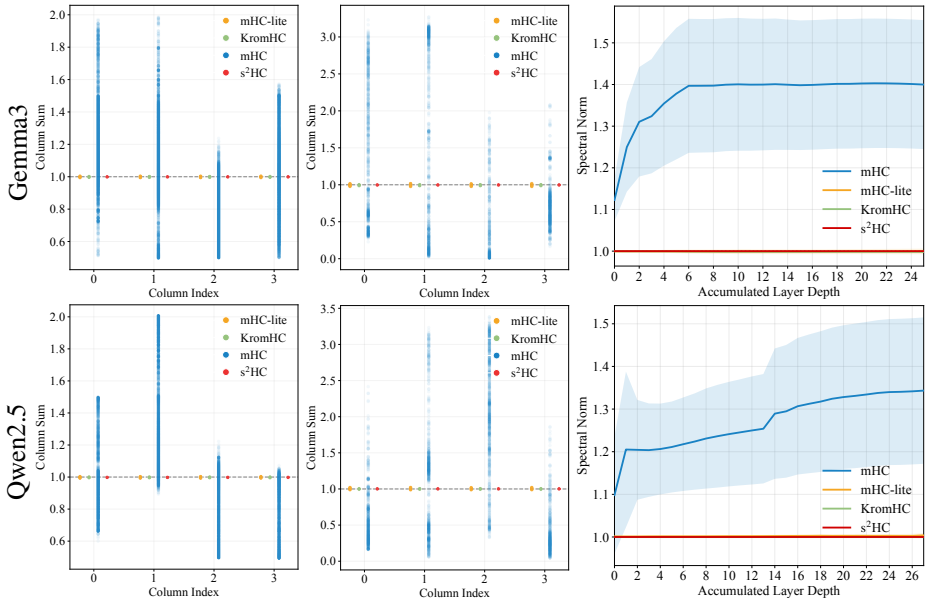

Figure 14: More stability analysis of constrained residual matrices for mHC, mHC-lite, KromHC and s^{2}HC. Statistics are computed from the trained Gemma3 and Qwen2.5 models with 4 residual streams. **Left:** Column-sum distribution of layer-wise residual matrices $\mathcal{H}_l^{\text{res}}$ for each layer. **Middle:** Column-sum distribution of the full-model composite mapping ($\prod_{l=0}^{25} \mathcal{H}_{25-l}^{\text{res}}$ for Gemma3 and $\prod_{l=0}^{27} \mathcal{H}_{27-l}^{\text{res}}$ for Qwen2.5). **Right:** Spectral norm of the cumulative composite mapping ($\| \prod_{l=0}^{L-1} \mathcal{H}_{L-1-l}^{\text{res}} \|_2$) across the first L layers. Lines denote means across samples while shaded regions indicate standard deviations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's main contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitation of our work in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Full proofs for all propositions and auxiliary theoretical results are provided in Appendices D.1 through D.4.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experimental details, including model architectures, training hyperparameters, and evaluation setups, are fully disclosed in Section 6 and Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The base models and the FineWebEdu dataset used in our experiments are publicly accessible, and we have released our code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All experimental details, including model architectures, training hyperparameters, and evaluation setups, are fully disclosed in Section 6 and Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to the massive computational requirements of LLM pre-training, we report results from a single training run, which is aligned with common practice in the related literature. To ensure reliability and reproducibility, all downstream evaluations are conducted using the widely adopted Language Model Evaluation Harness [26] across a diverse set of independent benchmarks. Furthermore, for our analytical experiments regarding matrix properties, our statistics are computed across 1024 data samples to ensure statistical reliability.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed information regarding the hardware (4 NVIDIA H200 GPUs), DDP training strategy, and total training compute (token budget) is provided in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research is a technical study on LLM model architecture using publicly available datasets and adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We have discussed the impacts of our work in Appendix C.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our research focuses on foundational architecture design using moderate-scale models and public datasets, which do not pose a high risk for significant misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external assets, including the FineWebEdu dataset and base models, are properly cited in the paper, and their open-source licenses are respected.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code for our proposed method is included with documentation and instructions for reproduction.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our core methodology was developed by the authors.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.