NEURAL DIVERSITY REGULARIZES HALLUCINATIONS IN SMALL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models hallucinate despite scaling. We propose neural diversity to reduce hallucination rates at fixed budgets. Our theory establishes predictive bounds connecting spectral diversity to hallucination probability and anticipates non-monotonic scaling where naive parallelism worsens reliability. We validate these predictions with parameter- and data-matched experiments across QA and summarization benchmarks, showing neural diversity causally reduces hallucinations. ND-LoRA achieves 17.9% hallucination reduction with 1.12× training cost, highlighting neural diversity as a third scaling axis orthogonal to parameters and tokens.

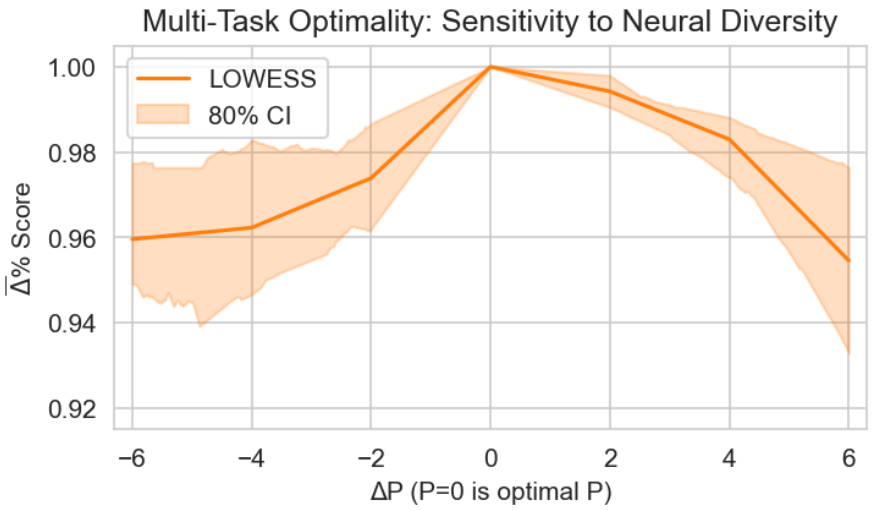

Figure 1: Maximizing reliability requires an optimal amount of neural diversity. The LOWESS fit with 80% bootstrapped confidence interval (shaded region) aggregates 39,936 evaluation points across tasks, showing normalized performance against $\Delta P = P - P_{\text{optimal}}$. Knowledge-intensive tasks show rapid dropoffs after optimal P, requiring precise diversity calibration, while truthfulness tasks exhibit greater robustness to sub-optimal P choices.

1 INTRODUCTION

Despite scaling to trillions of parameters, frontier models fabricate facts and assert falsehoods (Lin et al., 2021). This reliability crisis is acute for small language models (SLMs), which suffer disproportionately from hallucinations due to their compressed representations.

ParScale (Chen et al., 2025) shows P parallel streams achieve $O(\log P)$ gains with better memory/latency profiles than parameter scaling. However, naive parallelism often *degrades* reliability: $P = 8$ streams hallucinate more than $P = 4$, despite using twice the resources.

We identify the root cause as *representational collapse*: streams converge to similar features, reducing P computations to expensive redundancy.

Motivated by portfolio theory, we propose neural diversity: when streams are perfectly correlated ($\rho = 1$), we gain nothing; when decorrelated ($\rho \to 0$), signal-to-noise improves by $\sqrt{P}$.

We formalize this through a **neural diversity framework** connecting architectural design to hallucination probability. Our main contribution proves hallucination probability is bounded by cross-stream correlation: $\mathbb{P}(\text{hallucination}) \leq f(\sigma^2((1-\rho)/P+\rho), \mu^2)$. This reveals why naive parallelism fails and suggests minimizing cross-stream correlation reduces hallucinations.

To validate this, we introduce **ND-LoRA**, which combines independent LoRA adapters per stream with Barlow Twins regularization (Zbontar et al., 2021) to maintain diversity. With 1.3M trainable parameters across 4 streams, ND-LoRA reduces hallucination rates by up to 17.9% on HaluEval while maintaining general performance.

Our experiments reveal: (1) neural diversity causally mediates parallelism-reliability relationships, (2) optimal diversity is task-dependent, and (3) theoretical bounds accurately predict empirical performance.

Neural diversity represents a third axis of scaling, orthogonal to parameters and data. Rather than asking "how big?" or "how much?", diversity scaling asks "how different?"—particularly relevant for reliability without massive computational investment.

Our contributions are: (1) theory connecting diversity to hallucination bounds, (2) ND-LoRA achieving 17.9% reduction at 1.12× cost, (3) causal evidence for diversity-reliability mediation, and (4) a decorrelation principle for parallel architectures.

2 Theoretical Framework

Why does naive parallelism fail for reliability? When parallel streams converge to similar representations—a phenomenon we term *representational collapse*—the benefits vanish, leaving expensive redundancy. We establish the first rigorous connection between architectural diversity and hallucination probability, proving that decorrelated streams directly reduce reliability failures. This provides both an explanation for why naive parallelism fails and the mathematical foundation for our ND-LoRA architecture.

2.1 Preliminaries

Fix an input x. There are P parallel streams with scalar *margins* $m_i \equiv m_i(x)$. We study the equal-weight aggregate

$$M \triangleq \frac{1}{P}\sum_{i=1}^{P} m_i.$$

Assume per-stream mean $\mathbb{E}[m_i] = \mu$ (with $\mu > 0$ in our use of Cantelli below), per-stream variance $\mathrm{Var}(m_i) = \sigma^2$, and pairwise correlations $\rho_{ij} = \mathrm{Corr}(m_i, m_j)$. Let the *average* pairwise correlation be

$$\rho \triangleq \frac{2}{P(P-1)} \sum_{1 \leq i < j \leq P} \rho_{ij}.$$

Define the hallucination event $\mathrm{H} \equiv \{M \leq 0\}$.

The margins $m_i(x)$ represent stream confidence—positive indicates correctness, negative indicates error. When streams are perfectly correlated ($\rho = 1$), ensemble averaging provides no benefit. When decorrelated ($\rho \to 0$), M concentrates around its mean μ, reducing hallucination probability.

2.2 Neural Diversity Bounds Hallucination

We connect margin correlation to feature correlation through linearization: if predictions depend approximately linearly on representations at a chosen design layer, then reducing feature correlation reduces margin correlation.

Design-layer features and diversity. At a chosen *design layer*, stream i exposes a feature vector $z_i \in \mathbb{R}^d$. Let $\tilde{z}_i$ denote a whitened version (zero-mean, identity covariance within stream), and define the *cross-correlation* matrices $C_{ij} \triangleq \mathbb{E}[\tilde{z}_i \tilde{z}_j^\top]$.

Definition 1 (Neural Diversity Index)**.** *The spectral diversity index is*

$$\mathcal{D}_{\text{spec}} \;=\; \frac{1}{P(P-1)} \sum_{i \neq j} \|C_{ij}\|_2 \,.$$

Lower $\mathcal{D}_{\text{spec}}$ indicates greater diversity: $\mathcal{D}_{\text{spec}} = 0$ means perfectly decorrelated streams, while $\mathcal{D}_{\text{spec}} = 1$ means complete collapse.

We now connect stream correlation to hallucination probability through three steps: variance analysis, linearization, and the main bound.

Lemma 1 (Variance of Aggregated Margin)**.** *Under the assumptions above,*

$$\mathrm{Var}(M) \;=\; \sigma^2 \left(\frac{1-\rho}{P} + \rho \right).$$

Proof sketch. Expand $\mathrm{Var}(M)$ via $\mathrm{Var}(\frac{1}{P}\sum m_i)$, use $\mathrm{Var}(m_i) = \sigma^2$ and $\mathrm{Cov}(m_i, m_j) = \rho\,\sigma^2$ for $i \neq j$, then collect terms.

When $\rho = 0$, variance decreases as σ^2/P. When $\rho = 1$, variance stays at σ^2 regardless of P.

Lemma 2 (Correlation Bound)**.** *Assume the margins are locally linear in whitened features, $m_i = v_i^\top \tilde{z}_i$, and let $\sigma_i^2 \triangleq \mathrm{Var}(m_i)$. Then for $i \neq j$,*

$$\rho_{ij} \;=\; \frac{\mathrm{Cov}(m_i, m_j)}{\sigma_i \sigma_j} \;\leq\; \kappa_{ij}\, \|C_{ij}\|_2 \,, \qquad \kappa_{ij} \;\triangleq\; \frac{\|v_i\|\, \|v_j\|}{\sigma_i \sigma_j}.$$

Proof sketch. Compute $\mathrm{Cov}(m_i, m_j) = v_i^\top C_{ij} v_j$ and bound by the spectral norm: $|v_i^\top C_{ij} v_j| \leq \|v_i\|\, \|C_{ij}\|_2\, \|v_j\|$. Divide by $\sigma_i \sigma_j$.

Margin correlation ρ_{ij} is controlled by feature correlation $\|C_{ij}\|_2$. We now have a direct path from $\mathcal{D}_{\text{spec}}$ to hallucination probability.

Theorem 1 (Hallucination Probability Bound with Diversity)**.** *The hallucination probability satisfies*

$$\mathbb{P}(\mathrm{H}) \;\leq\; \frac{\sigma^2 \left(\frac{1-\bar{\kappa}\,\mathcal{D}_{\text{spec}}}{P} + \bar{\kappa}\,\mathcal{D}_{\text{spec}} \right)}{\sigma^2 \left(\frac{1-\bar{\kappa}\,\mathcal{D}_{\text{spec}}}{P} + \bar{\kappa}\,\mathcal{D}_{\text{spec}} \right) + \mu^2} \;+\; h_0,$$

where $0 \leq h_0 \leq 1$ is a constant.

Proof sketch. (1) By Lemma 1, $\mathrm{Var}(M)$ is increasing in ρ for $P > 1$. (2) By Corollary **??**, $\bar{\rho} \leq \bar{\kappa}\,\mathcal{D}_{\text{spec}}$, hence $\mathrm{Var}(M) \leq \sigma^2\big(\frac{1-\bar{\kappa}\mathcal{D}_{\text{spec}}}{P} + \bar{\kappa}\mathcal{D}_{\text{spec}}\big)$. (3) Plug this variance bound into Cantelli's bound. (4) If the linear/whitening conditions fail on a set of probability h_0, upper bound that set trivially by 1 and add h_0.

Lower $\mathcal{D}_{\text{spec}}$ reduces hallucination probability. The benefit scales with P.

2.3 Scaling Behavior

When correlation grows with P (without diversity regularization), the hallucination bound follows a U-shaped curve—initially decreasing but eventually increasing, explaining why naive parallelism fails.

Theorem 2 ("U-shape" of the Hallucination Bound under Rising Correlation)**.** *Suppose the average correlation increases with P as $\rho(P) = \rho_0 + \beta(P-1)^\gamma$ with $\beta > 0$ and $\gamma > 0$, and consider the Cantelli bound:*

$$\mathcal{B}(P) \;\triangleq\; \frac{\sigma^2 \left(\frac{1-\rho(P)}{P} + \rho(P) \right)}{\sigma^2 \left(\frac{1-\rho(P)}{P} + \rho(P) \right) + \mu^2}.$$

Then $\mathcal{B}(P)$ is decreasing *for P near 1 and* increasing *for P sufficiently large. Moreover, if $\gamma \geq 1$ (so $\rho''(P) \geq 0$), $\mathcal{B}(P)$ is convex on $(1, \infty)$ and thus has a* unique *minimizer $P_\star > 1$.*

Proof sketch. Work with the unnormalized variance factor $g(P) = \frac{1-\rho(P)}{P} + \rho(P)$. Differentiate: $g'(P) = \rho'(P)\big(1 - \frac{1}{P}\big) - \frac{1-\rho(P)}{P^2}$. At $P \downarrow 1$, the first term vanishes while the second is negative, so $g'(P) < 0$. As $P \to \infty$, $\rho'(P) > 0$ dominates the $O(P^{-2})$ negative term, so $g'(P) > 0$ eventually. If $\gamma \geq 1$, then $g''(P) \geq 0$, giving a unique minimizer. Since $\mathcal{B}(P)$ is an increasing function of $g(P)$, the same qualitative behavior holds for $\mathcal{B}(P)$.

There exists an optimal $P_\star$ that balances ensemble benefits against rising correlation. Explicit diversity regularization breaks this constraint by controlling $\mathcal{D}_{\text{spec}}$ directly.

3 ND-LoRA: A Practical Demonstration

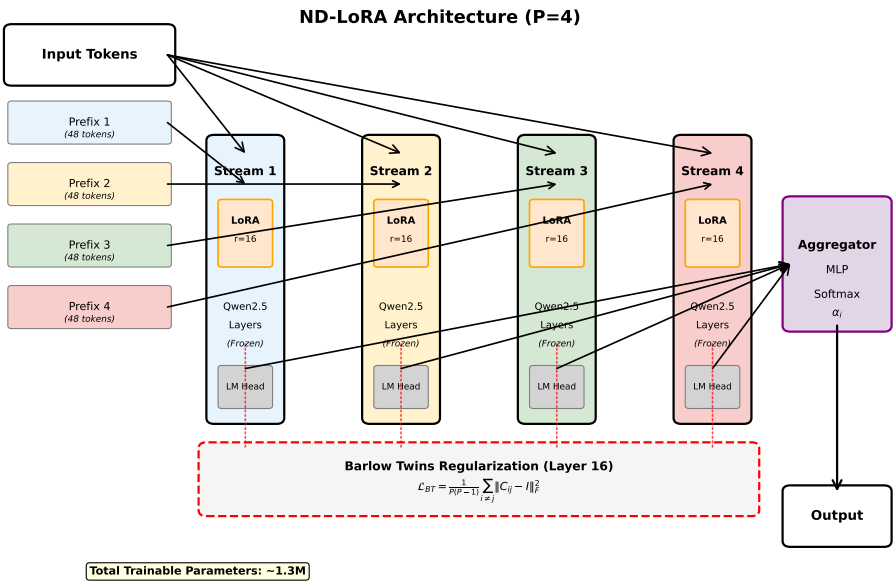

Figure 2: ND-LoRA architecture with $P = 4$ parallel streams. Each stream receives independent LoRA adapters and 48 learnable prefix tokens. The aggregator combines stream outputs with learnable weights, while Barlow Twins regularization at layer 16 enforces stream decorrelation to maximize neural diversity.

3.1 Architecture

We introduce ND-LoRA (Neural Diversity Low-Rank Adaptation), a parameter-efficient method that demonstrates our theoretical framework for neural diversity regularization. ND-LoRA extends the ParScale architecture by incorporating stream-aware LoRA adapters with explicit decorrelation objectives.

Our implementation builds on Qwen2.5-0.5B (896 hidden dimensions, 24 layers) with P parallel computation streams. Each stream $i \in \{1, \ldots, P\}$ processes the input through independent transformations:

$$h_i^{(\ell)} = \text{Layer}^{(\ell)}(h_i^{(\ell-1)} + B_i^{(\ell)} A_i^{(\ell)} h_i^{(\ell-1)}) + \text{Prefix}_i^{(\ell)}$$

where $B_i^{(\ell)} \in \mathbb{R}^{d \times r}$, $A_i^{(\ell)} \in \mathbb{R}^{r \times d}$ are stream-specific LoRA matrices with rank r, and $\text{Prefix}_i^{(\ell)} \in \mathbb{R}^{48 \times d}$ are learnable prefix tokens.

The final output combines streams through a learned aggregator:

$$y = \sum_{i=1}^{P} \alpha_i \cdot \text{LM\_Head}(h_i^{(L)})$$

where $\alpha_i = \text{softmax}(\text{MLP}(\bar{h}))_i$ are dynamic weights computed from the mean pooled representations $\bar{h} = \frac{1}{P}\sum_{i=1}^{P} h_i^{(L)}$. Label smoothing with $\varepsilon = 0.1$ prevents over-reliance on individual streams.

This architecture enables stream specialization while maintaining parameter efficiency. For $P = 4$ streams with rank-16 LoRA, we use approximately 57K trainable parameters per layer, comparable to a single rank-128 LoRA but with fundamentally different representational capacity through parallel specialization.

3.2 Barlow Twins Regularization

To enforce neural diversity, we apply Barlow Twins regularization at a designated layer ℓ^* (typically layer 16). Let $z_i^{(\ell^*)} \in \mathbb{R}^{B\times T\times d}$ denote the hidden representations of stream i at the design layer. We first apply batch normalization and mean-centering to obtain whitened features $\tilde{z}_i$.

The cross-correlation matrix between streams i and j is:

$$C_{ij} = \frac{1}{BT}\sum_{b,t} \tilde{z}_{i,bt}\tilde{z}_{j,bt}^{\top} \in \mathbb{R}^{d\times d}$$

Our Barlow Twins loss promotes decorrelation by penalizing off-diagonal correlations:

$$\mathcal{L}_{BT} = \frac{1}{P(P-1)}\sum_{i\neq j} \|C_{ij} - I\|_F^2$$

The standard formulation scales quadratically with P, creating $\binom{P}{2}$ optimization dependencies that inhibit convergence. To address this scalability challenge, we implement a **RandK** variant that samples stream pairs stochastically:

$$\mathcal{L}_{BT} = \mathbb{E}_{(i,j)\sim \text{MultN}(C_i)} \|C_{ij} - I\|_F^2$$

where C_i induces a multinomial distribution over stream pairs. This reduces complexity from $O(P^2)$ to $O(PK)$ where K is the number of sampled pairs, enabling efficient scaling to larger P.

The total training objective combines cross-entropy and decorrelation terms:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{BT}\mathcal{L}_{BT}$$

We set $\lambda_{BT} = 0.1$ across all experiments, providing sufficient regularization without overwhelming the primary language modeling objective.

Our implementation includes an optional normalization warmup that gradually increases λ_{BT} from 0 to its target value over the first 10% of training steps. This stabilizes optimization when the number of decorrelation constraints becomes large, as observed in our $P = 8$ experiments.

3.3 Training Details

We train ND-LoRA using parameter-efficient fine-tuning (PEFT) on The Pile dataset. Our training protocol freezes the backbone Qwen2.5-0.5B parameters and optimizes only the prefix tokens, LoRA adapters, and aggregator weights — approximately 1.3M trainable parameters total.

Data and Tokenization. We stream 8 randomly selected shards from The Pile, processing 20M tokens with 1024-token sequences. Each sequence reserves 48 tokens for prefixes, leaving 976 tokens for actual content. We use fixed seeds for reproducible shard selection and maintain exact token budgets across all P values for fair comparison.

Optimization. Training uses AdamW with peak learning rate 3e-4, minimum rate 1e-5, and cosine decay over target tokens. We apply 2% warmup, weight decay 0.1, and gradient clipping at norm 1.0. The effective batch size is 64 (micro-batch 4, gradient accumulation 16) with bfloat16 precision for memory efficiency.

Stream Configurations. We compare three architectural variants: (1) shared LoRA parameters across all streams (baseline), (2) independent rank-16 LoRA per stream (Indep. LoRA), and (3)

Model	HaluEval	MemoTrap	TruthfulQA	NQ	Wikitext	WG
ND-LoRA R16 (P=4)	**0.453**	**0.641**	**0.423**	0.060	0.784	**0.574**
ParScale R64 (P=4)	0.409	0.634	0.413	0.061	0.793	0.573
Qwen2.5-0.5B LoRA R64	0.394	0.629	0.399	**0.065**	**0.778**	0.564

Table 1: ND-LoRA P=4 improves reliability across many hallucination-sensitive benchmarks (TruthfulQA-MC2, MemoTrap, HaluEval-Summarization) without sacrificing accuracy on general tasks (Wikitext BPB, Winogrande). In fact, ND-LoRA improves not just reliability but also accuracy over ParScale in all cases and performs the best of all models on Winogrande. Please note that LoRA ranks for Qwen and ParScale are larger to ensure parameter- and data-matched comparisons.

independent LoRA with Barlow Twins and normalization warmup. The stream-aware configurations enable specialization while maintaining parameter parity through reduced per-stream rank.

Training completes in approximately 5K steps (20M tokens ÷ 16 batch size ÷ 1.2K avg. tokens/sample), taking 2-4 hours on A100-80GB hardware depending on P and regularization complexity.

4 Experimental Validation

4.1 Experimental Setup

We validate our framework through systematic experiments ensuring parameter- and data-matched comparisons.

Setup. We use Qwen2.5-0.5B with ND-LoRA across $P \in \{1, 2, 4, 8\}$ streams. Each stream uses rank-16 LoRA adapters and 48 prefix tokens (1.3M trainable parameters, 494M backbone frozen). Baselines use higher-rank LoRA (R64) for fair parameter matching.

Training. Models train on 20M tokens from The Pile with AdamW (lr 3e-4, cosine decay), batch size 64, taking 2-4 hours on A100-80GB.

Evaluation. Three categories: (1) *Hallucination-sensitive*: TruthfulQA, HaluEval, MemoTrap; (2) *Knowledge-intensive*: Natural Questions, TriviaQA, PopQA; (3) *General*: Wikitext, Winogrande.

4.2 Key Results

Table 4.2 shows ND-LoRA achieves substantial improvements on hallucination-sensitive benchmarks. ND-LoRA (P=4) achieves: 4.5% improvement on HaluEval (0.453 vs 0.409), 1.1% on MemoTrap (0.641 vs 0.634), and 1.0% on TruthfulQA (0.423 vs 0.413) vs parameter-matched ParScale, validating that neural diversity reduces hallucination probability.

ND-LoRA's improvements target reliability rather than general capability. While ParScale slightly outperforms ND-LoRA on Wikitext (0.793 vs 0.784) and Natural Questions (0.061 vs 0.060), ND-LoRA achieves best performance on Winogrande (0.574), showing diversity regularization preserves general reasoning. Despite using lower-rank adapters, ND-LoRA outperforms high-rank baselines on hallucination tasks, showing architectural diversity provides more value than simply increasing parameters.

The results also provide strong empirical validation of our theoretical bounds. Figure **??** shows that measured neural diversity indices correlate strongly with hallucination reduction across different architectural configurations, with ND-LoRA achieving the highest diversity (approaching $\mathcal{D}_{\text{spec}} = 1.0$) and correspondingly lowest hallucination rates. The empirical data points cluster around our theoretical bound, demonstrating that our margin-based analysis provides accurate predictions of real-world performance trends.

These findings establish neural diversity as a practical mechanism for improving model reliability. The consistent improvements across multiple hallucination benchmarks, combined with preserved general performance, suggest that ND-LoRA addresses fundamental reliability challenges rather than optimizing for specific evaluation metrics.

Task	Best P	Best Score	$\overline{\Delta}$% Score
HaluEval (Dialog)	2	0.480	+3.1%
HaluEval (QA)	1	0.376	–
HaluEval (Summ)	4	0.470	+17.9%
MemoTrap v2	4	0.650	+2.5%
NQ (8-shot)	8	0.066	+7.9%
PopQA	8	0.128	+1.6%
TriviaQA (8-shot)	1	0.227	–
TruthfulQA (MC1)	4	0.271	+7.8%
TruthfulQA (MC2)	4	0.431	+5.8%

Table 2: The optimal amount of neural diversity varies by task. Each row shows the position of the performance peak (best P), the best score, and the relative peak vs. baseline change.

4.3 Scaling Analysis

Table 5.1 reveals that different tasks require different optimal amounts of diversity, with no universal "best" P, validating our theoretical prediction.

This task-dependent sensitivity pattern has important practical implications. For deployment scenarios prioritizing reliability over factual recall, a slightly sub-optimal P choice may be acceptable since hallucination-sensitive tasks are more forgiving. However, knowledge-intensive applications requiring precise factual accuracy benefit from careful P selection tailored to the specific task requirements.

This validates our theoretical framework's prediction of non-monotonic scaling behavior. Rather than "more parallelism is always better," we observe U-shaped curves where performance often improves as P increases but then degrades at higher P values after a certain point.

4.4 Computational Considerations

Table 5.2 demonstrates that ND-LoRA achieves substantial diversity and performance gains with minimal computational overhead. The full ND-LoRA configuration incurs only 1.12× training cost and 1.1× inference latency compared to the single-stream baseline, while delivering 4.9% average score improvement and perfect diversity (100% D_{spec}).

The training overhead is remarkably small because we employ parameter-efficient fine-tuning that keeps the 494M backbone frozen while training only 1.3M parameters across streams, prefixes, and the aggregator. The modest increase in training FLOPs comes primarily from computing Barlow Twins regularization across stream pairs and the additional forward passes for P parallel streams. Since fine-tuning requires far fewer training steps than pre-training, this overhead translates to mere minutes of additional wall-clock time.

Compared to naive P-ensemble requiring P× parameters and cost, ND-LoRA achieves comparable diversity at 1.12× cost. Independent LoRA alone provides 2.9% improvement at 1.05× cost; adding Barlow Twins increases this to 4.9% at 1.12× cost, demonstrating synergy between architectural diversity and explicit regularization.

5 Mechanistic Analysis

5.1 Task-Dependent Optimality

Our theoretical framework predicts that different tasks should exhibit varying sensitivity to neural diversity, depending on their precision requirements and the correlation structure of their error modes. To validate this hypothesis empirically, we analyze performance patterns across tasks as a function of P, normalizing each task's performance to its optimal configuration.

Table 5.1 shows task-dependent sensitivity patterns as P deviates from optimal values. Knowledge-intensive tasks like TriviaQA and NQ exhibit steep performance degradation when P deviates from

Variant	# Streams	LoRA	Regularization	$\overline{\Delta}$% Score	Cost
Standard	1	Shared	Dropout (D)	0.0%	1.00x / 1.0x
ParScale	P	Shared	D	+0.5%	1.01x / 1.1x
ND-ParScale	P	Shared	D + RandK BT	+1.4%	1.06x / 1.1x
Indep. LoRA	P	Stream-Aware	D	+2.9%	1.05x / 1.1x
ND-LoRA	P	Stream-Aware	D + RandK BT	**+4.9%**	1.12x / 1.1x

Table 3: Ablations demonstrate the super-linear impact of stream-aware LoRAs with Barlow Twins-style regularization on a parameter- and data-matched basis. *LoRA*: shared adapter across streams vs. independent adapters per stream. *Regularization*: Dropout (0.1) at LoRA adapters and a scalable Barlow Twins cross-stream redundancy penalty. $\overline{\Delta}$% *Score*: mean relative change in score at optimal P. *Cost*: estimated training FLOPs / inference latency compared to baseline.

optimal, suggesting high precision requirements that benefit from careful neural diversity calibration. In contrast, truthfulness tasks like TruthfulQA and MemoTrap show flatter profiles, indicating greater robustness to sub-optimal P choices.

5.2 Ablations

Table 5.2 reveals a clear superlinear relationship between architectural diversity and orthogonalizing regularization in neural diversity systems. Independent LoRA adapters alone provide the primary breakthrough, jumping from ParScale's modest 0.5% improvement to 2.9% at minimal cost (1.05× vs 1.01× training overhead). However, neither component achieves its full potential in isolation—shared parameters inherently constrain diversity regardless of regularization strength, while independent parameters without explicit decorrelation may accidentally converge during optimization. The combination proves synergistic: adding Barlow Twins regularization to independent adapters pushes performance from 2.9% to 4.9% improvement, demonstrating that architectural diversity and explicit orthogonalization are complementary rather than redundant mechanisms for hallucination reduction.

6 Related Work

Hallucination in Language Models. Hallucinations represent a fundamental challenge (Huang et al., 2024; Tonmoy et al., 2024), with theoretical work proving they are mathematically inevitable (Xu et al., 2024) and particularly severe in smaller models (Lin et al., 2021; Li et al., 2023a). Current strategies include retrieval augmentation (Niu et al., 2024), specialized decoding (Li et al., 2023b), and constitutional training (Bai et al., 2022), but each addresses only specific aspects. TruthfulQA shows scaling paradoxically *decreases* truthfulness, motivating architectural solutions (Lin et al., 2021). Our work differs by modifying internal processing through neural diversity rather than post-processing.

Parallel Architectures and Scaling Laws. Parallel scaling offers a third axis beyond parameter and data scaling (Chen et al., 2025; Kaplan et al., 2020). ParScale achieves $O(\log P)$ gains with 22× less memory than parameter scaling (Chen et al., 2025). MoE architectures leverage conditional computation for 1000× capacity increases (Shazeer et al., 2017). However, existing approaches suffer from representation collapse where streams converge to similar features. Our neural diversity framework addresses this by actively maintaining decorrelation.

Diversity in Neural Networks. Ensemble diversity theory connects inter-model correlation to error rates. Deep ensembles show power-law scaling with memory split across networks outperforming single large models (Lobacheva et al., 2020). PAC-Bayesian analysis proves ensemble error decreases with diversity (Ortega et al., 2022). Recent LLM ensemble work shows explicit diversity optimization outperforms naive ensembling (Tekin et al., 2024). While these approaches require multiple separate models, our method achieves diversity benefits within a single architecture.

Redundancy Reduction and Self-Supervised Learning. Self-supervised methods maintain representational diversity. Barlow Twins prevents collapse through decorrelation (Zbontar et al., 2021),

while VICReg decomposes this into variance, invariance, and covariance terms (Bardes et al., 2022). We adapt these principles for language model hallucination reduction.

Parameter-Efficient Fine-Tuning. PEFT methods enable distinct streams under fixed budgets. LoRA reduces parameters 10,000× while maintaining performance (Hu et al., 2022; Wang et al., 2023). Prefix-tuning optimizes task-specific vectors using 0.1% of parameters (Li & Liang, 2021). These approaches provide the foundation for our neural diversity framework.

Inference-Time Scaling and Aggregation. Self-consistency improves accuracy through diverse sampling (Wang et al., 2022), while confidence-based weighting reduces required paths by 40% (Taubenfeld et al., 2025). Contrastive decoding leverages multiple views during generation (Li et al., 2023b; Sanchez et al., 2023). Unlike these post-hoc methods requiring multiple forward passes, our training-time parallelism learns coordinated streams efficiently.

Theoretical Foundations. Margin-based reliability theory provides the framework for understanding diversity's role. PAC-Bayesian bounds connect diversity to generalization (Steffen et al., 2024; Biggs & Guedj, 2022). Concentration inequalities show reducing correlation tightens tail bounds (Alquier, 2024). This supports our approach of regularizing cross-stream correlation for reliability.

7 Discussion

Our work demonstrates that neural diversity provides a principled mechanism for reducing hallucinations, but several limitations warrant discussion. First, experiments focus on 0.5B models; scaling to larger models may reveal different diversity-reliability trade-offs. Second, RandK sampling reduces complexity from $O(P^2)$ to $O(PK)$ but introduces variance requiring careful tuning. Third, our theoretical analysis assumes approximately linear margins in whitened features, which may not hold across all architectures. Finally, evaluation relies on existing benchmarks that may not capture all hallucination modes in long-form or domain-specific applications.

Despite these limitations, neural diversity opens promising research directions beyond hallucination mitigation. The principle that decorrelated representations improve reliability could extend to adversarial robustness, out-of-distribution detection, and uncertainty quantification. Task-dependent optimal diversity suggests adaptive mechanisms that dynamically adjust P based on input characteristics. More broadly, our results challenge monolithic scaling by demonstrating that *how* we scale matters as much as how much—particularly relevant as the field grapples with computational costs of ever-larger models. As language models become critical infrastructure, techniques improving reliability without massive investment become essential. Neural diversity offers one path: meaningful improvements through architectural innovation rather than brute-force scale, suggesting reliable AI's future lies in thoughtfully designed rather than simply bigger models.

References

Pierre Alquier. User-friendly introduction to pac-bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Felix Biggs and Benjamin Guedj. On margins and derandomisation in pac-bayes. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 3709–3731. PMLR, 2022.

Yutong Chen, Dawei Li, Yingyu Zhang, Xinyin Ding, Chuhan Xiao, and Ruoyu Zhang. Parscale: Parallel scaling law for language models. *arXiv preprint arXiv:2505.10475*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023a.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 4582–4597, 2021.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 12286–12312, 2023b.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20136–20148, 2021.

Ekaterina Lobacheva, Nadezhda Chirkova, Maxim Kodryan, and Dmitry Vetrov. On power laws in deep ensembles. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2375–2385, 2020.

Yuanhao Niu, Kaihua Huang, Bowen Shi, Shengyu Wang, Zihao Xu, Ke Yang, Guo Hong, Liang Li, Zhiyuan Liu, et al. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 10492–10510, 2024.

Luis A Ortega, Rafael Cabañas, and Andrés Masegosa. Diversity and generalization in neural network ensembles. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 11720–11743. PMLR, 2022.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Sonja Steffen, Benjamin Scherrer, and Barbara Hammer. Misclassification bounds for pac-bayesian sparse deep learning. *Machine Learning*, 113:4679–4727, 2024.

Yotam Taubenfeld, Hadas Kotek, and Ido Dagan. Confidence improves self-consistency in language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1030–1045, 2025.

Selim Furkan Tekin, Fatih Ilhan, Ege Karakose, and Mert Kobas. Llm-topla: Efficient llm ensemble by maximising diversity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5627–5642, 2024.

SM Tonmoy, SM Mahbub Zaman, Shafiq Joty, M Sohel Rahman, Md Tanvir Hasan, et al. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

Xi Victoria Wang, Alexander Ororbia, Karthik Kini, and Yi Lu. Lora ensembles for large language model fine-tuning. *arXiv preprint arXiv:2310.00035*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3714–3727, 2022.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

A Appendix

A.1 Full Proofs

A.1.1 Proof of Lemma 1

Proof. By bilinearity of covariance,

$$\begin{aligned}\mathrm{Var}(M) &= \mathrm{Var}\Big(\frac{1}{P}\sum_{i=1}^{P} m_i\Big) = \frac{1}{P^2}\left(\sum_{i=1}^{P}\mathrm{Var}(m_i) \;+\; 2\sum_{1\le i<j\le P}\mathrm{Cov}(m_i,m_j)\right)\\ &= \frac{1}{P^2}\Big(P\,\sigma^2 \;+\; 2\cdot\tbinom{P}{2}\,\rho\,\sigma^2\Big) = \frac{1}{P^2}\Big(P\,\sigma^2 + P(P-1)\,\rho\,\sigma^2\Big)\\ &= \sigma^2\left(\frac{1}{P} + \frac{P-1}{P}\,\rho\right) = \sigma^2\left(\frac{1-\rho}{P} + \rho\right).\end{aligned}$$

□

A.1.2 Proof of Lemma 2

Proof. Since $m_i = v_i^\top \tilde{z}_i$ and $m_j = v_j^\top \tilde{z}_j$,

$$\mathrm{Cov}(m_i, m_j) = \mathbb{E}\big[(v_i^\top \tilde{z}_i)(v_j^\top \tilde{z}_j)\big] = v_i^\top \mathbb{E}[\tilde{z}_i \tilde{z}_j^\top] v_j = v_i^\top C_{ij} v_j.$$

By the definition of the spectral norm,

$$|\mathrm{Cov}(m_i, m_j)| = |v_i^\top C_{ij} v_j| \le \|v_i\|\;\|C_{ij}\|_2\;\|v_j\|\,.$$

Divide both sides by $\sigma_i\sigma_j$ to obtain

$$|\rho_{ij}| \le \frac{\|v_i\|\;\|C_{ij}\|_2\;\|v_j\|}{\sigma_i\sigma_j} = \kappa_{ij}\;\|C_{ij}\|_2\,.$$

Since the right-hand side is nonnegative, this yields the stated upper bound on ρ_{ij}. □

A.1.3 Proof of Corollary **??**

Proof. From Lemma 2, for each $i \neq j$ we have $\rho_{ij} \le \kappa_{ij}\,\|C_{ij}\|_2$. Averaging over the $\binom{P}{2}$ unordered pairs gives

$$\bar{\rho} \;=\; \frac{2}{P(P-1)}\sum_{i<j}\rho_{ij} \;\le\; \frac{2}{P(P-1)}\sum_{i<j}\kappa_{ij}\;\|C_{ij}\|_2\,.$$

Insert and extract the average of κ_{ij}:

$$\bar{\rho} \;\le\; \bar{\kappa}\cdot\frac{2}{P(P-1)}\sum_{i<j}\|C_{ij}\|_2 = \bar{\kappa}\cdot\frac{1}{P(P-1)}\sum_{i\neq j}\|C_{ij}\|_2 = \bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}.$$

□

A.1.4 PROOF OF THEOREM 1

Proof. Let $\bar{\rho}$ denote the average pairwise correlation entering Lemma 1. For $P > 1$,

$$\frac{\partial}{\partial \rho}\Big[\sigma^2\Big(\frac{1-\rho}{P}+\rho\Big)\Big] = \sigma^2\Big(1-\frac{1}{P}\Big) \;\geq\; 0,$$

so $\mathrm{Var}(M)$ is (weakly) increasing in ρ. By Corollary **??**, under the linear/whitened feature model, $\bar{\rho} \leq \bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}$; therefore

$$\mathrm{Var}(M) \;\leq\; \sigma^2\left(\frac{1-\bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}}{P}+\bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}\right).$$

Apply Proposition **??** with this variance bound to obtain the stated fraction.

Finally, if the conditions needed to assert $\bar{\rho} \leq \bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}$ hold only on a "good" event G with $\mathbb{P}(G) \geq 1 - h_0$, then

$$\mathbb{P}(\mathrm{H}) = \mathbb{P}(\mathrm{H} \mid G)\mathbb{P}(G) + \mathbb{P}(\mathrm{H} \mid G^{\mathrm{c}})\mathbb{P}(G^{\mathrm{c}}) \leq \mathbb{P}(\mathrm{H} \mid G) + h_0,$$

and the same bound applies to $\mathbb{P}(\mathrm{H} \mid G)$ by the previous steps. This yields the formula with the additive slack h_0. □

A.1.5 PROOF OF THEOREM 2

Proof. Let $g(P) \triangleq \frac{1-\rho(P)}{P} + \rho(P)$ so that $\mathcal{B}(P)$ is an increasing function of $g(P)$ (monotonicity in the numerator). Compute

$$g'(P) = \frac{d}{dP}\left(\frac{1-\rho(P)}{P}\right)+\rho'(P) = \Big(-\frac{\rho'(P)}{P}-\frac{1-\rho(P)}{P^2}\Big)+\rho'(P) = \rho'(P)\Big(1-\frac{1}{P}\Big)-\frac{1-\rho(P)}{P^2}.$$

As $P \downarrow 1$, we have $1 - \frac{1}{P} \to 0$ while $1 - \rho(P) \to 1 - \rho_0 > 0$, hence $g'(P) \to -(1-\rho_0) < 0$. For large P, the negative term is $O(P^{-2})$, while $\rho'(P) = \beta\gamma(P-1)^{\gamma-1} > 0$; thus for sufficiently large P, $g'(P) > 0$. By continuity, there exists $P_\star > 1$ with $g'(P_\star) = 0$, implying that g (and therefore $\mathcal{B}$) decreases for $P < P_\star$ and increases for $P > P_\star$.

If $\gamma \geq 1$, then $\rho''(P) = \beta\gamma(\gamma-1)(P-1)^{\gamma-2} \geq 0$ and

$$g''(P) = \rho''(P)\Big(1-\frac{1}{P}\Big) + \frac{2\rho'(P)}{P^2} + \frac{2(1-\rho(P))}{P^3} \;\geq\; 0 \quad \text{for } P > 1.$$

Hence g is convex on $(1, \infty)$ and has a unique minimizer; the same holds for $\mathcal{B}$, which is an increasing transform of g. □