NEURAL DIVERSITY REGULARIZES HALLUCINATIONS IN SMALL LANGUAGE MODELS

Kushal Chakrabarti
South Park Commons
San Francisco, CA 94107, USA
kushalc@obviouslywrong.org

Nirmal Balachundhar
South Park Commons
San Francisco, CA 94107, USA
nbalachundhar@gmail.com

ABSTRACT

Language models continue to hallucinate despite increases in parameters, compute, and data. We propose *neural diversity* — decorrelated parallel representations — as a principled mechanism that reduces hallucination rates at fixed parameter and data budgets. Inspired by portfolio theory, where uncorrelated assets reduce risk by $\sqrt{P}$, we prove hallucination probability is bounded by representational correlation: $\mathbb{P}(\mathrm{H}) \leq f(\sigma^2((1-\rho(P))/P + \rho(P)), \mu^2)$, which predicts that language models need an optimal amount of neurodiversity. To validate this, we introduce ND-LoRA (Neural Diversity Low-Rank Adaptation), combining parallel LoRA adapters with Barlow Twins regularization, and demonstrate that ND-LoRA *reduces hallucinations by up to 25.6% (14.6% avg.)* without degrading general accuracy. Ablations show LoRA adapters and regularization act synergistically, causal interventions prove neurodiversity as the mediating factor and correlational analyses indicate scale: a 0.1% neural correlation increase is associated with a 3.8% hallucination increase. Finally, task-dependent optimality emerges: different tasks require different amounts of optimal neurodiversity. Together, our results highlight neural diversity as a third axis of scaling — orthogonal to parameters and data — to *improve the reliability of language models at fixed budgets*.

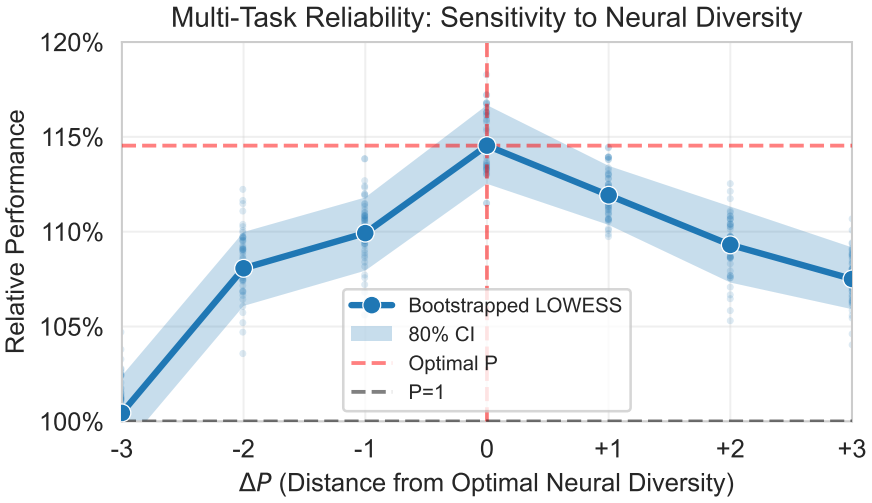

Figure 1: **Maximizing reliability requires an optimal amount of neural diversity.** We vary parallel decorrelated representations ($P \in \{1, 2, 4, 8\}$) and measure performance across 6 hallucination benchmarks. The x-axis shows distance from the performance-maximizing $P_\star$: $\Delta P = P - P_\star$. Aggregating 182,850 evaluation points (LOWESS fit, 80% bootstrap CI, 1000 bootstrap resamplings), the U-shaped curve shows performance improves rapidly with neural diversity, peaking at a specific optimum and then degrading slowly with excessive parallelism.

1 INTRODUCTION

Despite scaling to trillions of parameters, language models hallucinate persistently (Lin et al., 2021). This reliability crisis is acute for small language models (SLMs) — increasingly favored for edge deployment (Chen et al., 2024) — whose compressed representations make them especially vulnerable to hallucinations, with even well-resourced efforts like GPT-OSS 20B exhibiting 91% hallucination rates on factual benchmarks (OpenAI, 2025).

Existing scaling methods — parameters, inference-time compute, and parallel architectures like ParScale (Chen et al., 2025) — predominantly target *accuracy* (first-moment improvements in perplexity and task performance). Yet *reliability* (second-moment reduction in hallucinations and factual errors) remains largely unaddressed. While ParScale achieves $O(\log P)$ performance gains with better memory and latency, it provides no explicit mechanism for reducing hallucinations. Without explicit regularization, *representational collapse* (Jing et al., 2022) drives parallel streams toward identical features, leaving reliability gains unrealized despite added computational resources.

Drawing from portfolio theory, we propose that decorrelated representations reduce hallucination risk just as uncorrelated assets reduce financial risk. When streams are perfectly correlated ($\rho = 1$), parallelism provides no benefit; when decorrelated ($\rho \to 0$), signal-to-noise improves by $\sqrt{P}$.

We formalize this through a **neural diversity framework** proving 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 existing scaling approaches don't address reliability and how explicit decorrelation provides a mechanism to reduce hallucinations at fixed parameter budgets.

To validate this, we introduce **ND-LoRA**, combining independent LoRA adapters with Barlow Twins regularization (Zbontar et al., 2021). Despite adding only 0.004% more computation, ND-LoRA reduces hallucinations by up to 25.6% (14.6% avg.) while preserving general performance.

Our experiments establish: (1) neural diversity causally reduces hallucination probability; (2) there is an optimal amount of neural diversity and that optimal amount is task-dependent; and (3) the relationship between neural diversity and hallucination probability is theoretically provable.

Neural diversity represents a third scaling axis beyond parameters and data. While traditional scaling asks "how big?" and data scaling "how much?", diversity scaling asks "how different?" — crucial for achieving reliability without massive computational investment.

Our contributions are:

- **Theory**: The first principled connection between neural spectral diversity and hallucination probability that matches empirically observed hallucination rates
- **Method**: ND-LoRA, a parameter-efficient demonstration achieving up to 25.6% hallucination reduction and maintaining general-purpose accuracy at 1.00004× training cost
- **Analysis**: Evidence of the concavity of P in industry-standard benchmarks and the presence of task-dependent optimal P revealing when neural diversity helps most (and hurts)
- **Framework**: A general principle for parallel architectures — that maintaining decorrelation is as important as adding capacity — with implications beyond hallucination mitigation

2 A THEORY OF NEURAL DIVERSITY

Why don't existing scaling methods improve reliability? Without explicit diversity mechanisms, gradient descent drives parallel streams toward similar representations through *representational collapse* (Jing et al., 2022), leaving reliability gains unrealized. We establish the first rigorous connection between architectural diversity and hallucination probability, proving that decorrelated streams reduce hallucinations and providing mathematical foundations for ND-LoRA.

2.1 PRELIMINARIES

We develop a framework connecting parallel neural computation to hallucination probability. Consider a language model processing input x to produce an output (e.g., answer to a question, summary

of a document). We say the model *hallucinates* when it generates content unsupported by its input or training knowledge — for instance, fabricating facts, misattributing quotes, or creating non-existent references.

Our architecture employs P parallel computational pathways called *streams*, each processing the same input through the same model but in perturbed ways. Each stream i produces a scalar *margin* $m_i \equiv m_i(x)$ representing its confidence: positive margins indicate the stream believes the output is correct, while negative margins signal likely errors. These streams are combined through a weighted average to produce the final prediction:

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

We model each stream with mean confidence $\mathbb{E}[m_i] = \mu$ (with $\mu > 0$ for correct outputs), per-stream variance $\mathrm{Var}(m_i) = \sigma^2$, and pairwise correlations $\rho_{ij} = \mathrm{Corr}(m_i, m_j)$ measuring how similarly streams behave. The *average pairwise correlation* captures overall stream similarity:

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

We define the hallucination event as $\mathrm{H} \equiv \{M \le 0\}$—when the aggregated margin crosses zero, indicating the ensemble prediction is incorrect.

The key intuition is diversification: when streams are perfectly correlated ($\rho = 1$), they make identical errors and averaging provides no benefit. When decorrelated ($\rho \to 0$), independent errors cancel out and M concentrates around its mean μ, reducing hallucination probability as P increases.

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*

$$\mathcal{D}_{\mathrm{spec}} = \frac{1}{P(P-1)} \sum_{i \ne j} \|C_{ij}\|_2 . \tag{3}$$

Lower $\mathcal{D}_{\mathrm{spec}}$ indicates greater diversity: $\mathcal{D}_{\mathrm{spec}} = 0$ means perfectly decorrelated streams, while $\mathcal{D}_{\mathrm{spec}} = 1$ 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,*

$$\mathrm{Var}(M) = \sigma^2 \left(\frac{1-\rho}{P} + \rho \right). \tag{4}$$

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 $i \ne j$, 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 $i \ne j$,*

$$\rho_{ij} = \frac{\mathrm{Cov}(m_i, m_j)}{\sigma_i \sigma_j} \le \kappa_{ij} \, \|C_{ij}\|_2 , \qquad \kappa_{ij} \triangleq \frac{\|v_i\| \, \|v_j\|}{\sigma_i \sigma_j}. \tag{5}$$

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 $\mathcal{D}_{\mathrm{spec}}$ to hallucination probability.

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

$$\mathbb{P}(\mathrm{H}) \;\leq\; \frac{\sigma^2 \left(\frac{1-\bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}}{P} + \bar{\kappa}\,\mathcal{D}_{\mathrm{spec}} \right)}{\sigma^2 \left(\frac{1-\bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}}{P} + \bar{\kappa}\,\mathcal{D}_{\mathrm{spec}} \right) + \mu^2} \;+\; h_0, \tag{6}$$

where $\bar{\kappa}$ is the average correlation scaling factor and $0 \leq h_0 \leq 1$ is a constant accounting for approximation error.

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

This gets us the first half of our theoretical result: Lower $\mathcal{D}_{\mathrm{spec}}$ reduces hallucination probability and, additionally, the benefit scales with P.

2.3 Scaling Behavior

Next, we demonstrate that under common circumstances, the hallucination bound follows a U-shaped curve — initially decreasing with higher P, but starts increasing eventually. Consider the case where the correlation itself increases with P, say, due to optimizer constraints:

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:*

$$\mathcal{B}(P) \;\triangleq\; \frac{\sigma^2 \left(\frac{1-\rho(P)}{P} + \rho(P) \right)}{\sigma^2 \left(\frac{1-\rho(P)}{P} + \rho(P) \right) + \mu^2}. \tag{7}$$

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 $\mathcal{B}(P)$.

All together, Theorems 1 and 2 show that there exists an optimal $P_\star$ that minimizes hallucinations. Next, we show how to construct an architecture and training protocol to find $P_\star$.

3 ND-LoRA: A Practical Demonstration

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 with stream-aware LoRA adapters and explicit decorrelation objectives.

3.1 Architecture

Our implementation builds on ParScale with P parallel computation streams. Each stream $i \in \{1, \ldots, P\}$ uses 48 learnable prefix tokens prepended to the input sequence that flow through all layers via the attention mechanism, along with stream-specific LoRA adapters applied at each layer:

$$h_i^{(\ell)} = \mathrm{Layer}^{(\ell)}(h_i^{(\ell-1)} + B_i^{(\ell)} A_i^{(\ell)} h_i^{(\ell-1)}) \tag{8}$$

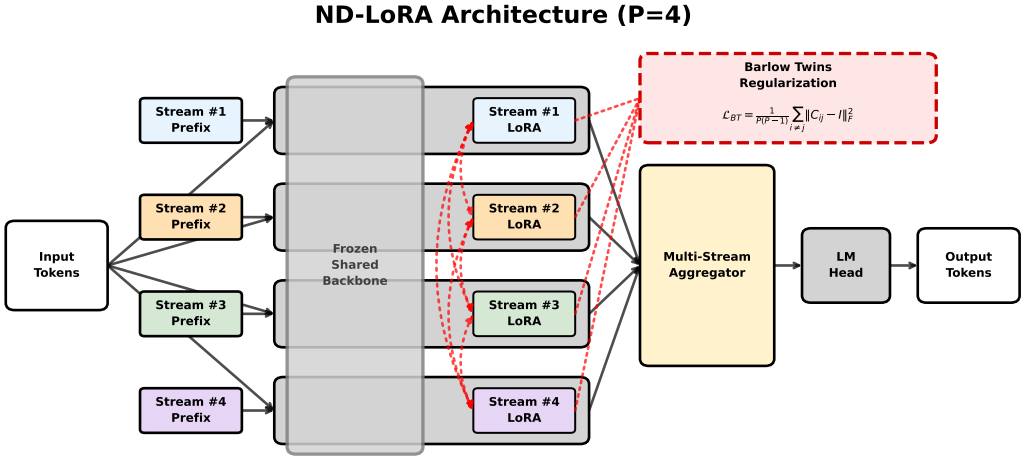

Figure 2: **ND-LoRA schematic for $P = 4$ parallel streams.** Each stream receives independent LoRA adapters and learnable prefix tokens. The aggregator combines stream outputs with learnable weights, while Barlow Twins regularization incentivizes decorrelation between stream outputs.

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. The final output combines streams through a learned aggregator:

$$y = \text{LM\_Head}\left(\sum_{i=1}^{P} \alpha_i \cdot h_i^{(L)}\right) \tag{9}$$

where $\alpha_i = (1-\varepsilon)\cdot \text{softmax}(\text{MLP}([h_1^{(L)},\ldots,h_P^{(L)}]))_i + \varepsilon/P$ are dynamic weights with label smoothing ($\varepsilon = 0.1$) computed from the concatenated stream representations. This prevents attention collapse by ensuring minimum weight ε/P for each stream.

This architecture enables stream specialization while maintaining parameter efficiency. For $P = 2$ streams with rank-16 LoRA, we use approximately 29K trainable parameters per layer, comparable to a single rank-32 LoRA but with fundamentally different representational capabilities.

3.2 Barlow Twins Regularization

To encourage neural diversity, we apply Barlow Twins regularization at a pre-specified design layer ℓ^*. 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:

$$C_{ij} = \frac{1}{BT}\sum_{b,t} \tilde{z}_{i,bt}\tilde{z}_{j,bt}^{\top} \in \mathbb{R}^{d\times d} \tag{10}$$

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

$$\mathcal{L}_{BT} = \frac{1}{P(P-1)}\sum_{i\neq j} \|C_{ij} - I\|_F^2 \tag{11}$$

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:

$$\mathcal{L}_{BT} = \mathbb{E}_{(i,j)\sim \text{MultN}(C_i)}\|C_{ij} - I\|_F^2 \tag{12}$$

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:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{BT}\mathcal{L}_{BT} \tag{13}$$

Model	HaluEval	MemoTrap	TruthfulQA	NQ	Wikitext	WG
ND-LoRA R16 (P=2)	**0.481***	**0.666***	**0.442***	0.055	0.784	**0.574**
ParScale R32 (P=2)	0.439	0.638	0.412	0.059	0.793	0.564
Qwen LoRA R32	0.400	0.634	0.403	**0.065**	**0.778**	0.572

Table 1: **Even at $P = 2$ streams, ND-LoRA achieves up to 20.2% relative hallucination reduction vs. parameter-matched baseline.** Across hallucination benchmarks, ND-LoRA shows statistically significant improvements (HaluEval-Summarization +20.2%*, MemoTrap +5.1%*, TruthfulQA-MC2 +9.5%*) while maintaining competitive Winogrande, NQ, and Wikitext general-purpose capabilities. Baselines use higher LoRA ranks for parameter parity. * indicates $p < 0.05$.

4 Experimental Validation

4.1 Experimental Setup

We validate our theoretical framework through systematic hallucination reduction experiments using neural diversity regularization with parameter- and data-matched comparisons.

Model and Architecture. We use Qwen2.5-0.5B (896 hidden dimensions, 24 layers) with ND-LoRA across $P \in \{1, 2, 4, 8\}$ parallel streams applied to QKV self-attention modules and a design layer of 20 for de-correlation loss. Each stream uses independent rank-16 LoRA adapters and 48 prefix tokens, totaling 5-20M trainable parameters with 494M backbone frozen. Baseline methods use higher-rank LoRA (R32-R128) for parameter matching.

Training Protocol. Models train on 20M tokens from The Pile (8 random shards, fixed seeds). We use 1024-token sequences, AdamW optimization (peak lr 3e-4, cosine decay, 2% warmup), batch size 64, bfloat16 precision, design layer 16. Training completes in 5K steps (30 minutes on A100).

Evaluation Benchmarks. We evaluate across: (1) *Hallucination-sensitive*: TruthfulQA (Lin et al., 2021), HaluEval (Li et al., 2023a), MemoTrap (McKenzie et al., 2023); (2) *Knowledge-intensive*: Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023); (3) *General capability*: Wikitext BPB (Merity et al., 2017), Winogrande (Sakaguchi et al., 2020). This tests if neural diversity improves reliability without sacrificing general performance.

Statistical Methodology. We evaluate significance using McNemar's test for binary classification tasks and two-tailed bootstrap tests with 10,000 samples for other tasks. Improvements marked with * are significant at $p < 0.05$.

4.2 Key Results

Table 1 demonstrates ND-LoRA achieves substantial improvements on hallucination-sensitive benchmarks while maintaining competitive general performance. ND-LoRA with $P = 2$ streams achieves statistically significant improvements on HaluEval-Summarization (0.481* vs 0.400, $p < 0.001$, 8.1% absolute / 20.2% relative), TruthfulQA-MC2 (0.442* vs 0.403, $p = 0.030$, 3.9% absolute / 9.5% relative) and MemoTrap (0.666* vs 0.634, $p < 0.001$, 3.2% absolute / 5.1% relative) vs parameter-matched Qwen, validating our theoretical prediction.

Although ND-LoRA's improvements specifically target reliability benchmarks, they preserve general capabilities. Qwen slightly outperforms on Wikitext (0.784 vs 0.778) and Natural Questions (0.055 vs. 0.065), but ND-LoRA wins on Winogrande (0.574). This task-selective pattern aligns with our framework for non-monotone behaviors and task-dependent optimality.

Parameter efficiency is evident comparing ND-LoRA R16 ($P = 2$) against Qwen2.5-0.5B LoRA R32. Despite lower-rank adapters, ND-LoRA consistently outperforms the high-rank baseline on hallucination tasks, demonstrating that architectural diversity provides more value at equal capacity. This shows representational diversity, not parameter count, drives reliability gains in our experiment.

These findings establish neural diversity as a practical reliability mechanism. Consistent improvements across hallucination benchmarks with preserved general performance suggest ND-LoRA addresses fundamental reliability challenges rather than metric-specific optimization. Figure 3 demon-

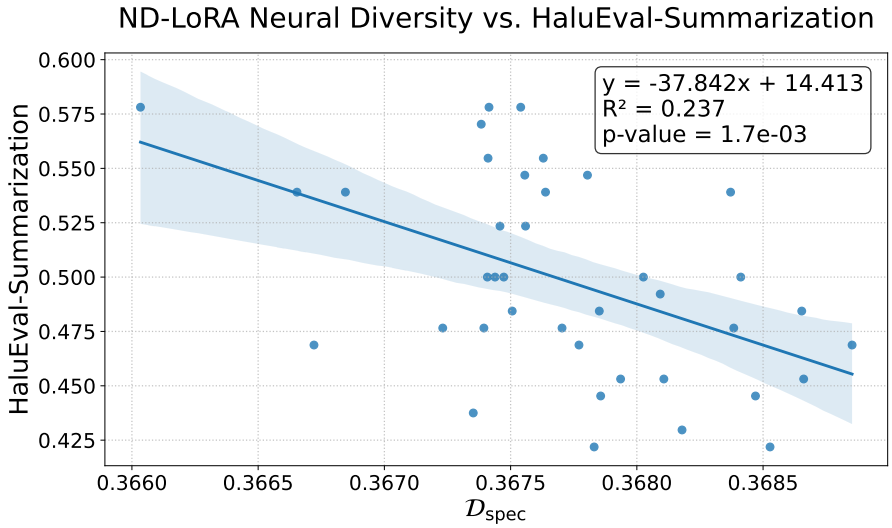

Figure 3: **Reliability improves as neural diversity increases (lower $\mathcal{D}_{\text{spec}}$).** Specifically, spectral diversity ($\mathcal{D}_{\text{spec}}$) is negatively correlated with HaluEval-Summarization performance (slope=-37.842, R²=0.237, p=0.002), consistent with $\mathbb{P}(\text{H}) \propto \mathcal{D}_{\text{spec}}$ in Theorem 1.

Category	Task	Best P	Best Score	Δ% Score	Sig.
Hallucination	HaluEval (Dialog)	4	0.516	+12.8%	***
	HaluEval (QA)	4	0.451	+23.4%	***
	HaluEval (Summ)	4	0.502	+25.6%	***
	MemoTrap v2	8	0.689	+8.8%	***
	TruthfulQA (MC1)	2	0.269	+7.3%	
	TruthfulQA (MC2)	2	0.442	+9.5%	*
Knowledge	NQ (8-shot)	1	0.066	–	
	NQ-swap	8	0.554	+0.8%	
	PopQA	1	0.111	–	
	TriviaQA (8-shot)	1	0.192	–	

Table 2: **The optimal amount of neural diversity varies by task.** Hallucination tasks show large gains with optimal $P \in \{2, 4, 8\}$ (up to +25.6%), while knowledge tasks mostly peak at $P = 1$. Even among hallucination tasks, optimal P differs: HaluEval variants peak at $P = 4$, TruthfulQA at $P = 2$, MemoTrap at $P = 8$. Each row shows the performance peak (best P), best score, relative increase vs. $P = 1$ baseline, and statistical significance (*** for $p < 0.001$, * for $p < 0.05$).

strates strong empirical correlation between neural diversity and performance, building intuition for the causal relationship established in Section 5.

4.3 Scaling Analysis

Our theoretical framework predicts U-shaped scaling curves where hallucination probability initially decreases with P but increases as correlation rises without diversity regularization. Figure 1 empirically validates this across tasks, revealing a critical insight: maximizing reliability requires an *optimal amount* of neural diversity — more diversity is not always better.

4.4 Task-Dependent Optimality

Further, the optimal diversity is task-dependent. Table 2 reveals striking task-dependent sensitivity patterns relative to the $P = 1$ baseline. Hallucination-focused tasks show the largest gains: HaluEval Summarization achieves +25.6% relative improvement at $P = 4$, HaluEval QA shows +23.4%

Task	$\Delta\mathcal{D}_{\text{spec}}$	Δ Score	SE	d	p-value	Sig.	N
HaluEval-Summ	0.024	-0.005	0.010	0.007	1.6×10^{-5}	***	512
MemoTrap v2	0.031	-0.003	0.010	0.000	8.2×10^{-5}	***	512
TruthfulQA-MC2	0.025	-0.007	0.009	0.018	3.3×10^{-7}	***	512

Table 3: **Artificial corruption of neural diversity establishes statistical causality.** Reducing neural diversity ($\Delta\mathcal{D}_{\text{spec}} > 0$) causes accuracy drops across tasks with high statistical significance ($p < 0.001$) via paired t-tests with Fisher meta-analysis (N=4 sub-experiments × 128 samples each).

at $P = 4$, and TruthfulQA MC2 shows +9.5% at $P = 2$ while MemoTrap benefits from higher diversity ($P = 8$, +8.8%). Notably, knowledge-intensive tasks like PopQA, TriviaQA and NQ show no improvement over baseline, which is expected as ND-LoRA does not add new sources of knowledge or try to improve recall of existing knowledge. This heterogeneity demonstrates that different tasks require different amounts of neural diversity to maximize reliability, with hallucination-focused tasks generally benefiting most from decorrelated representations.

5 Mechanistic Analysis

5.1 Neural Diversity as the Causal Mediator

To establish causality beyond correlation, we perform artificial corruption interventions that directly manipulate cross-stream similarity.

Experiment Design. Starting with a pre-trained ND-LoRA $P = 4$ model, we inject a corruption hook at the RMSNorm layer that randomly substitutes the hidden state at randomly-chosen positions in a given stream from another stream , reducing $\mathcal{D}_{\text{spec}}$ while preserving activation magnitudes. We evaluate on a matched basis: each corrupted evaluation is paired with an uncorrupted baseline using identical samples and resampling indices. Across 4 sub-experiments with different random seeds, we collect N=128 paired samples per task. This paired design maximizes statistical power by controlling sample-level variance, analyzed via paired t-tests with Fisher meta-analysis.

Results. Table 3 provides statistically robust evidence that neural diversity causally affects performance. All three tasks show highly significant accuracy drops ($p < 0.001$) when stream-level substitution reduces diversity ($\Delta\mathcal{D}_{\text{spec}} \approx 0.025$). While effect sizes are modest (0.3% to 0.7% score reduction) — likely because artificial stream substitution creates out-of-distribution corruption patterns — the statistical significance establishes causality beyond correlational association.

5.2 Ablations

To isolate the contributions of ND-LoRA, we systematically ablate ND-LoRA components at fixed $P = 4$ streams. All variants maintain parameter parity through LoRA rank adjustments, enabling fair comparison. We measure run-time spectral diversity ($\mathcal{D}_{\text{spec}}$) at the aggregation layer using evaluation samples, quantifying actual cross-stream correlation during inference.

Table 4 reveals a super-linear combination: independent LoRA (+2.9%) and Barlow Twins (+1.4%) sum to 4.3% but achieve 4.9% when combined (Stream LoRA-BT) — a 14% bonus. Targeting KVQ attention amplifies this further by 2.6× to +12.8% (ND-LoRA at fixed $P = 4$; maximum gains reach 14.6% when optimizing P per-task, see Table 2). Neither component alone suffices: ParScale's near-complete collapse ($\mathcal{D}_{\text{spec}} = 0.999$) yields only +0.5%, while Stream LoRA without regularization achieves +2.9%, both less than a quarter of ND-LoRA's final impact. This establishes that both architectural capacity and explicit regularization are necessary for full impact.

Counterintuitively, ND-LoRA achieves best performance (+12.8%) with *higher* $\mathcal{D}_{\text{spec}} = 0.376$ than Stream LoRA-BT's 0.140 (lowest across all variants). This reveals that strategic localization to representational bottlenecks matters more than maximizing global decorrelation: focusing LoRA and Barlow Twins on KVQ attention modules — where information flow is controlled — provides 2.6× amplification over applying them uniformly across all layers. Consistent with Table 2, this demonstrates neural diversity is a task-dependent resource requiring strategic allocation to critical computational pathways rather than uniform application.

Variant	Streams	LoRA	Regul.	Target	$\mathcal{D}_{\text{spec}}$	$\overline{\Delta}$% Score	Δ Cost
Standard	1	Single	D	All	–	0.0%	**1.0x / 1.0x**
ParScale	P	Single	D	All	0.9991	+0.5%	1.00003x / 1.1x
ParScale-BT	P	Single	D + BT	All	0.9988	+1.4%	1.00003x / 1.1x
Stream LoRA	P	Stream	D	All	0.2678	+2.9%	1.00003x / 1.1x
Stream LoRA-BT	P	Stream	D + BT	All	**0.1400**	+4.9%	1.00004x / 1.1x
ND-LoRA	P	Stream	D + BT	KVQ	0.3755	**+12.8%**	1.00004x / 1.1x

Table 4: **Ablations reveal super-linear combination of impact.** Stream LoRA (+2.9%) and Barlow Twins (+1.4%) combine super-linearly (+4.9%), and focusing on KVQ attention amplifies to +12.8%. *LoRA*: single shared vs. P stream-aware adapters. *Regularization*: Dropout vs. Barlow Twins. *Target*: All layers vs. KVQ attention only. $\mathcal{D}_{spec}$: spectral diversity (lower better). $\overline{\Delta}$% *Score*: avg. change (hallucination benchmarks).

5.3 Computational Considerations

ND-LoRA achieves substantial reliability gains with negligible overhead (1.00004$\times$ training, 1.1$\times$ inference). Three factors drive efficiency: (a) 20M token training amortizes to 0.004% of 1T pretraining; (b) frozen 494M backbone makes gradients through 1.3M parameters nearly free; (c) ND-LoRA and ParScale require identical FLOPs during inference, simply loading different LoRA adapters per stream. Unlike P-model ensembles with $P\times$ training cost, ND-LoRA achieves neural diversity within a single architecture. See Appendix A.2 for details.

6 Related Work

Hallucination in Language Models. Hallucinations represent a fundamental challenge in language model deployment (Huang et al., 2024; Tonmoy et al., 2024), proven mathematically inevitable in computable models (Xu et al., 2024; Kalai & Vempala, 2024), with particular severity in smaller models (Lin et al., 2021; Li et al., 2023a). Mechanistic investigations reveal hallucinations arise from internal representation failures (Yu et al., 2024; Ferrando et al., 2025). Current mitigation strategies—retrieval augmentation (Niu et al., 2024), specialized decoding (Li et al., 2023b; Manakul et al., 2023; Wei et al., 2024), and constitutional training (Bai et al., 2022)—address only specific aspects. MoE-LLaVA improves by 1.1% on POPE but requires $8\times$ parameters (Zhang et al., 2024). Our work achieves hallucination reduction through neural diversity within a single architecture at fixed parameter budgets.

Parallel Architectures and Scaling Laws. Parallel scaling offers a third axis beyond parameters and data (Chen et al., 2025; Kaplan et al., 2020). ParScale achieves $O(\log P)$ gains with 22× less memory and 6× less latency (Chen et al., 2025), while inference-optimal scaling laws favor smaller models with more data (Sardana & Frankle, 2024) and MoE achieves 1000× capacity through conditional computation (Shazeer et al., 2017). However, these approaches target accuracy rather than reliability. We show parallel scaling reduces hallucinations through explicit neural diversity.

Diversity in Neural Networks. Deep ensembles provide improved accuracy and uncertainty estimates (Lakshminarayanan et al., 2017), with power-law scaling showing 2% gains (Lobacheva et al., 2020) and PAC-Bayesian analysis proving error decreases with $\sqrt{\text{diversity}}$ (Ortega et al., 2022). LLM ensembles achieve 2.2% absolute MMLU improvements (Tekin et al., 2024), while negative correlation learning demonstrates diversity must be actively encouraged (Liu & Yao, 1999). However, prior work targets general accuracy and requires multiple models. Our method achieves substantially larger gains (8.1% absolute / 25.6% relative on hallucination benchmarks) at lower cost (1.00004$\times$ vs. $P\times$ training) within a single architecture.

Redundancy Reduction and Self-Supervised Learning. Self-supervised methods provide differentiable diversity mechanisms. Barlow Twins prevents collapse through dimension decorrelation (Zbontar et al., 2021), VICReg decomposes into variance, invariance, and covariance terms (Bardes et al., 2022), and understanding dimensional collapse (Jing et al., 2022) reveals why explicit decorrelation is necessary. We adapt these principles for hallucination reduction beyond visual learning.

Parameter-Efficient Fine-Tuning. PEFT methods enable distinct streams under fixed budgets. LoRA reduces parameters by 10,000× with multiple adapters inducing specialization (Hu et al., 2022; Wang et al., 2023), prefix-tuning uses 0.1% of parameters (Li & Liang, 2021), and BitFit shows minimal differences enable specialization when regularized (Ben Zaken et al., 2022). BatchEnsemble and LoRA-Ensemble achieve diversity through clever parameterization (Wen et al., 2020; Mühlematter et al., 2024), providing our technical foundation.

Inference-Time Scaling and Aggregation. Self-consistency improves reasoning by 17.9% through diverse sampling (Wang et al., 2022), though resampling has limitations with imperfect verifiers (Stroebl et al., 2024). Confidence-based weighting reduces required paths by 40% (Taubenfeld et al., 2025), while contrastive decoding and classifier-free guidance achieve improvements equivalent to parameter doubling (Li et al., 2023b; Sanchez et al., 2023). Unlike post-hoc aggregation requiring multiple forward passes, our training-time parallelism learns coordinated streams.

Theoretical Foundations. PAC-Bayesian bounds show minimax-optimal rates connecting diversity to generalization (Steffen et al., 2024; Biggs & Guedj, 2022; Ortega et al., 2022), while concentration inequalities demonstrate that reducing predictor correlation tightens tail bounds on error events (Alquier, 2024). This supports our approach of explicitly regularizing cross-stream correlation for factual reliability.

7 Discussion

Neural diversity establishes a principled framework bridging portfolio theory and deep learning to develop models that are architecturally robust to noise. Our work provides both theoretical grounding — proving hallucination probability is bounded by cross-stream correlation — and empirical validation through ND-LoRA's 25.6% hallucination reduction at negligible computational cost. The framework reveals why existing scaling approaches target accuracy over reliability and how explicit decorrelation through neural diversity provides a mechanism to reduce hallucinations at fixed parameter budgets, introducing a third scaling axis orthogonal to parameters and data. Unlike traditional ensemble methods (Lakshminarayanan et al., 2017), which require multiple separate models, our approach achieves diversity benefits within a single architecture.

At the same time, our demonstration has limitations. Computational constraints restricted experiments to 0.5B models; larger-scale production-ready implementations are part of planned future work. Benchmark scope focuses on English hallucination and knowledge tasks, with some tasks showing no benefit from increased neural diversity, highlighting task-dependent efficacy that requires further investigation. Finally, while our theoretical framework predicts non-monotonicity, it doesn't provide a satisfying explanation of why different tasks have different optimality points.

As language models become critical infrastructure, reliability techniques that avoid massive resource investment become essential. While hallucinations may be mathematically inevitable (Kalai & Vempala, 2024), neural diversity demonstrates they can be systematically reduced through architectural design. Our results show reliable AI may emerge from thoughtfully designed architectures rather than brute-force scaling alone, opening new avenues for efficient, trustworthy language models.

Author Contributions

Kushal Chakrabarti led the overall project design, core implementation of ND-LoRA, experimental design, and paper writing. Nirmal Balachundhar contributed theoretical insights and implementation for novel neural diversity techniques, experiment implementation and analysis, and paper writing.

Acknowledgments

We thank the research community at South Park Commons for valuable discussions and feedback throughout this project. We are grateful to Augustine Mavor-Parker, Bhav Ashok, David Sontag, Iman Modaressi, Jaclyn Lunger, John Bohannon, Nelson Ray and Patrick Bozeman for their insightful comments on earlier drafts. This work was supported by South Park Commons and Modal, who generously provided compute resources for our experiments.

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.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 1–9, 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.

Yu Chen, Yufan Wang, Shengming Zhang, et al. A review on edge large language models: Design, execution, and applications. *arXiv preprint arXiv:2410.11845*, 2024.

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.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. *arXiv preprint arXiv:2411.14257*, 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.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2022.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*, 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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

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.

Yong Liu and Xin Yao. Ensemble learning via negative correlation. *Neural Networks*, 12(10): 1399–1404, 1999.

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.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, Alisa Liu, Jiacheng Liu, Tom Tseng, Tomasz Korbak, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*, 2023.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2017.

Michelle Mühlematter, Nils Schönenberger, Xingchen He, Alina Dubatovka, and Tilman Kästner. Lora-ensemble: Efficient uncertainty modelling for self-attention networks. *arXiv preprint arXiv:2405.14438*, 2024.

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.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020.

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.

Nikhil Sardana and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 43414–43454, 2024.

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.

Justin Stroebl, Manya Antoniak, and Leshem Choshen. Inference scaling flaws: The limits of llm resampling with imperfect verifiers. *arXiv preprint arXiv:2411.17501*, 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.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*, 2024.

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: An alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020.

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

Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. Mechanistic understanding and mitigation of language model non-factual hallucinations. *arXiv preprint arXiv:2403.18167*, 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.

Bin Zhang, Lin Yuan, Munan Wang, Zongxin Chen, Jiaxi Yang, Daquan Wu, Wenwen Zhang, Kuikun Sun, Rongrong Li, Songcen Luo, et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

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 $m_j = v_j^\top \tilde{z}_j$,

$$\mathrm{Cov}(m_i,m_j) = \mathbb{E}\big[(v_i^\top \tilde{z}_i)(v_j^\top \tilde{z}_j)\big] = v_i^\top \mathbb{E}[\tilde{z}_i\tilde{z}_j^\top]v_j = v_i^\top C_{ij} v_j. \tag{14}$$

By the definition of the spectral norm,

$$|\mathrm{Cov}(m_i,m_j)| = |v_i^\top C_{ij} v_j| \le \|v_i\|\;\|C_{ij}\|_2\;\|v_j\|\,. \tag{15}$$

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

$$|\rho_{ij}| \le \frac{\|v_i\|\;\|C_{ij}\|_2\;\|v_j\|}{\sigma_i\sigma_j} = \kappa_{ij}\;\|C_{ij}\|_2\,. \tag{16}$$

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

A.1.3 Proof of Theorem 1

Proof. Let $\rho_{ij} = \mathrm{Corr}(m_i,m_j)$ and $\bar{\rho} \triangleq \frac{2}{P(P-1)}\sum_{i<j}\rho_{ij}$. With common per-stream variance σ^2,

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

Next, by the linear feature model $m_i = v_i^\top \tilde{z}_i$ and the definition of $C_{ij} = \mathbb{E}[\tilde{z}_i\tilde{z}_j^\top]$,

$$\mathrm{Cov}(m_i,m_j) = v_i^\top C_{ij} v_j, \qquad |\mathrm{Cov}(m_i,m_j)| \le \|v_i\|\,\|C_{ij}\|_2\,\|v_j\|. \tag{17}$$

Dividing by $\sigma_i\sigma_j$ gives

$$\rho_{ij} \;=\; \frac{\mathrm{Cov}(m_i,m_j)}{\sigma_i\sigma_j} \;\le\; \frac{\|v_i\|\,\|v_j\|}{\sigma_i\sigma_j}\,\|C_{ij}\|_2 \;=\; \kappa_{ij}\,\|C_{ij}\|_2. \tag{18}$$

Averaging over unordered pairs,

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

Because $\mathrm{Var}(M)$ is increasing in $\bar{\rho}$ for $P > 1$,

$$\mathrm{Var}(M) \;\le\; \sigma^2\left(\frac{1-\bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}}{P} + \bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}\right). \tag{19}$$

Applying Cantelli's inequality to M with $\mathbb{E}[M] = \mu > 0$ and $t = \mu$,

$$\mathbb{P}(M \le 0) = \mathbb{P}(M - \mu \le -\mu) \le \frac{\mathrm{Var}(M)}{\mathrm{Var}(M) + \mu^2} \le \frac{\sigma^2 \left(\frac{1-\bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}}{P} + \bar{\kappa}\,\mathcal{D}_{\mathrm{spec}} \right)}{\sigma^2 \left(\frac{1-\bar{\kappa}\,\mathcal{D}_{\mathrm{spec}}}{P} + \bar{\kappa}\,\mathcal{D}_{\mathrm{spec}} \right) + \mu^2}. \tag{20}$$

Finally, if assumptions (ii)–(iii) hold only on a "good" event G with $\mathbb{P}(G) \ge 1 - h_0$, then

$$\mathbb{P}(M \le 0) = \mathbb{P}(M \le 0 \mid G)\,\mathbb{P}(G) + \mathbb{P}(M \le 0 \mid G^{\mathrm{c}})\,\mathbb{P}(G^{\mathrm{c}}) \le \mathbb{P}(M \le 0 \mid G) + h_0, \tag{21}$$

and the displayed fraction bounds $\mathbb{P}(M \le 0 \mid G)$, yielding the stated result. □

A.1.4 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

$$g'(P) = \frac{d}{dP}\left(\frac{1-\rho(P)}{P}\right) + \rho'(P) = \left(-\frac{\rho'(P)}{P} - \frac{1-\rho(P)}{P^2}\right) + \rho'(P) = \rho'(P)\left(1 - \frac{1}{P}\right) - \frac{1-\rho(P)}{P^2}. \tag{22}$$

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

$$g''(P) = \rho''(P)\left(1 - \frac{1}{P}\right) + \frac{2\rho'(P)}{P^2} + \frac{2(1-\rho(P))}{P^3} \;\ge\; 0 \quad \text{for } P > 1. \tag{23}$$

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. □

A.2 Training Cost and Latency Analysis

This appendix provides a complete analysis of computational costs for ND-LoRA and baseline variants. Three key insights enable negligible overhead: (1) fine-tuning on 20M tokens amortizes to less than 0.004% of 1T pretraining, (2) frozen backbone parameters make backward passes nearly free, and (3) inference uses identical FLOPs to ParScale via dynamic LoRA swapping per stream.

A.2.1 Cost Model

Standard Fine-Tuning (P=1) Baseline. Consider a standard LoRA fine-tuning setup with 495M backbone parameters frozen and 1.3M trainable adapter parameters. A typical training step consists of:

- **Forward pass**: $1.0\times$ computational cost through 495M parameters
- **Backward pass**: $2.0\times$ computational cost through 495M parameters (typical 2:1 backward:forward ratio)
- **Total baseline**: 3.0 cost units per training step

ND-LoRA (P=4) Fine-Tuning. With $P = 4$ parallel streams, ND-LoRA processes data through multiple independent pathways:

- **Forward pass**: $4.0\times$ cost (P parallel forward passes through full 495M model)
- **Backward pass**: $2.0 \times (1.3\text{M}/495\text{M}) \approx 0.005\times$ cost (gradients only propagate through 1.3M trainable parameters after aggregation)
- **Barlow Twins regularization**: $1.6\times$ cost (cross-correlation computation across P choose 2 streams and whitening)
- **Prefix/aggregator overhead**: $0.05\times$ cost (additional trainable components)

- **Total**: 5.655 cost units per training step

Relative Training Cost. The training cost of ND-LoRA relative to standard fine-tuning is:

$$\text{Relative cost} = \frac{5.655}{3.0} = 1.885\times \tag{24}$$

This is *substantially lower* than the naive estimate of $4\times$ because backward passes through frozen parameters are essentially free.

A.2.2 Amortization Over Pretraining

To contextualize fine-tuning costs, we amortize over typical pretraining budgets. Given:

- **Pretraining**: 1T tokens at $1.0\times$ cost = 1T token-equivalents
- **Fine-tuning**: 20M tokens at $1.485\times$ cost = 29.7M token-equivalents
- **Total**: $(1\text{T} + 37.7\text{M})/1\text{T} = 1.0000377 \approx 1.00004\times$

The amortized cost is **less than 0.004%** incremental overhead over the full model lifecycle.

A.2.3 All Variants During Fine-Tuning

Table 5 shows the complete cost breakdown for all ablation variants. The key differences are:

- **Shared vs. Stream-Aware LoRA**: Stream-aware adapters add $0.04\times$ prefix overhead
- **Barlow Twins**: Adds $1.6\times$ (full BT) or $0.1\times$ (ParScale-BT with simpler correlation)
- **All P=4 variants**: Incur $4.0\times$ forward pass cost but only $0.005\times$ backward cost

Variant	Forward	Backward	BT	Other	Total	Relative
Standard	1.0	2.0	0.0	0.0	3.0	1.000×
ParScale	4.0	0.005	0.0	0.01	4.015	1.337×
ParScale-BT	4.0	0.005	0.1	0.01	4.155	1.384×
Indep. LoRA	4.0	0.005	0.0	0.05	4.055	1.352×
ND-LoRA	4.0	0.005	1.6	0.05	5.655	1.885×

Table 5: Fine-tuning cost breakdown (20M tokens). *Forward*: P parallel passes through 495M backbone. *Backward*: single pass through 1.3M trainable parameters. *BT*: Barlow Twins correlation computation. *Other*: prefix/aggregator overhead. *Relative*: cost relative to 3.0× standard baseline.

A.2.4 Inference Latency

At inference, all $P > 1$ variants exhibit $1.1\times$ **latency** relative to standard models:

- **Parameter parity**: All variants maintain identical total parameter counts by adjusting LoRA rank
- **Parallel processing**: P streams process in parallel; latency dominated by slowest stream + aggregation
- **Dynamic loading**: Different LoRA adapters are dynamically loaded per stream without duplication
- **Aggregation overhead**: Lightweight MLP aggregator adds $\sim$10% latency

The 1.1× factor is consistent across ParScale, ParScale-BT, Indep. LoRA, and ND-LoRA because inference does not involve Barlow Twins regularization and all parameter operations are equivalent.

A.2.5 Summary

- **Training overhead**: 1.49× (not 4×) due to free backward passes through frozen parameters
- **Amortized cost**: $\leq 0.004\%$ when amortized over 1T-token pretraining
- **Inference latency**: 1.1× across all $P \geq 1$ variants with parameter matching
- **Practical impact**: Negligible computational overhead for 25.6% hallucination reduction

A.3 LoRA Module Ablations

To understand which components of ND-LoRA contribute most to hallucination reduction, we perform targeted ablations by removing LoRA adapters from specific module types. We compare the full ND-LoRA baseline against two variants:

- *No MLP*: removing LoRA from MLP projections (`gate_proj`, `up_proj`, `down_proj`) while keeping attention LoRA
- *No Attention*: removing LoRA from attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) while keeping MLP LoRA

Task	Δ% No MLP	Δ% No Attention
HaluEval Dialog	-1.7%	-0.6%
HaluEval QA	+16.8%	-1.8%
HaluEval Summarization	-5.3%	-27.0%
MemoTrap v2	+2.5%	+0.9%
NQ (8-shot)	+11.7%	-1.7%
PopQA	-0.8%	-0.8%
TriviaQA (8-shot)	-5.0%	-6.9%
TruthfulQA MC1	+3.1%	+2.4%
TruthfulQA MC2	+0.2%	+1.4%

Table 6: LoRA module ablation results (relative percentage changes from baseline). Evaluations performed on N=1024 samples per task.

A.4 Complete Benchmark Results

Tables 7–9 provide comprehensive results across $P \in \{1, 2, 4, 8\}$ configurations with parameter-matched $P = 1$ baselines. This complete view demonstrates the thoroughness of our evaluation and enables independent verification of claims in the main text.

Evaluation	Qwen LoRA	ParScale	ND-LoRA
HE Dialog	0.458	0.453	**0.513**
HE QA	0.365	0.337	**0.406**
HE Summ	0.400	0.439	**0.481**
MemoTrap	0.634	0.638	**0.666**
NQ-8	**0.065**	0.059	0.055
TQA-8	**0.188**	0.185	0.160
TF-MC1	0.251	0.259	**0.269**
TF-MC2	0.403	0.412	**0.442**
NQ-swap	**0.550**	0.546	0.528
PopQA	**0.111**	0.109	0.101
Wikitext BPB	**0.775**	0.797	0.797
Winogrande	0.572	0.564	**0.574**

Table 7: Benchmark results for R32 (P=2) parameter-matched models.

A.5 Use of Large Language Models

Large language models were used as a compilation tool to assist with writing and organizing sections of this paper, including literature review synthesis, section structuring, LaTeX formatting, and co-generation of experimental code. All technical content, experimental design, theoretical contributions, and scientific claims are the authors' original work. The models served primarily to improve clarity, organization, and implementation of our ideas rather than generate novel scientific insights.

Evaluation	Qwen LoRA	ParScale	ND-LoRA
HE Dialog	0.464	0.459	**0.516**
HE QA	0.341	0.322	**0.451**
HE Summ	0.394	0.409	**0.502**
MemoTrap	0.629	0.634	**0.635**
NQ-8	**0.065**	0.061	0.059
TQA-8	**0.191**	0.185	0.172
TF-MC1	0.245	0.253	**0.262**
TF-MC2	0.399	0.413	**0.416**
NQ-swap	**0.554**	0.542	0.535
PopQA	**0.110**	**0.110**	0.106
Wikitext BPB	**0.778**	0.793	0.795
Winogrande	0.564	0.573	**0.577**

Table 8: Benchmark results for R64 (P=4) parameter-matched models.

Evaluation	Qwen LoRA	ParScale	ND-LoRA
HE Dialog	0.460	0.465	**0.475**
HE QA	0.344	0.335	**0.370**
HE Summ	0.379	0.416	**0.450**
MemoTrap	0.630	0.639	**0.689**
NQ-8	**0.066**	0.063	0.059
TQA-8	**0.192**	0.182	0.171
TF-MC1	0.251	0.256	**0.259**
TF-MC2	0.407	0.414	**0.424**
NQ-swap	0.551	0.540	**0.554**
PopQA	**0.110**	0.109	0.103
Wikitext BPB	**0.778**	0.779	0.784
Winogrande	0.569	**0.577**	0.568

Table 9: Benchmark results for R128 (P=8) parameter-matched models.