A MECHANISTIC EXPLANATION OF CHSH VIOLATIONS IN CLASSICAL NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Non-classical correlations, certified by violations of Bell–CHSH-type inequalities entailed by classical probability theory, are generally assumed to lie beyond the expressive capacity of classical neural networks. In this work, we show that a CHSH-style statistic S in standard feedforward multi-task networks can exceed the classical bound of 2 inside a specific "emergence window", thereby exhibiting non-classical correlations. We further provide a mechanistic explanation of this phenomenon: under the shared-capacity bottleneck of multi-task learning, gradient competition couples hidden neurons; under ordinary feedforward evaluation, easy and hard task pairs become unequally learnable, which unbalances the four correctness correlations and drives $S > 2$. We observe this pattern in numerical simulation, text classification, and image recognition, and further verify that the window is governed by the alignment between network capacity and task complexity. Using CHSH-type inequalities as a probe, this work shows that the learning dynamics of classical networks can also give rise to CHSH-style correlational excesses, and offers a conceptual basis for non-classical statistical modeling in quantum-inspired methods.

1 INTRODUCTION

Non-classical correlations describe statistical dependencies that cannot be reproduced by any local hidden-variable model (Clauser & Shimony, 1978). The Clauser–Horne–Shimony–Holt (CHSH) inequality (Clauser et al., 1969) turns this boundary into a quantitative test: any statistics embeddable in a single classical (Kolmogorov) probability space must satisfy $S \leq 2$ (Brunner et al., 2014; Fine, 1982). A measured excess over this bound is therefore often read as a signature of non-classicality. In this paper we adopt the informational reading of such excesses as contextuality rather than physical nonlocality (Abramsky & Brandenburger, 2011; Barbosa et al., 2022): CHSH-type violations certify the failure of a non-contextual classical model, not the presence of quantum entanglement in silicon.

Classical neural networks are trained and evaluated under classical probability (Goodfellow et al., 2016), and are therefore widely presumed unable to exhibit non-classical correlations (Zhang et al., 2025; Kak, 2007). This presumption has erected a conceptual barrier at the intersection of quantum theory and artificial intelligence: the underlying logic of classical networks is often taken to be at odds with non-classicality (Kak, 2007). Although cross-disciplinary efforts in this area have produced promising results, most quantum–ML approaches still rely on injecting Hilbert-space structure, tensor networks, or other explicit quantum ingredients into the architecture (Kriváchy et al., 2020; Stoudenmire & Schwab, 2016; Zhang et al., 2023). This leaves a more fundamental question for quantum-inspired deep learning: is there a deeper connection between classical neural networks and non-classicality that does not presuppose the explicit addition of quantum components? The answer bears directly on the conceptual rationale of quantum-inspired machine learning.

To probe this question, we use Bell–CHSH-type inequalities as a diagnostic and construct a CHSH-style statistic S in feedforward networks to reflect correctness correlations arising in multi-task learning. Direct empirical evidence shows that, across controlled numerical simulations, multi-label text classification (AAPD), and multi-label image recognition (COCO), standard feedforward multi-task networks can produce $S > 2$ inside a finite capacity "emergence window", thereby exhibiting non-classical statistical correlations in the CHSH sense. Notably, this non-classicality is typically transient: it appears while training loss is still relatively high and falls back toward or below 2 after

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

the loss converges. We also emphasize that a CHSH violation is interpreted here as non-classicality at the probabilistic level in the network's learning dynamics, not as evidence of physical nonlocality.

We then give a mechanistic explanation of this phenomenon. Under the shared-capacity bottleneck of multi-task learning, gradient competition couples hidden neurons; under ordinary feedforward evaluation, this coupling prevents a shared representation from supporting all multi-task objectives equally well, which unbalances the four correctness correlations that enter S and drives $S > 2$. We further explain why the emergence window arises: it is governed by the relative match between task complexity and network capacity. When capacity is overly abundant relative to the task, objectives tend to decouple in the shared layer, so competition and coupling are too weak to induce the imbalance above; when capacity is overly scarce, the network underfits broadly, correctness correlations collapse together, and $S > 2$ likewise fails to persist. Only in an intermediate regime is competition strong enough to induce coupling and correlational imbalance, while the network can still attain usable predictive accuracy on most task settings, so that S exceeds 2. In summary, CHSH violations in classical networks arise from internal learning dynamics under multi-task training and do not rely on explicit quantum components; yet the conditions for their appearance are stringent and depend on the alignment between task difficulty and network capacity.

We conduct targeted experiments to support this explanation. First, gradient–S trajectories show that elevated S (i.e., non-classical correlations) concentrates in a moderate gradient-competition regime, rather than when competition is too weak or too strong. Second, we examine factors that shape the emergence window: capacity sweeps show that peak S follows an inverted-U as a function of network capacity (hidden width), and that harder tasks shift the window toward wider networks; under approximately matched parameter budgets, globally coupled architectures (e.g., fully connected networks and LSTMs) more readily produce $S > 2$, whereas locally constrained ones (e.g., CNNs and attention models) suppress it.

Our contributions are threefold:

- We adopt the CHSH inequality as a diagnostic probe and show that a CHSH-style statistic S exceeds 2 in classical feedforward multi-task networks across numerical simulation, text classification (AAPD), and image recognition (COCO), indicating non-classical statistical correlations inside a finite emergence window.
- Starting from gradient competition under a shared-capacity bottleneck, we explain this classical excess over the CHSH bound and, through a capacity–task alignment perspective, account for why the phenomenon appears only inside that window.
- Through gradient–S trajectories, capacity/complexity sweeps, and architecture comparisons under matched parameter budgets, we further analyze the factors that govern the non-classical window.

2 Related Work

2.1 Non-classical Correlations

Non-classical correlations were quantified in the context of Bell's theorem (Bell, 1964), which showed that no local hidden-variable model can reproduce all quantum predictions (Einstein et al., 1935). The CHSH inequality (Clauser et al., 1969) provides an experimentally testable form of this result, establishing a classical upper bound of $S \leq 2$ under the assumptions of locality, realism, and measurement independence (Brunner et al., 2014). Quantum systems can violate this bound via entanglement, reaching Tsirelson's limit of $2\sqrt{2}$ (Cirel'son, 1980).

In machine learning, a line of research exploits the powerful fitting capacity of classical feed-forward neural networks to detect whether a given data distribution contains intrinsic non-classical correlations. The working hypothesis is that if a classical network, after sufficient training, exhibits persistently large fitting error on a distribution (even when the network capacity is increased), that distribution likely cannot be reproduced by any local hidden-variable model. Kriváchy et al. (Kriváchy et al., 2020) pioneered this approach by training a neural network as an "oracle" to learn classical strategies for reproducing a probability distribution; persistent failure of the network certifies nonlocality. Subsequent studies have extended this methodology to identify nonclassicality of quantum states of light from experimental data (Gebhart et al., 2021), detect nonlocality in chain-shaped quantum

networks via deep learning (Zhang et al., 2024), and recognize quantum contextuality using Bayesian neural networks (Wasilewski et al., 2023). Emonts et al. (Emonts et al., 2024) also employ neural networks to analyze the effects of topological boundary conditions on Bell nonlocality. In all these works, the neural network serves as an external probe for analyzing external data. In sharp contrast, our work investigates whether a classical neural network can endogenously develop non-classical correlations within its own learning dynamics, without any quantum assumptions or architectural modifications.

2.2 Quantum-Inspired Models and Emergent Dynamics

Quantum-inspired models incorporate structures such as Hilbert spaces, tensor networks, or complex-valued representations into classical architectures to improve expressivity (Stoudenmire & Schwab, 2016; Schuld & Petruccione, 2021; Zhang et al., 2023; Novikov et al., 2015). These methods have been successfully applied across NLP, vision, and recommendation systems (Phukan & Ekbal, 2025; Ling et al., 2024; Aizpurua et al., 2026; Tang, 2019; Debi & Makmal, 2025).

However, such approaches rely on explicitly embedding quantum formalism into network design, raising questions about their conceptual necessity (da Silva & de Oliveira, 2016). This motivates the study of whether non-classical behaviors can emerge purely from learning dynamics.

From this perspective, prior work in multi-task learning has extensively studied gradient interference and capacity constraints (Yu et al., 2020; Wang et al., 2020; Hashimoto et al., 2024), typically as optimization challenges. In contrast, we reinterpret gradient competition as a generative mechanism: under capacity bottlenecks, it couples shared representations and unbalances the correctness correlations that enter a CHSH-style statistic, thereby driving $S > 2$.

3 Preliminaries

We briefly review the CHSH inequality and the Non-Classical Network (NCnet) protocol of Zhao et al. (Zhao et al., 2026), which we use as a controlled multi-task setting for constructing a CHSH-style statistic from classical feedforward networks.

3.1 The CHSH Inequality

The Clauser-Horne-Shimony-Holt (CHSH) inequality (Clauser et al., 1969) provides a quantitative test for non-classical correlations. Consider two parties, Alice and Bob, each choosing between two measurement settings ($i \in \{1, 2\}$ for Alice, $j \in \{1, 2\}$ for Bob) and obtaining binary outcomes $A_i, B_j \in \{-1, +1\}$. Define $\mathbb{E}[A_i B_j]$ as the correlation between the outcomes. The CHSH statistic is

$$S = \mathbb{E}[A_1 B_1] + \mathbb{E}[A_1 B_2] + \mathbb{E}[A_2 B_1] - \mathbb{E}[A_2 B_2]. \tag{1}$$

Under local hidden-variable assumptions, $|S| \leq 2$ must hold (Brunner et al., 2014). An excess $|S| > 2$ means that the four correlations cannot arise from a single classical joint distribution over (A_1, A_2, B_1, B_2) compatible with those assumptions. In quantum mechanics, entangled states can reach at most the Tsirelson (Cirel'son) bound $|S| \leq 2\sqrt{2} \approx 2.828$ (Cirel'son, 1980).

In this paper, A_i and B_j are *correctness* indicators of neural predictions rather than physical measurement outcomes, and S is a corresponding *CHSH-style* diagnostic of multi-task learning dynamics. Because the entries of S are assembled from separately trained pair settings (Section 3.2), we do not claim a standard Bell experiment: the classical threshold 2 is used as a reference for correlational imbalance, and the algebraic range of Eq. (1) for ± 1-valued factors is $[-4, 4]$ (so values above Tsirelson's bound are possible and observed). We make no claim of physical nonlocality.

3.2 The Non-Classical Network

Zhao et al. (Zhao et al., 2026) proposed NCnet as a simple feedforward architecture for studying when such an S can exceed 2. As illustrated in Figure 1, NCnet consists of:

- **Input layer**: four binary variables $X_1, X_2, X_3, X_4 \in \{0, 1\}$.

- **Hidden layer**: n neurons with ReLU activation, shared between two task heads.
- **Output layer**: two independent task heads, corresponding to Alice and Bob.

The underlying targets are

$$\begin{aligned} \text{Alice:} \quad & \alpha_1 = X_1, \quad \alpha_2 = X_1 \oplus X_2, \\ \text{Bob:} \quad & \beta_1 = X_3, \quad \beta_2 = X_3 \oplus X_4, \end{aligned} \tag{2}$$

where $\oplus$ denotes XOR. For each pair setting (α_i, β_j) with $i, j \in \{1, 2\}$, we train a *separate* two-head network with hard parameter sharing to predict α_i and β_j jointly. The four settings are therefore four related multi-task problems rather than a single four-head model. After training on setting (i, j), we define correctness indicators on held-out inputs for that network,

$$A_i^{(i,j)} = \begin{cases} +1 & \text{if the prediction of } \alpha_i \text{ is correct,} \\ -1 & \text{otherwise,} \end{cases} \qquad B_j^{(i,j)} = \begin{cases} +1 & \text{if the prediction of } \beta_j \text{ is correct,} \\ -1 & \text{otherwise.} \end{cases} \tag{3}$$

Each pairwise correlation is estimated within its own trained setting, $C_{ij} := \widehat{\mathbb{E}}[A_i^{(i,j)} B_j^{(i,j)}]$, and we assemble the CHSH-style statistic

$$S = C_{11} + C_{12} + C_{21} - C_{22}. \tag{4}$$

When unambiguous we still write $C(A_i, B_j)$ for C_{ij}. Because $A_i^{(i,1)}$ and $A_i^{(i,2)}$ come from different trained networks, Eq. (4) is *not* a standard Bell–CHSH statistic over one joint outcome tuple; it is a diagnostic that scores imbalance among the four pair-wise correctness correlations under matched protocols. Throughout the paper, "$S > 2$" and "CHSH-style non-classical correlations" refer to this assembled diagnostic.

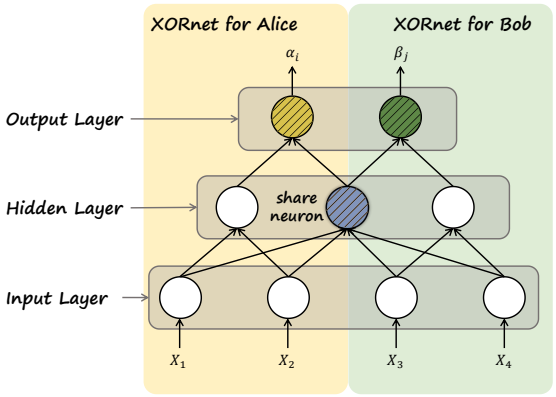


Figure 1: Architecture of NCnet (Zhao et al., 2026).

Zhao et al. (Zhao et al., 2026) report that, in their setting, $S > 2$ appears only near a critical hidden size (e.g., $n = 3$) and vanishes for smaller or larger capacities. Using this protocol as a controlled starting point, Section 4 documents when $S > 2$ arises across broader data regimes and that the excess is typically transient; Section 5 provides a mechanistic explanation, which Section 6 supports experimentally.

4 Direct Evidence of CHSH-Style Non-Classical Correlations

Under the NCnet protocol of Zhao et al. (Zhao et al., 2026), we provide direct empirical evidence that a CHSH-style statistic $S > 2$ can appear across diverse settings—synthetic Data Types 1–3 as well as multi-label image recognition (COCO) and text classification (AAPD). Importantly, the excess is typically *transient*: it arises mid-training and usually recedes by convergence. A mechanistic explanation and finer-grained controls follow in Sections 5–6.

4.1 Experimental Setup

We evaluate the CHSH-style diagnostic of Section 3.2. For each pair setting (α_i, β_j) we train a separate two-head network, estimate $C_{ij} = \widehat{\mathbb{E}}[A_i^{(i,j)} B_j^{(i,j)}]$ on held-out inputs, and assemble $S = C_{11} + C_{12} + C_{21} - C_{22}$. We consider three synthetic generators (Data Types 1–3; Appendix B). On COCO and AAPD we apply the same protocol by associating input features with the four settings; the table rows labeled Type 1–3 denote three matched feature/label groupings of increasing structural complexity under that protocol (Appendix D.2).

Unless otherwise noted, we train single-hidden-layer ReLU networks with two output heads, optimized by Adam with mean squared error (MSE). Each configuration is repeated over 20 random seeds. Learning rates, hidden widths, and related hyperparameters are summarized in Appendix C.

Table 1: CHSH-style statistics on synthetic data types. Values are mean ± standard deviation across seeds. $S_{\max}$ is the peak S during training; $S > 2$ duration is consecutive epochs with $S > 2$; S_{fin} is S at the last epoch; Loss at $S_{\max}$ is the training loss at the peak-S epoch; Violation rate is the fraction of seeds with some $S > 2$.

Data Type	Maximum $S_{\max}$ (mean ± std)	$S > 2$ duration (epochs)	Final S_{fin} (mean ± std)	Loss at $S_{\max}$ (mean ± std)	Violation Rate (%)
1	**2.131** ± 0.245	3.2 ± 5.7	1.825 ± 0.225	0.458 ± 0.214	35%
2	**2.296** ± 0.156	4.5 ± 5.1	1.972 ± 0.084	0.335 ± 0.195	80%
3	**2.053** ± 0.116	2.5 ± 2.7	1.033 ± 0.425	0.247 ± 0.024	70%

Table 2: CHSH-style statistics on COCO.

Data Type	Maximum $S_{\max}$ (mean ± std)	$S > 2$ duration (epochs)	Final S_{fin} (mean ± std)	Loss at $S_{\max}$ (mean ± std)	Violation Rate (%)
1	**3.565** ± 0.192	2.6 ± 0.8	1.978 ± 0.027	1.000 ± 0.001	75.0%
2	**3.997** ± 0.005	2.2 ± 1.2	1.151 ± 0.943	1.002 ± 0.002	90.0%
3	**3.890** ± 0.013	3.6 ± 1.6	2.041 ± 1.217	1.001 ± 0.001	70.0%

4.2 Results

Table 1 reports CHSH-style statistics on the three synthetic data types; Tables 2 and 3 report the corresponding statistics on COCO and AAPD. Figure 2 illustrates training dynamics for synthetic Data Type 2. In all tables, "$S > 2$ duration" is the average number of consecutive epochs with $S > 2$ (a temporal length, not the capacity emergence window discussed later).

Existence across domains. In every setting we consider—synthetic Types 1–3, COCO, and AAPD—mean $S_{\max}$ exceeds 2. Synthetic peaks remain moderate (about 2.05–2.30), whereas several real-data configurations reach much larger peaks (often above 3, and approaching the algebraic upper value 4 in some COCO/AAPD rows). As noted in Section 3.1, values above Tsirelson's bound are expected for this assembled diagnostic. Thus large CHSH-style excesses are not confined to toy generators.

Transience. Despite large peaks, S_{fin} typically falls back toward or below 2. For example, synthetic Type 2 goes from 2.296 to 1.972, and COCO Type 2 from ≈ 3.997 to ≈ 1.151; AAPD shows the same qualitative drop. The $S > 2$ duration stays short (usually a few epochs). The excess is therefore typically transient and does not characterize the converged model.

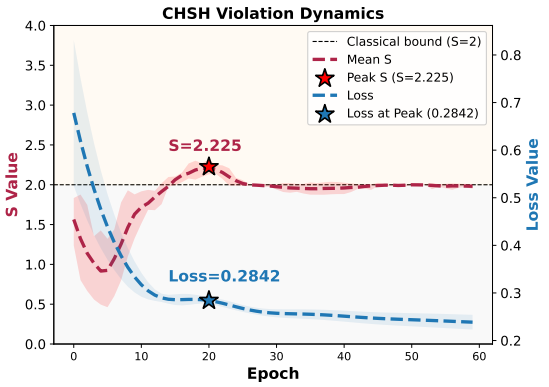


Figure 2: Mean ± standard deviation of CHSH-style S (red) and training loss (blue) over epochs for Data Type 2. Red/blue stars mark peak $S_{\max}$ and the loss at that epoch. Shaded regions denote ±1 standard deviation.

Mid-training peak. Figure 2 shows that S crosses 2 while training loss is still relatively high, then recedes as optimization continues. The same mid-training pattern appears for Types 1 and 3 in Appendix D. Peak S is therefore aligned with an intermediate stage of learning rather than with the final optimum.

Dependence on data structure. Within each domain, violation rates and peak heights vary by type (e.g., synthetic and COCO Type 2 are among the strongest; some Type 1/Type 3 settings are weaker). Hence both the occurrence and the strength of $S > 2$ depend on the input distribution, rather than appearing uniformly.

Table 3: CHSH-style statistics on AAPD.

Data Type	Maximum $S_{\max}$ (mean $\pm$ std)	$S > 2$ duration (epochs)	Final S_{fin} (mean $\pm$ std)	Loss at $S_{\max}$ (mean $\pm$ std)	Violation Rate (%)
1	**3.698** $\pm$ 0.069	2.6 $\pm$ 1.0	1.001 $\pm$ 0.495	0.999 $\pm$ 0.004	75.0%
2	**3.744** $\pm$ 0.126	2.4 $\pm$ 1.9	1.432 $\pm$ 0.702	1.001 $\pm$ 0.003	85.0%
3	**3.186** $\pm$ 0.288	3.0 $\pm$ 1.4	1.473 $\pm$ 0.856	1.000 $\pm$ 0.008	60.0%

These results establish that transient $S > 2$ arises across synthetic and real multi-task settings, yet the underlying mechanism remains unclear. We therefore analyze the phenomenon from a multi-task learning perspective: Section 5 develops a mechanistic explanation, and Section 6 tests it.

5 A Mechanistic Explanation of CHSH-Style Non-Classical Correlations

Section 4 shows that a CHSH-style statistic can exceed 2, but does not explain why. We give a working explanation in the NCnet setting (Figure 1). Write

$$C(A_i, B_j) \;:=\; C_{ij} = \widehat{\mathbb{E}}\big[A_i^{(i,j)} B_j^{(i,j)}\big] \tag{5}$$

for the pair-wise correctness correlation entering the assembled S of Section 3.2. As in that section, each pair (α_i, β_j) is trained as a separate two-head multi-task problem; S is assembled from the four C_{ij}. Under a shared-capacity bottleneck, gradient competition couples hidden neurons; under ordinary feedforward evaluation, this coupling makes the four task pairs unequally learnable, which unbalances the four correctness correlations and drives $S > 2$—but only inside an intermediate capacity regime. We develop this explanation in four steps below.

Step 1: Representational Bottleneck and Gradient Competition In NCnet the four targets have unequal demands: each identity map needs only one ReLU unit, while a single XOR needs at least two, and α_2, β_2 act on disjoint inputs. Hosting both XOR features without forced overlap therefore requires about $2 + 2 = 4$ hidden neurons; identity heads can reuse units from those XOR subcircuits and need not add extra width in this estimate. When the width n falls short, a *representational bottleneck* forces feature sharing and interferes with updates on shared weights. Within a trained Alice–Bob pair we quantify this conflict on neuron h_k by

$$\mathcal{G}_k \;=\; 1 - \cos\big(\nabla_{\mathbf{w}_k} \mathcal{L}_A,\; \nabla_{\mathbf{w}_k} \mathcal{L}_B\big), \tag{6}$$

where $\mathcal{L}_A$ and $\mathcal{L}_B$ are the Alice and Bob losses (undefined gradients are ignored in practice). Then $\mathcal{G}_k \in [0, 2]$: 0 means aligned head gradients and 2 means fully opposed ones, matching the two-head probe in Section 6.1.

Step 2: From Gradient Competition to Neural Coupling Competition on shared parameters induces statistical dependence among hidden units. For neurons h_k, h_l that tend to serve different heads, let ρ_{kl} be their Pearson correlation and $\bar{\mathcal{G}}_{kl} = \frac{1}{2}(\mathcal{G}_k + \mathcal{G}_l)$. When head gradients on shared weights are strongly opposed, updates cannot specialize the two units independently and tend to push them toward co-activation. Differentiating ρ_{kl} along gradient flow yields an exact but high-dimensional identity (Appendix A.1); we summarize the dominant balance by the phenomenological ODE

$$\frac{d\rho_{kl}}{dt} = \alpha \bar{\mathcal{G}}_{kl}(1 - \rho_{kl}^2) - \beta \rho_{kl}, \tag{7}$$

with $\alpha, \beta > 0$. Its stable root in $(0, 1)$ is

$$\rho_{kl}^*(\bar{\mathcal{G}}_{kl}) = \frac{-\beta + \sqrt{\beta^2 + 4\alpha^2 \bar{\mathcal{G}}_{kl}^2}}{2\alpha \bar{\mathcal{G}}_{kl}} \quad (\bar{\mathcal{G}}_{kl} > 0), \tag{8}$$

with $\rho_{kl}^*(0) = 0$. Thus ρ_{kl}^* increases with $\bar{\mathcal{G}}_{kl}$ and approaches 1 under strong, sustained competition: bottleneck conflict promotes *neural coupling*.

Step 3: From Neural Coupling to Correlational Imbalance Step 2 established that bottleneck competition can drive $\rho_{kl} \approx 1$: the two hidden units rise and fall together, so the shared layer no longer carries two independently usable features. What remains is essentially a single co-activation mode $z \propto h_k + h_l$—an approximately one-dimensional code. Evaluation is still an ordinary forward pass: inputs are fed through the network, head correctness yields A_i and B_j, and the four correlations $C(A_i, B_j)$ are estimated; no latent is resampled or adapted at test time.

Under this collapsed code the four pair settings are no longer equally solvable. We call settings that include an identity map (e.g., (α_1, β_j) or (α_i, β_1)) the *easier* pairs: both heads can often read a simple input bit from the same z, so the corresponding $C(A_i, B_j)$ stays near 1. The *hard* pair $(\alpha_2, \beta_2) = (X_1 \oplus X_2,\ X_3 \oplus X_4)$ needs two independent nonlinear features on disjoint inputs; a one-dimensional shared code cannot implement both maps at once, so at least one head errs frequently and $C(A_2, B_2)$ is depressed. Coupling therefore unbalances the four CHSH terms under feedforward evaluation (Appendix A.2): easier-pair correlations stay high while the hard-pair correlation lags.

Step 4: From Correlational Imbalance to $S > 2$ Step 3 leaves easier-pair correlations high and the hard-pair correlation depressed. Collect the former into $C_+ = \frac{1}{3}\big(C(A_1, B_1) + C(A_1, B_2) + C(A_2, B_1)\big)$ and write $C_{22} = C(A_2, B_2)$. Then

$$S = 3C_+ - C_{22} = 2C_+ + \mu_{\text{eff}}, \qquad \mu_{\text{eff}} := C_+ - C_{22}. \tag{9}$$

In the working regime where the easier pairs remain accurate ($C_+ \approx 1$), one has $S \approx 2 + \mu_{\text{eff}}$, so $S > 2$ if and only if $\mu_{\text{eff}} > 0$, i.e., when the hard pair lags ($C_{22} < C_+$). Removing the bottleneck lets C_{22} catch up, so $\mu_{\text{eff}} \to 0$ and S returns toward 2. The same identity clarifies mid-training transience: as optimization continues, either specialization reduces coupling or C_{22} improves, shrinking μ_{eff} even if an earlier checkpoint had $S > 2$.

Emergence window. The same gap explains why $S > 2$ appears only under intermediate capacity. If the network is wide enough for heads to specialize, coupling stays weak and C_{22} catches up with C_+; if capacity is too scarce or competition too destructive, even the easier pairs fail and C_+ collapses with C_{22}. Only in between is coupling strong enough to depress the hard pair while easier pairs remain accurate—the finite *emergence window* tested in Section 6. Appendix A.3 records a dual-factor formalization of this argument.

6 Empirical Support for the Mechanistic Explanation

We now test the links in Section 5. Gradient–S trajectories probe the competition stage; capacity and complexity sweeps probe the predicted emergence window; and an architecture comparison tests whether inductive biases that promote global coupling facilitate $S > 2$.

6.1 Gradient Competition and the Emergence of Non-classical Correlation

Experimental Setup We track how two-head gradient competition co-evolves with the assembled CHSH-style statistic during training on Data Type 2. Matching Eq. (6), for each epoch and each of the four pair settings we compute a competition intensity $\mathcal{G}$ from the cosine discrepancy between Alice and Bob gradients on that setting's shared hidden layer; we then evaluate the four C_{ij} on held-out inputs and assemble S as in Section 3.2. Reported $\mathcal{G}$–S scatter plots pool epochs across settings and seeds for that data type.

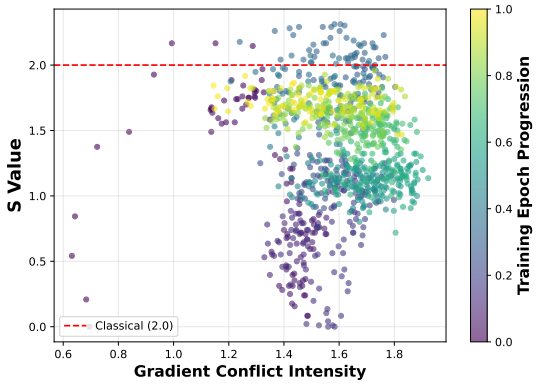


Figure 3: Joint evolution of two-head gradient competition $\mathcal{G}$ and CHSH-style S during training for Data Type 2.

Results Figure 3 visualizes the joint evolution of gradient competition and the CHSH-style statistic during optimization for Data Type 2. Each point corresponds to one optimization epoch, while the color progression indicates training time.

At the early stage of training (dark-colored points), the trajectory is broadly scattered: the shared representation has not yet stabilized, and both $\mathcal{G}$ and S fluctuate substantially.

As optimization proceeds, the dynamics concentrate into a structured region centered around $\mathcal{G} \approx 1.5 \sim 1.7$. Within this regime, higher-S states concentrate near the upper envelope of the trajectory, even though—as documented in Section 4—the excess over 2 remains transient along the training path. The strongest excesses do not appear under either minimal or maximal gradient interaction; elevated S instead concentrates at a moderate level of head conflict.

When the trajectory moves toward larger conflict ($\mathcal{G} \gtrsim 1.8$), S systematically decreases toward classical values near or below 2. This matches the emergence-window remark in Section 5: overly strong competition can destroy the coordinated shared code that keeps easier pairs accurate, so high $\mathcal{G}$ alone need not yield $S > 2$. Overall, the pattern is consistent with Steps 1–2 (bottleneck competition and coupling pressure) together with the requirement that easier-pair accuracy remain intact.

Additional experiments for Data Type 1 and Data Type 3 are provided in Appendix E. These configurations illustrate complementary boundary regimes. The $\mathcal{G}$–S trajectories corroborate the competition stage of the explanation; direct measurements of hidden-unit coupling ρ_{kl} are left to future work.

6.2 Network Capacity and Task Complexity Factors

Experimental Setup To probe the emergence window predicted in Section 5, we vary the number of hidden neurons in a single-hidden-layer ReLU network while keeping other hyperparameters fixed, thereby changing the effective capacity stress $\xi = \mathcal{C}_{\text{task}}/\mathcal{C}_{\text{network}}$. For each width we track S during training and record its maximum $S_{\max}$ over the trajectory, averaged over 20 random seeds.

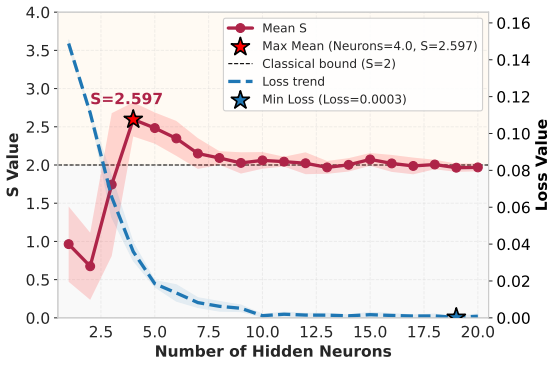


Figure 4: Capacity-induced CHSH violation for Data Type 2.

Results Figure 4 shows a clear non-monotonic pattern for Data Type 2: $S > 2$ occurs only in an intermediate width regime. When the network is under-parameterized, even easier pairs fail and the four correlations collapse together, so S stays near 2. When it is over-parameterized, the bottleneck is removed, head competition weakens, C_{22} can catch up with C_+, and S again returns toward the classical bound.

The peak ($S_{\max} = 2.597$) occurs at an intermediate width (Neurons $= 4$), consistent with the Step 1 estimate that hosting both XOR features without forced overlap needs about four hidden units, and with the finite emergence window in Section 5.

Additional Results Appendix F repeats the capacity sweep for Data Types 1 and 3. Increasing task complexity systematically shifts the window toward wider networks, consistent with control by the ratio ξ rather than by absolute width alone.

6.3 Architectural Inductive Biases

Experimental Setup. The mechanism in Section 5 relies on *global* competition in a shared representation: if an architecture isolates features across space or tokens, a width bottleneck need not force the head coupling that unbalances the four correlations. To test this, we compare four architectures on the COCO CHSH-style protocol—Fully Connected (FC), LSTM, Attention (ATT), and CNN—while

Table 4: Architecture comparison on the CHSH-style COCO protocol (matched ∼75K parameters).

Model	#Params	S Value	Violation Rate	Loss
FC	75K	2.62 ± 1.31	65%	0.15
LSTM	73K	2.23 ± 1.22	64.7%	0.13
ATT	76K	1.72 ± 1.15	29.4%	0.12
CNN	72K	1.44 ± 1.23	23.5%	0.13

matching the parameter budget to ∼75K. Training and evaluation follow the same multi-task pair protocol as above; we report mean S, violation rate, and loss across seeds.

Results. Table 4 shows a clear split along inductive bias. FC yields the strongest excess ($S = 2.62 \pm 1.31$, violation rate 65%): dense global weights route all features through one shared map, which is the setting in which head competition and coupling are expected to be strongest. LSTM also exceeds 2 ($S = 2.23 \pm 1.22$) via a shared recurrent state. By contrast, ATT and CNN stay below the classical bound on average ($S = 1.72$ and 1.44; violation rates 29.4% and 23.5%). Local receptive fields (CNN) and content-based routing (Attention) let sub-tasks specialize in largely separate pathways, weakening the shared-bottleneck pathway to correlational imbalance. Thus architectural inductive bias acts as a third control—alongside capacity and data complexity—on whether the $S > 2$ regime appears.

7 Conclusion

This work uses a CHSH-style probe to study non-classical statistical correlations arising during the training of classical neural networks. First, under the NCnet protocol, we show that a CHSH-style statistic S in feedforward multi-task networks can exceed the classical bound of 2 on synthetic data as well as on real multi-task settings (text and images). The excess typically appears mid-training while loss remains relatively high, and recedes as training converges. We then give a feedforward mechanistic explanation: a shared-capacity bottleneck induces head-wise gradient competition and hidden-unit coupling, which makes easy and hard task pairs unequally learnable, unbalances the four correctness correlations, and drives $S > 2$. Finally, targeted analyses support this account: elevated S concentrates under moderate gradient competition; capacity sweeps exhibit an inverted-U emergence window that shifts toward wider networks as tasks become harder; and globally coupled architectures facilitate the excess, whereas locally constrained ones suppress it.

These findings show that CHSH-style correlational excesses need not rely on explicit quantum components: they can emerge from the optimization dynamics of classical multi-task learning under a capacity bottleneck. This offers a new perspective on optimization and representation learning in neural networks, and a conceptual motivation for quantum-inspired diagnostics of multi-task interference.

8 Limitations

First, the mechanistic account is developed mainly under controlled multi-task settings with simplified architectures. Although we observe consistent emergence across several datasets, generality to large language models and other large-scale systems remains open.

Second, our evaluation uses a static, pairwise CHSH-style protocol in which the four C_{ij} come from separately trained pair networks. A stricter Bell-style test would require a single model that exposes all four setting-conditioned outcomes; extending the present diagnostic in that direction (and to sequential or reinforcement-learning settings) is left for future work.

Third, the construction and interpretation of S remain quantum-inspired diagnostics: they borrow CHSH-type inequalities to characterize correlational structure during training, but are not yet connected to physical quantum circuits or quantum learning systems. Extending this probe to quantum circuits and quantum–classical hybrid models is a natural next step.

AI USE STATEMENT

In the preparation of this manuscript, the authors used GPT-4 for language polishing and grammar correction, for manuscript checking, and for assisting with the organization and clarification of the logical structure of the paper.

All content generated or assisted by the LLM has been thoroughly reviewed, verified, and edited by the human authors. The authors take full responsibility for the accuracy, originality, and integrity of all content in this submission.

ETHICS STATEMENT

Potential misinterpretation of the CHSH terminology used in this paper—in particular, reading $S > 2$ as physical quantum nonlocality—is discussed in Appendix G.

REPRODUCIBILITY STATEMENT

We provide the materials needed to reproduce our results in the appendix. Synthetic data-generation procedures are detailed in Appendix B; training hyperparameters and optimization settings in Appendix C; and additional empirical analyses (gradient–S regimes, capacity sweeps across data types, and real-data protocol notes) in Appendices E, F, and D.2. Theoretical derivations supporting the mechanistic explanation appear in Appendices A.1–A.3.

REFERENCES

Samson Abramsky and Adam Brandenburger. The sheaf-theoretic structure of non-locality and contextuality. *New Journal of Physics*, 13(11):113036, 2011.

Borja Aizpurua, Sukhbinder Singh, and Román Orús. Classical neural networks on quantum devices via tensor network disentanglers: A case study in image classification. *Physical Review Research*, 8(2):023030, 2026.

Rui Soares Barbosa, Tom Douce, Pierre-Emmanuel Emeriau, Elham Kashefi, and Shane Mansfield. Continuous-variable nonlocality and contextuality. *Communications in Mathematical Physics*, 391 (3):1047–1089, 2022.

John S Bell. On the einstein podolsky rosen paradox. *Physics Physique Fizika*, 1(3):195, 1964.

Nicolas Brunner, Daniel Cavalcanti, Stefano Pironio, Valerio Scarani, and Stephanie Wehner. Bell nonlocality. *Reviews of modern physics*, 86(2):419–478, 2014.

Boris S Cirel'son. Quantum generalizations of bell's inequality. *Letters in Mathematical Physics*, 4 (2):93–100, 1980.

John F Clauser and Abner Shimony. Bell's theorem. experimental tests and implications. *Reports on Progress in Physics*, 41(12):1881, 1978.

John F Clauser, Michael A Horne, Abner Shimony, and Richard A Holt. Proposed experiment to test local hidden-variable theories. *Physical review letters*, 23(15):880, 1969.

Adenilton J da Silva and Wilson R de Oliveira. Comments on "quantum artificial neural networks with applications". *Information Sciences*, 370–371:120–122, 2016.

Shlomi Debi and Adi Makmal. Variational quantum recommendation system with embedded latent vectors. *Scientific Reports*, 15(1):32907, 2025.

Albert Einstein, Boris Podolsky, and Nathan Rosen. Can quantum-mechanical description of physical reality be considered complete? *Physical review*, 47(10):777, 1935.

Patrick Emonts, Mengyao Hu, Albert Aloy, and Jordi Tura. Effects of topological boundary conditions on bell nonlocality. *Physical Review A*, 110(3):032201, 2024.

Arthur Fine. Joint distributions, quantum correlations, and commuting observables. *Journal of Mathematical Physics*, 23(7):1306–1310, 1982.

Valentin Gebhart, Lukas M. Müller, and Jan Sperling. Identifying nonclassicality from experimental data using artificial neural networks. *arXiv preprint arXiv:2101.07112*, 2021.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Koji Hashimoto, Yuji Hirono, Jun Maeda, and Jojiro Totsuka-Yoshinaka. Neural network representation of quantum systems. *Machine Learning: Science and Technology*, 5(4):045039, 2024.

Subhash Kak. Quantum mechanics and artificial intelligence. In *Intelligent Computing Everywhere*, pp. 81–101. Springer, 2007.

Tamás Kriváchy, Yu Cai, Daniel Cavalcanti, Arash Tavakoli, Nicolas Gisin, and Nicolas Brunner. A neural network oracle for quantum nonlocality problems in networks. *npj Quantum Information*, 6 (1):70, 2020.

Ya-Qi Ling, Jun-Hu Zhang, Li-Hua Zhang, Yan-Ran Li, and Hui-Lei Huang. Image classification using hybrid classical-quantum neutral networks. *International Journal of Theoretical Physics*, 63 (5):125, 2024.

Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. Tensorizing neural networks. *Advances in neural information processing systems*, 28, 2015.

Arpan Phukan and Asif Ekbal. A survey of quantum natural language processing: From compositional models to nisq-era empiricism. In *Proceedings of the QuantumNLP {:} Integrating Quantum Computing with Natural Language Processing*, pp. 65–75, 2025.

Maria Schuld and Francesco Petruccione. *Machine learning with quantum computers*, volume 676. Springer, 2021.

Edwin Stoudenmire and David J Schwab. Supervised learning with tensor networks. *Advances in neural information processing systems*, 29, 2016.

Ewin Tang. A quantum-inspired classical algorithm for recommendation systems. In *Proceedings of the 51st annual ACM SIGACT symposium on theory of computing*, pp. 217–228, 2019.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. *arXiv preprint arXiv:2010.05874*, 2020.

Jan Wasilewski, Tomasz Paterek, and Karol Horodecki. Uncertainty of feed forward neural networks recognizing quantum contextuality. *Journal of Physics A: Mathematical and Theoretical*, 56(45): 455305, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Jun-Jie Zhang, Dong-Xiao Zhang, Jian-Nan Chen, Long-Gang Pang, and Deyu Meng. On the uncertainty principle of neural networks. *iScience*, 28(4), 2025.

Peng Zhang, Hui Gao, Jing Zhang, Dawei Song, et al. Quantum-inspired neural language representation, matching and understanding. *Foundations and Trends® in Information Retrieval*, 16(4-5): 318–509, 2023.

Ying Zhang, Jin-chuan Hou, and Kan He. Detecting nonlocal correlations in chain-shaped quantum networks via a deep-learning method. *Physical Review A*, 110(6):062609, 2024.

Hanyu Zhao, Yang Wu, and Yuexian Hou. On emergences of non-classical statistical characteristics in classical neural networks. *arXiv preprint arXiv:2603.04451*, 2026.

A Supplementary Material for Theoretical Framework

A.1 Derivation of Neural Correlation Dynamics

This appendix records how the phenomenological coupling model in Section 5 is motivated from gradient flow. It is *not* a uniqueness theorem: the ODE is a low-dimensional summary of an exact but intractable identity.

A.1.1 Microscopic Parameter Dynamics

Consider N task losses $\{\ell_p\}$ sharing parameters θ (in NCnet, $N = 2$ heads within a trained pair). Under continuous-time gradient flow,

$$\frac{d\theta}{dt} = -\sum_{p=1}^{N} \lambda_p \nabla_\theta \ell_p, \tag{10}$$

with task weights $\lambda_p > 0$. For hidden units h_k, h_l, the Pearson correlation $\rho_{kl} = \rho_{kl}(\theta)$ evolves by the chain rule:

$$\frac{d\rho_{kl}}{dt} = \sum_m \frac{\partial \rho_{kl}}{\partial \theta_m} \frac{d\theta_m}{dt} = -\sum_{p=1}^{N} \lambda_p \sum_m \frac{\partial \rho_{kl}}{\partial \theta_m} \frac{\partial \ell_p}{\partial \theta_m}. \tag{11}$$

Each summand measures whether minimizing ℓ_p increases or decreases ρ_{kl}. Partitioning coordinates into those that increase ρ_{kl} under task p and those that decrease it yields an exact decomposition into driving and decaying contributions, but the partition itself depends on θ and on the data.

A.1.2 Phenomenological Reduction

Two qualitative regularities suggest a scalar summary. First, when head gradients on shared weights are strongly opposed—large $\bar{\mathcal{G}}_{kl} = \frac{1}{2}(\mathcal{G}_k + \mathcal{G}_l)$ in Eq. (6)—the driving contributions that raise ρ_{kl} tend to dominate. Second, as $|\rho_{kl}| \to 1$, further increases become geometrically harder, which we encode by a factor $(1 - \rho_{kl}^2)$; regularization and specialization pressures that pull representations apart are summarized by a linear decay $-\beta\rho_{kl}$. These considerations motivate

$$\frac{d\rho_{kl}}{dt} = \alpha \bar{\mathcal{G}}_{kl}(1 - \rho_{kl}^2) - \beta\rho_{kl}, \tag{12}$$

with $\alpha, \beta > 0$. Setting the right-hand side to zero and taking the root in $(0, 1)$ gives

$$\rho_{kl}^*(\bar{\mathcal{G}}_{kl}) = \frac{-\beta + \sqrt{\beta^2 + 4\alpha^2 \bar{\mathcal{G}}_{kl}^2}}{2\alpha \bar{\mathcal{G}}_{kl}} \quad (\bar{\mathcal{G}}_{kl} > 0), \tag{13}$$

with the convention $\rho_{kl}^*(0) = 0$. The map $\bar{\mathcal{G}}_{kl} \mapsto \rho_{kl}^*$ is strictly increasing and saturates at 1. We emphasize that Eq. (12) is a phenomenological model consistent with Eq. (11), not an identity derived without additional assumptions. In particular, it describes the *coupling* branch of the story; the collapse of easy-pair accuracy under destructive competition is handled by $\eta(\xi)$ in Appendix A.3.

A.2 Feedforward Explanation of the Correctness Gap

Section 5 explains how coupling unbalances the four correctness correlations under standard feedforward evaluation; Section 5 turns that gap into $S > 2$. After training on a pair (α_i, β_j), A_i and B_j are correctness indicators on a forward pass. When the hidden units serving the two heads are strongly correlated ($\rho_{kl} \approx 1$), they co-activate and the shared representation collapses to an approximately one-dimensional mode $z \propto h_k + h_l$ rather than two independently usable features. Easier pairs that include an identity target can still read a simple input bit from this common z, so both heads are correct on most inputs and $C(A_i, B_j) \approx 1$. The hard pair (α_2, β_2) requires two independent XOR features on disjoint inputs; a rank-constrained coupled code cannot realize both maps simultaneously, so at least one head errs on a non-negligible fraction of inputs and $C(A_2, B_2)$ drops. The gap $\mu_{\text{eff}} = C_+ - C_{22}$ in Section 5 is exactly this easy/hard discrepancy. No test-time latent update is involved.

A.3 Formal Characterization of the Emergence Window

We develop a dual-factor formalization of the emergence-window remark at the end of Section 5.

A.3.1 Capacity Stress and Dual Factors

Let $\xi = \mathcal{C}_{\text{task}}/\mathcal{C}_{\text{network}}$. For basic NCnet, a convenient proxy is $\xi \approx 4/d_{\text{eff}}$ with d_{eff} an effective hidden width, but the argument only needs that larger width decreases ξ. Recalling $\mu_{\text{eff}} = C_+ - C_{22}$ from Section 5, we write

$$\mu_{\text{eff}}(\xi) = \mu_{\text{int}}(\xi)\,\eta(\xi), \tag{14}$$

as a phenomenological factorization. Here $\mu_{\text{int}}(\xi) \geq 0$ measures how strongly a bottleneck *opens* the easy/hard gap through coupling (Step 2–3): it is nondecreasing in ξ on the regime where easier pairs remain solvable. The factor $\eta(\xi) \in [0, 1]$ measures residual easy-pair solvability: it is nonincreasing in ξ, and also shrinks when competition becomes destructive (large unstable $\mathcal{G}$) so that C_+ itself collapses. Thus η is the term that reconciles "more competition $\Rightarrow$ more coupling" with the empirical observation that *excessive* $\mathcal{G}$ yields small S.

A.3.2 Boundary Vanishing

Lemma 1 (Boundary Vanishing)**.** *Assume μ_{int} is bounded and $\eta(\xi) \to 0$ as $\xi \to \infty$, while $\mu_{\text{int}}(\xi) \to 0$ as $\xi \to 0$. Then $\mu_{\text{eff}}(\xi) \to 0$ at both extremes.*

Proof. As $\xi \to 0$ (abundant capacity), heads can specialize, $\bar{\mathcal{G}} \to 0$, $\rho^* \to 0$, and $C_{22} \to C_+$, so $\mu_{\text{int}} \to 0$ and hence $\mu_{\text{eff}} \to 0$. As $\xi \to \infty$ (severe bottleneck), even easy pairs fail, so C_+ falls with C_{22}; equivalently $\eta \to 0$, and the product vanishes regardless of how large μ_{int} becomes. □

A.3.3 Existence of an Intermediate Window

Theorem 1 (Emergence Window)**.** *Assume μ_{eff} is continuous on $(0, \infty)$, satisfies Lemma 1, and is* unimodal*: there is ξ^* such that μ_{eff} is nondecreasing on $(0, \xi^*]$ and nonincreasing on $[\xi^*, \infty)$, with $\mu_{\text{eff}}(\xi^*) > 0$. Then for every $\varepsilon \in \big(0, \mu_{\text{eff}}(\xi^*)\big)$ there exist $0 < \xi_{\text{low}}(\varepsilon) < \xi^* < \xi_{\text{high}}(\varepsilon) < \infty$ such that $\mu_{\text{eff}}(\xi) > \varepsilon$ if and only if $\xi \in \big(\xi_{\text{low}}(\varepsilon), \xi_{\text{high}}(\varepsilon)\big)$. Under the working regime $C_+ \approx 1$, this is exactly the set of ξ with $S > 2 + \varepsilon$.*

Proof. Unimodality and continuity imply that each strict superlevel set of a unimodal function with limits 0 at both ends is a single bounded open interval containing ξ^*. The identification with $\{S > 2 + \varepsilon\}$ follows from $S = 2C_+ + \mu_{\text{eff}}$ when $C_+ \approx 1$. Continuity of μ_{eff} in effective width is a modeling assumption, empirically supported by the smooth inverted-U capacity curves in Section 6.2. □

Interpretation. Below ξ_{low}, coupling is too weak and C_{22} matches C_+. Above ξ_{high}, capacity (or stability) is insufficient to keep C_+ near 1, so the gap closes from above. Only in the intermediate window is coupling strong enough to depress the hard pair while leaving easier pairs intact. The same logic explains why S tracks a moderate-$\mathcal{G}$ ridge rather than maximizing $\mathcal{G}$: large $\mathcal{G}$ raises μ_{int} only while η remains high.

B Details of CHSH Data Generation

This appendix details the three synthetic input generators used in Section 4.1. The types increase structural complexity from deterministic maps to control-feature-dependent and more nonlinear mechanisms, enabling a controlled comparison of when CHSH-style $S > 2$ appears.

Notation. Each sample consists of binary variables $x_i \in \{0, 1\}$. The network outputs two targets (t_1, t_2) for a chosen pair setting $(a, b) \in \{1, 2\}^2$, matching the Alice–Bob heads in the NCnet protocol. Indicator functions are denoted by $\mathbb{I}(\cdot)$.

TYPE 1: DETERMINISTIC MAPPING

Algorithm 1: CHSH Data Generation for Type 1

Require: measurement settings (a, b), random seed s
Set random seed s
Initialize x as all 4-bit binary patterns (repeated)
for each sample **do**
  if $a = 1$ **then**
    $t_1 \leftarrow x_0$
  else
    $t_1 \leftarrow (x_0 \neq x_1)$
  end if
  if $b = 1$ **then**
    $t_2 \leftarrow x_2$
  else
    $t_2 \leftarrow (x_2 \neq x_3)$
  end if
end for
return x, (t_1, t_2)

This data type is a purely deterministic baseline (identity and XOR maps) with no additional control features. It is the lowest-complexity generator in the suite.

TYPE 2: CONTROL-FEATURE MAPPING

Algorithm 2: CHSH Data Generation for Type 2

Require: measurement settings (a, b), random seed s
Set random seed s
Initialize x_{base} as all 5-bit binary patterns (repeated)
for each sample **do**
  Sample $r \sim \text{Uniform}(0, 1)$
  if $r \leq 0.8$ **then**
    $(c_1, c_2) \leftarrow (0, 0)$
  else if $r \leq 0.9$ **then**
    $(c_1, c_2) \leftarrow (-2, 2)$
  else
    $(c_1, c_2) \leftarrow (2, -2)$
  end if
  if $a = 1$ **then**
    if $|c_1| \leq 1$ **or** $c_1 < -1$ **then**
      $t_1 \leftarrow x_0$
    else
      $t_1 \leftarrow 1 - x_0$
    end if
  else
    if $|c_2| \leq 1$ **or** $c_2 < -1$ **then**
      $t_1 \leftarrow (x_1 \neq x_2)$
    else
      $t_1 \leftarrow 1 - (x_1 \neq x_2)$
    end if
  end if
  if $b = 1$ **then**
    if $|c_1| \leq 1$ **or** $c_1 > 1$ **then**
      $t_2 \leftarrow 1 - x_0$
    else

```
        t_2 ← x_0
    end if
  else
    if |c_2| ≤ 1 or c_2 > 1 then
        t_2 ← 1 − (x_3 ≠ x_4)
    else
        t_2 ← (x_3 ≠ x_4)
    end if
  end if
end for
Concatenate (c_1, c_2) with x_base to form x
return x, (t_1, t_2)
```

This data type adds control features (c_1, c_2) that conditionally flip or rewrite the targets. The resulting maps are harder than Type 1 and raise capacity stress, but evaluation remains ordinary feedforward prediction: $S > 2$ is attributed to learning dynamics under a shared bottleneck (Section 5), not to test-time latent adaptation.

TYPE 3: NONLINEAR CONTROL-FEATURE MAPPING

Algorithm 3: CHSH Data Generation for Type 3

Require: measurement settings (a, b), random seed s

```
Set random seed s
Initialize x_base as all 6-bit binary patterns (repeated)
for each sample do
  Sample r ~ Uniform(0, 1)
  if r ≤ 0.6 then
    (c_1, c_2) ← (0, 0)
  else if r ≤ 0.8 then
    (c_1, c_2) ← (−1, 1)
  else
    (c_1, c_2) ← (1, −1)
  end if
  if a = 1 then
    if c_1 = 0 then
      t_1 ← x_0
    else if c_1 > 0 then
      t_1 ← x_0 · x_1
    else
      t_1 ← I(x_2 + x_1 > 0)
    end if
  else
    if c_2 = 0 then
      t_1 ← x_1
    else if c_2 > 0 then
      t_1 ← 1 − (x_2 ≠ x_1)
    else
      t_1 ← (x_2 ≠ x_3)
    end if
  end if
  if b = 1 then
    if c_1 = 0 then
      t_2 ← x_2
    else if c_1 > 0 then
      t_2 ← x_2 · x_3
```

```
        else
            t_2 ← 𝕀(x_4 + x_3 > 0)
        end if
    else
        if c_2 = 0 then
            t_2 ← x_3
        else if c_2 > 0 then
            t_2 ← 1 − (x_4 ≠ x_3)
        else
            t_2 ← (x_4 ≠ x_5)
        end if
    end if
end for
Concatenate (c_1, c_2) with x_base to form x
return x, (t_1, t_2)
```

This data type further introduces nonlinear transformations (e.g., multiplicative interactions and thresholds), increasing task complexity relative to Types 1–2 and shifting the capacity regime in which $S > 2$ tends to appear.

C Training Hyperparameters for CHSH Experiments

All CHSH experiments were conducted with single-hidden-layer neural networks. The main hyperparameters used for training are summarized in Table 5.

Table 5: Training hyperparameters for CHSH experiments.

Hyperparameter	**Value**
Number of output units	2
Number of epochs	60
Number of random seeds	20
Learning rate	0.01
Optimizer	Adam
Activation function	ReLU
Loss function	Mean Squared Error
Hidden layer sizes	3 (Type 1), 4 (Type 2), 15 (Type 3)
Batch size	Full-batch (all samples used at each step)

Notes:

- Each seed produces a full training trajectory of CHSH values across epochs.
- The hidden layer size was chosen based on preliminary scans to optimize the emergence of CHSH violation.
- Full-batch training ensures deterministic computation of S_{max} and S_{fin} per seed.

D Additional Experimental Results

D.1 Training Dynamics for Data Types 1 and 3

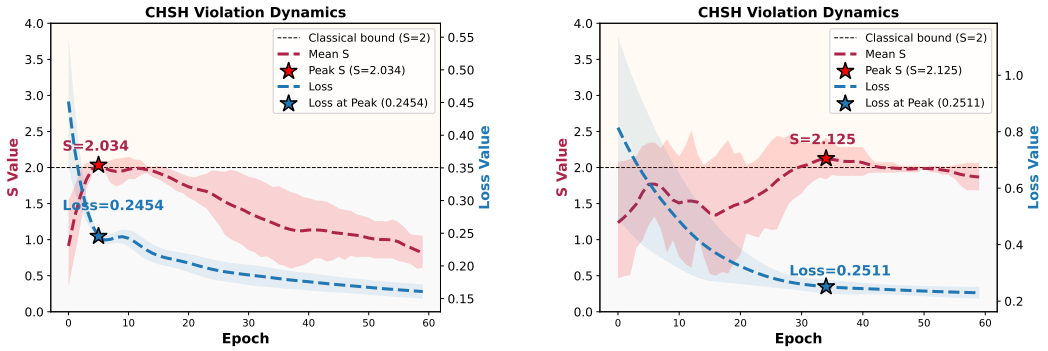
In addition to Data Type 2 in Section 4, Figure 5 reports training dynamics of CHSH-style S and loss for Data Types 1 and 3.

Overall, both data types exhibit qualitatively similar dynamics: the CHSH-style statistic S rises early in training, can temporarily exceed 2, and subsequently decreases toward or below 2 as optimization proceeds, while training loss decreases monotonically.

However, the strength and persistence of violations differ across data types. For Data Type 1, the excess is relatively weak ($S_{\max} \approx 2.131$ in Table 1) and confined to a short $S > 2$ duration, consistent with its lower violation rate.

For Data Type 3, the mean peak is $S_{\max} \approx 2.053$. The excess remains transient, with a shorter $S > 2$ duration than Data Type 2.

The mismatch between peak height and temporal persistence across types suggests that these two aspects need not move in lockstep, and that both capacity and data structure matter for when $S > 2$ appears.



(a) Data Type 1 ($S_{\max} \approx 2.131$). (b) Data Type 3 ($S_{\max} \approx 2.053$).

Figure 5: Training dynamics of CHSH-style S and training loss for Data Types 1 and 3. In both cases, S may temporarily exceed 2 during training, but the excess is not sustained and typically decays to $S \leq 2$.

D.2 Real-World Protocol Details

The COCO and AAPD results in Tables 2 and 3 (Section 4) use the same CHSH-style multi-task protocol as the synthetic experiments (Section 4.1): each pair setting (α_i, β_j) is trained as a separate two-head problem, correctness indicators define C_{ij}, and S is assembled as in Section 3.2. Training follows Appendix C.

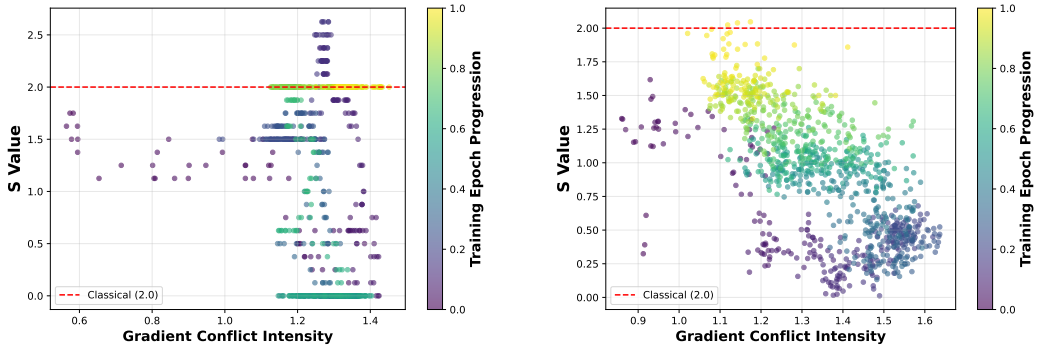
On AAPD (multi-label abstracts) and MS COCO (multi-label images), the four NCnet targets are realized by mapping $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ to concrete category/object labels. We use three matched groupings of increasing structural complexity; these are the Type 1–3 rows in the COCO/AAPD tables and are *not* the synthetic generators of Appendix B. Table 6 lists the mappings. For each Type, a sample's two-head targets indicate whether the corresponding labels are present; the four pair settings are trained separately as in Section 3.2.

Table 6: CHSH-style setting maps for AAPD categories and COCO object classes (Types 1–3).

Type	AAPD				COCO			
	α_1	α_2	β_1	β_2	α_1	α_2	β_1	β_2
1	`cs.AR`	`cs.CV`	`cs.NA`	`cs.SE`	Person	Bicycle	Car	Motorcycle
2	`cs.AI`	`cs.DM`	`cs.LG`	`stat.ML`	Laptop	Cell phone	Book	Clock
3	`cs.CC`	`cs.OH`	`cs.IR`	`math.AP`	Dog	Cat	Bird	Horse

E Additional Analysis of Gradient Competition

This appendix complements Section 6.1 with Data Types 1 and 3. Relative to the moderate-competition ridge on Data Type 2, Type 1 illustrates a weak-competition boundary and Type 3 a destructive-competition boundary; together they sketch insufficient interaction, an intermediate high-S regime, and destructive interference.



(a) Data Type 1: weak-competition boundary

(b) Data Type 3: destructive-competition boundary

Figure 6: Complementary $\mathcal{G}$–S regimes for Data Types 1 and 3. Left: weak competition yields only sparse, unstable $S > 2$. Right: competition remains substantial but S collapses, consistent with destructive interference.

E.1 Data Type 1: Weak-Competition Boundary

Figure 6(a) presents the optimization trajectory for Data Type 1. Compared with the main-text configuration, gradient interaction is substantially weaker.

The trajectory concentrates mainly in relatively low-competition regions ($\mathcal{G} < 1.3$). Occasional epochs briefly produce $S > 2$, but these states remain sparse and unstable; most runs settle near classical values around $S \approx 2$.

This indicates that weak head competition is insufficient to sustain the coupling needed for a stable easy/hard correlational gap. Heads can largely succeed independently, so the shared representation stays closer to a factorized solution.

E.2 Data Type 3: Destructive-Competition Boundary

Figure 6(b) shows Data Type 3. Competition remains non-negligible (often $\mathcal{G} \approx 1.3 \sim 1.6$), overlapping numerically with parts of the Type 2 cloud, yet S collapses toward much lower values, with many epochs below $S = 1$. Absolute $\mathcal{G}$ alone is therefore not a universal score across data types: what matters is whether moderate conflict coexists with usable easier-pair accuracy.

Early training is highly unstable, with large fluctuations alongside strong conflict. As optimization proceeds, the trajectory settles into low-S regions despite persistently large $\mathcal{G}$, consistent with destructive interference that destabilizes the shared code required for $S > 2$.

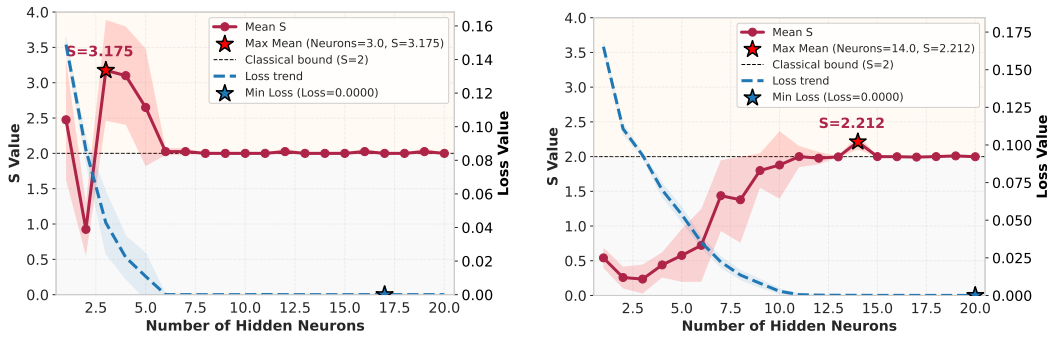
Taken together, the three data types sketch:

- a **weak-competition** boundary with only transient $S > 2$ (Data Type 1),
- an **intermediate** regime in which moderate $\mathcal{G}$ co-occurs with elevated S (Data Type 2 in Section 6.1),
- and a **destructive-competition** boundary where substantial $\mathcal{G}$ accompanies collapsed S (Data Type 3).

This pattern supports the claim in Section 5 that $S > 2$ requires not maximal competition, but intermediate conflict while easier pairs remain solvable.

F Effect of Task Complexity on the Emergence Window

In this section, we further analyze how task complexity affects the emergence of non-classical correlations under capacity constraints. We complement the main text (which focuses on Data Type 2)



(a) Data Type 1 (Low Complexity)

(b) Data Type 3 (High Complexity)

Figure 7: CHSH maximum value $S_{\max}$ as a function of hidden layer size for different task complexities. Both configurations exhibit a non-monotonic emergence pattern, but with systematically shifted emergence windows.

by reporting results for Data Type 1 and Data Type 3, representing lower and higher task complexity regimes, respectively.

Experimental Setup We follow the same protocol as Section 6.2: vary hidden width in a single-hidden-layer ReLU network with other hyperparameters fixed, and record $S_{\max}$ over training, averaged across 20 seeds. Data Type 1 is lower complexity; Data Type 3 is higher complexity (Appendix B).

Results Figure 7 shows the relationship between network capacity and CHSH-style $S_{\max}$ under different task complexities.

Across both settings we observe a non-monotonic pattern: $S > 2$ appears only inside a finite intermediate width regime, but the location of that window shifts with complexity.

For Data Type 1 the peak occurs near 3 hidden neurons; for Data Type 3 it shifts to about 14. Although both settings exhibit $S > 2$, the peak $S_{\max}$ decreases as complexity increases (2.131 vs. 2.053), suggesting that harder tasks require more width to enter the coupling regime while also making a large easy/hard correlational gap harder to sustain.

Mechanistic Interpretation These results support the capacity-stress explanation in Section 5: the window shifts with ξ, consistent with the dual-factor form $\mu_{\text{eff}} = \mu_{\text{int}}\eta$ in Appendix A.3.

As task complexity increases, $\mathcal{C}_{\text{task}}$ grows, so a larger network capacity is needed before bottleneck competition is strong enough to induce coupling and depress the hard pair. At the same time, harder tasks spread gradient conflict across more directions, which can reduce the peak gap $C_+ - C_{22}$ even inside the window. Thus the emergence window is set by the relative scaling $\xi = \mathcal{C}_{\text{task}}/\mathcal{C}_{\text{network}}$, not by absolute width alone.

G Broader Impact

This work studies the emergence of non-classical statistical correlations in classical neural networks, as quantified by violations of the CHSH inequality. We emphasize that our results are purely theoretical and empirical in nature, and do not imply the presence of physical quantum effects in machine learning systems.

From a positive perspective, this work provides a lens on representation learning under capacity constraints. Interpreting CHSH-style excesses as a diagnostic of multi-task interference and correlational imbalance offers a compact way to track how optimization under shared bottlenecks shapes statistical dependence among task heads. This may help analyze gradient interference, representation collapse, and capacity-driven non-monotonic regimes in multi-task networks.

More broadly, the notion of an "emergence window" highlights how complex statistical behaviors can arise only under intermediate regimes of model capacity. This perspective may be useful for studying other phenomena in deep learning where performance, generalization, and internal representation structure are governed by non-monotonic interactions between data complexity and model scale.

At the same time, we acknowledge potential risks of misinterpretation. In particular, the use of terminology such as "non-classical correlations" or "CHSH violation" may be incorrectly interpreted as evidence of quantum physical behavior in artificial neural networks. We stress that all such terms are used in a strictly mathematical and informational sense, referring to violations of classical probabilistic factorization assumptions rather than any physical nonlocality.

Finally, we do not foresee direct negative societal impacts arising from this work. However, as with many theoretical advances in deep learning, the framework could be misapplied if its scope is not properly understood. We therefore encourage careful interpretation of our results within the context of statistical learning theory and representation dynamics.