TRANSFORMER BLOCK COUPLING & its Correlation with Generalization in LLMs

Murdock Aubry*, Haoming Meng*, Anton Sugolov*, Vardan Papyan

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY

What is Coupling?

An empirical phenomenon where transformer blocks adopt a common basis and perform similar computations in coordination across depth and tokens.

Mathematical Formulation

- **Token embedding:** $x_i^l \in \mathbb{R}^d$ at layer l.

$$X^{l+1} = F_{\text{block}}^{l+1}(X^l) = X^l + f^{l+1}(X^l).$$

- **Block Jacobians:** The linearization at layer l.

$$J_{t_1 t_2}^l = \frac{\partial}{\partial x_{t_1}^{l-1}} \left(f^l(X^{l-1})\right)_{t_2} \in \mathbb{R}^{d\times d}.$$

- **Coupling:** Given Jacobians J_1, J_2, we compute their singular value decompositions

$$J_1 = U_1 S_1 V_1^T \qquad J_2 = U_2 S_2 V_2^T,$$

and quantify **coupling** of their top-K singular vectors using

$$m_K(J_1, J_2) = \frac{||U_{2,K}^T J_1 V_{2,K} - S_{1,K}||_F}{\|s_{1,K}\|_p} = \frac{||U_{2,K}^T U_1 S_1 V_1^T V_{2,K} - S_{1,K}||_F}{\|s_{1,K}\|_p}.$$

Measures how strongly the top-K singular vectors are aligned (diagonalizing J_1 with the top-K singular vectors of J_2).

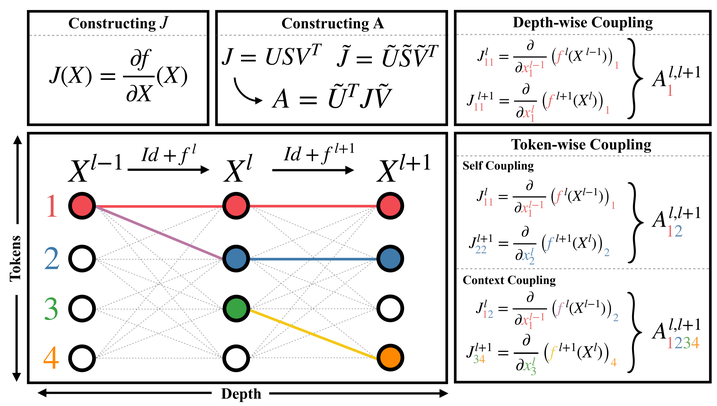
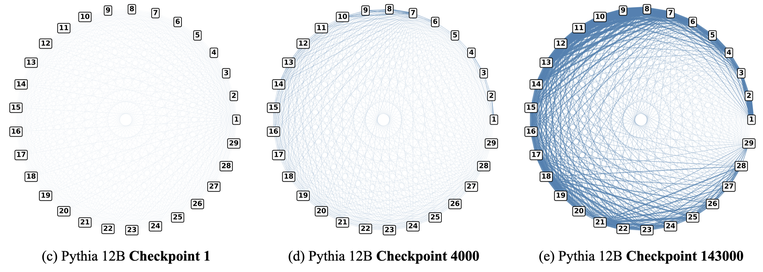


Figure 1. Measuring coupling through multiple token interactions in the transformer blocks.

Emergence of Coupling with Training



(c) Pythia 12B **Checkpoint 1** (d) Pythia 12B **Checkpoint 4000** (e) Pythia 12B **Checkpoint 143000**

Figure 2. Increased coupling of transformer blocks in Pythia 12B during training.

Correlation with Generalization

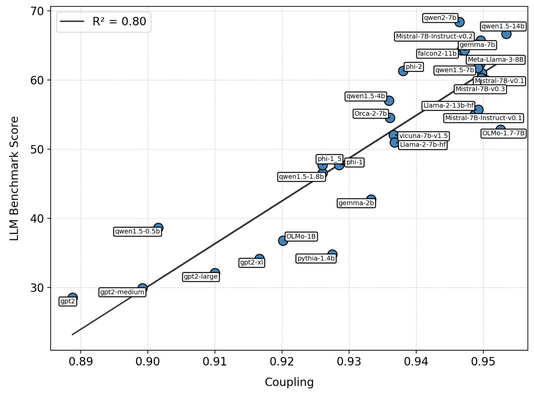
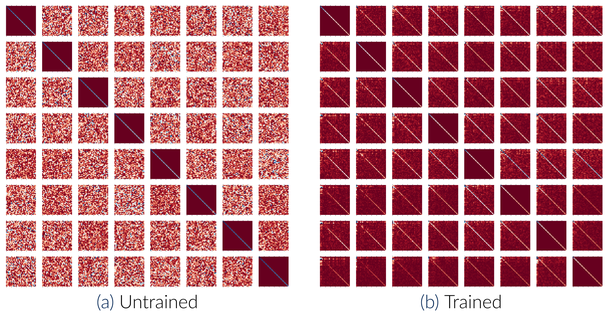


Figure 3. HuggingFace Open LLM leaderboard score plotted against coupling on dataset prompts.

Depth-wise coupling

For layers $1 \le l_1, l_2 \le L$, we plot A^{l_1,l_2}, observing **coupling across depth** of block Jacobians J^{l_1}, J^{l_2} operating on a fixed token embedding ($t_1 = t_2$).

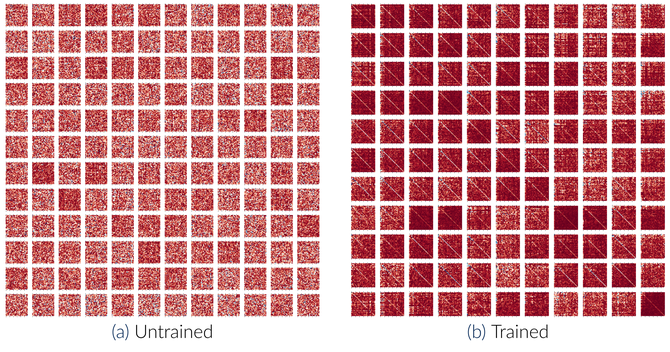


(a) Untrained (b) Trained

Figure 4. Llama-3-8B, coupling of the last token across layers 9 to 16.

Token-wise coupling

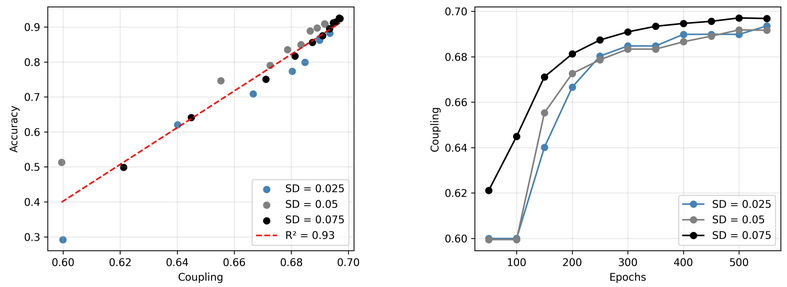
For the various types of token-wise coupling, we plot $A^{l_1,l_2}_{t_1,t_2}$ across tokens t_1, t_2, observing **coupling across tokens** of block Jacobians J_{t_1}, J_{t_2}, at fixed layers l_1 and l_2.



(a) Untrained (b) Trained

Figure 5. Llama-3-8B, context coupling with fixed output token across pairs of prompt input tokens, fixing two random layers.

Coupling in Vision Transformers



(a) Correlation with Generalization (b) Emergence with Training

Figure 6. Coupling in ViTs with varied stochastic depth settings.

Properties of Embedding Trajectories

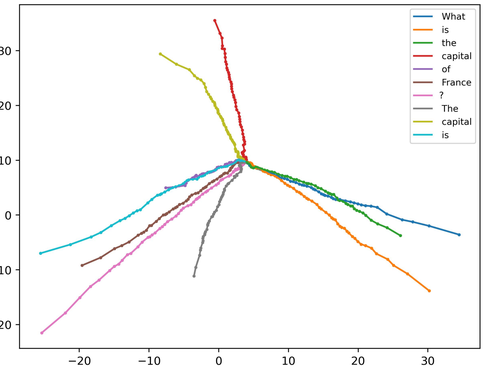


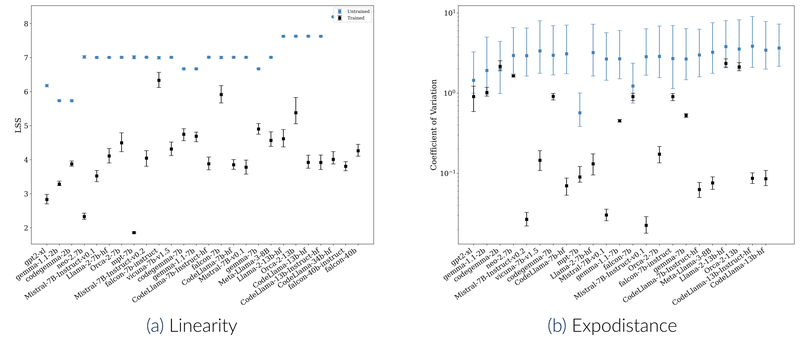
Figure 7. Llama-3-70B, PCA of token embedding trajectories with LLM layer depth.

- **Linearity** (Line-Shape Score): $\text{LSS}_i^{0,\dots,L} = \frac{L}{||\tilde{x}_i^L - \tilde{x}_i^0||_2}$ where

$$\tilde{x}_i^l = \tilde{x}_i^{l-1} + \frac{x_i^l - x_i^{l-1}}{\left|\left|x_i^l - x_i^{l-1}\right|\right|_2} \text{ for } l = 1, \dots, L$$

.

- **Exponential Growth** (expodistance): $\text{ED}_i = \frac{\text{Var}_l \alpha_i^l}{(\text{Avg}_l \alpha_i^l)^2}$ where $\alpha_i^l = \ln\left(\frac{\|x_i^l\|}{\|x_i^{l-1}\|}\right)$



(a) Linearity (b) Expodistance

Figure 8. Embedding trajectory metrics at initialization and after training.