Depth, Context, and Robustness in In-Context Regression

Murat Enes Erdoğan

ELEC/COMP 450-550 (Fall 2025): Introduction to Modern Learning Theory (Instructor: Asst. Prof. Zafer Doğan)

1. Problem Context

Motivation: Understanding how transformer architecture parameters (width, depth, context length) influence in-context learning (ICL) performance is crucial for efficient model scaling. The paper by Bordelon et al. (2025) addresses this by analyzing when depth provides benefits in ICL.

- This work addresses modern phenomena in neural scaling laws, specifically how architectural choices affect generalization in the overparameterized regime

Setting: Deep linear self-attention models performing in-context linear regression, where the model learns to predict from context without parameter updates.

- When does increasing depth L actually improve ICL, and when is it unnecessary?

2. Methodology & Theory

The paper introduces a "Reduced Gamma Model" that captures the essential behavior of deep linear attention:

$$f(x_*) = \frac{1}{LP} x_*^\top \Gamma \sum_{\ell=0}^{L-1} \left(I - L^{-1}\hat{\Sigma}\Gamma\right)^{\ell} X^\top y$$

Where $\hat{\Sigma} = \frac{1}{P} X^\top X$ is the empirical covariance.

Key Results:

- Result 2 (ISO)If context ratio α = P/D → ∞, depth L=1 achieves minimal loss—depth is unnecessary.
- Result 5 (FS): Models trained on fixed covariance are brittle to distribution shift.
- Result 6 (RRS): Random rotations force Γ = γI (isotropic), learning a general algorithm.

3. Replication Strategy

Tools Used: PyTorch with custom implementation of the Reduced Gamma Model.

Dataset: Synthetic Gaussian data:

- ISO: $x \sim N(0, I), \beta \sim N(0, I), y = (1/\sqrt{D})\beta \cdot x$
- FS: Power−law covariance Σ with eigenvalues $\lambda_k \sim k^{-\nu}$
- RRS: Randomly rotated $\Sigma_c = Q_c \Lambda Q_c^\top$

Scope:

- Replicated Figure 1 (ISO training dynamics for varying α and L)
- Replicated Figure 3(c) (FS brittleness to distribution shift)

4. Replication Results

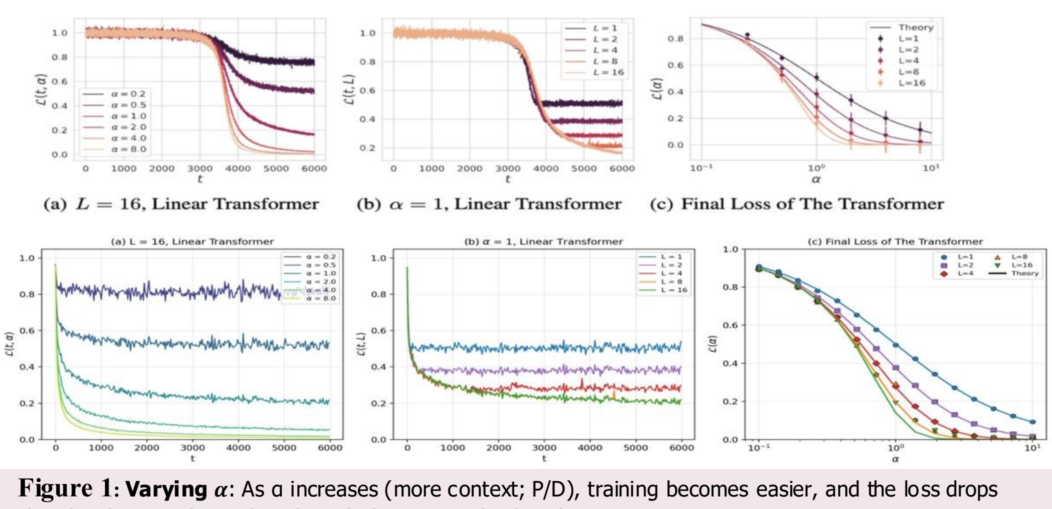

Figure 1: **Varying α**: As α increases (more context; P/D), training becomes easier, and the loss drops sharply, showing diminishing benefit from extra depth at long contexts.
Varying depth L at $\alpha = 1$: When context is moderate ($\alpha \approx 1$), deeper models reach lower loss faster (and slightly better final loss), showing depth helps most in the "not-enough-context-yet" regime.
Final loss vs α: Final loss decreases monotonically with α, and the experimental trend matches theory closely, confirming the paper's scaling-law prediction.

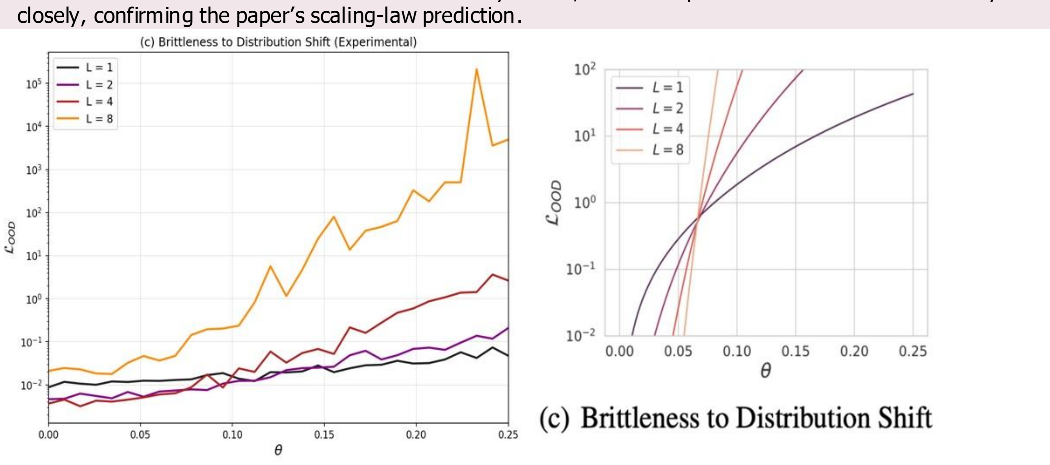

Figure 2: As the shift parameter **θ increases**, **OOD loss rises rapidly**, indicating the FS-trained model is specialized to the training covariance and degrades under distribution shift. Higher depth amplifies brittleness: deeper models show steeper blow-up in OOD loss (more fragile despite similar in-distribution fit).

5. Critical Analysis

Assumptions:

- Linear attention (no softmax nonlinearity), which enables analytical tractability but differs from practical transformers.
- Gaussian i.i.d. data distributions—real data has complex correlations.

Limitations:

- The reduced Γ model assumes aligned weight matrices ($W_x^T w_y = 0$), which may not hold in practice.
- Computational cost still scales with D for the matrix operations.

Discrepancies:

- In Figure 3(c) replication, experimental curves show similar trends but with some variance due to finite-sample effects (D=32 vs. theoretical $D \to \infty$).
- The FS model required initialization near optimal ($\Gamma \approx L\Sigma^{(-1)}$) for reliable convergence.

6. New Results

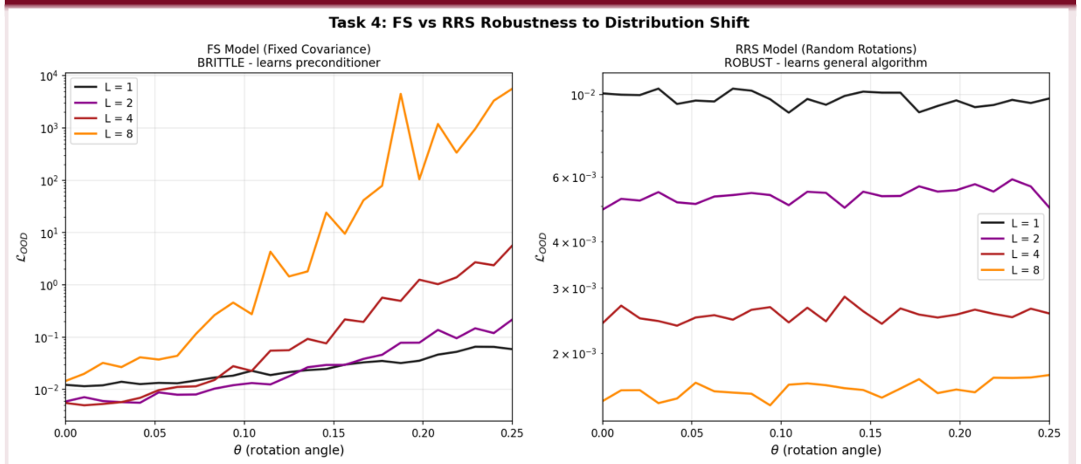

Figure 3: **FS model = brittle**: OOD loss increases strongly with θ, consistent with "memorizing the training geometry". **RRS model = robust**: training across rotated covariances makes the OOD curve **nearly flat**, supporting the idea that RRS encourages a **generic algorithmic solution (GD-like)** rather than covariance memorization.

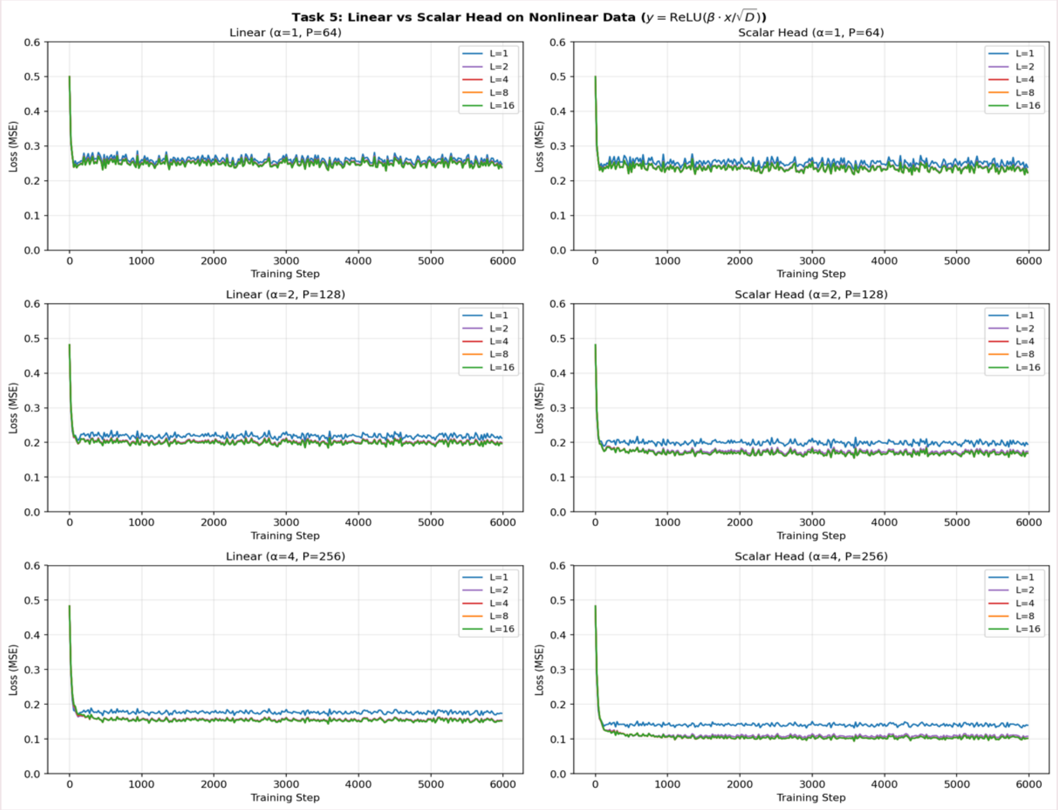

Figure 4: Nonlinear ISO target $y = ReLU(z)$ with $z = \beta \cdot x/\sqrt{D}$. Compare (i) **linear head** $\hat{y} = u + b_0$ (paper head; $u = w_o^\top h^L$) vs (ii) **scalar nonlinear head** $\hat{y} = u + c_1 ReLU(u) + c_2 ReLU(-u) + b$ (strict superset of linear). As $\alpha = P/D$ increases (more context), u becomes a better estimate of z, so the nonlinear head's advantage grows (up to ~ 33% lower MSE at $\alpha = 4$). Depth effects are comparatively modest for this single-index ReLU target.

7. Future Directions

- Test robustness under **non-i.i.d. data distributions** or real-world covariate structures.
- Explore **deeper nonlinear heads** (multi-layer) for the nonlinear ICL extension.

References

- Bordelon, Blake; Letey, Mary; Pehlevan, Cengiz. (2025). *Theory of Scaling Laws for In-Context Regression: Depth, Width, Context and Time. arXiv:2510.01098*
- Bach, F. (2024). *Learning Theory from First Principles.* MIT Press.