Weekly arXiv Picks | Week of June 6, 2026

This week's selection focuses on **Mechanistic Interpretability, Chain-of-Thought Theory, Optimization Dynamics, Edge of Stability, Physics of AI, and World Models**.

---

1. The Shape of Addition: Geometric Structures of Arithmetic in Large Language Models

arXiv: https://arxiv.org/abs/2606.03645

Why do large language models make arithmetic mistakes? This paper argues that addition is not represented as a discrete symbolic algorithm, but rather as a continuous geometric structure in the residual stream. The authors identify *Iso-Raw-Sum Trajectories (IRSTs)* and show that arithmetic errors emerge when these trajectories cross quantization boundaries. The work offers a geometric perspective on arithmetic reasoning and sheds light on why both correct and hallucinated answers can often be decoded from the same internal representations.

---

2. An Asymptotic Theory of Chain-of-Thought in In-Context Learning

arXiv: https://arxiv.org/abs/2606.03217

This paper develops a theoretical framework for understanding Chain-of-Thought (CoT) reasoning. By modeling CoT as an iterative refinement process, the authors derive asymptotic scaling laws relating reasoning depth to generalization performance. Their analysis predicts distinct regimes, including rapid improvement, saturation, and even performance degradation due to overthinking. A promising step toward a theory of reasoning-time scaling.

---

3. Spectral Asymptotics of Neural Network Loss Landscapes: An Exact Decomposition of the Curvature Exponent

arXiv: https://arxiv.org/pdf/2606.02596

The authors investigate the spectral structure of neural network loss landscapes and derive an exact decomposition of the curvature exponent. The work connects Hessian spectra, curvature growth, and optimization geometry, providing new theoretical tools for understanding training dynamics and second-order optimization.

---

4. Edge of Stability Selectively Shapes Learning Across the Data Distribution

arXiv: https://arxiv.org/pdf/2606.04212

Edge of Stability (EoS) has traditionally been viewed as a global optimization phenomenon. This paper shows that EoS can selectively influence different parts of the data distribution, accelerating learning for some subsets while slowing it for others. The results suggest that EoS plays an important role in determining what neural networks learn first.

---

5. Why Muon Outperforms Adam: A Curvature Perspective

arXiv: https://arxiv.org/pdf/2606.04662

Muon has recently emerged as a strong alternative to Adam for large-scale training. This paper explains its success through the lens of curvature. The key finding is that Muon achieves better optimization not simply through larger updates, but through improved handling of high-curvature directions, leading to more efficient and stable learning.

---

6. The Invisible Hand of Physics: When Video Diffusion Models Know More Than They Show

arXiv: https://arxiv.org/pdf/2606.05328

One of my favorite papers this week.

The authors show that video diffusion models often possess significantly more physical knowledge internally than is reflected in their generated videos. Even when outputs violate physical laws, latent representations can encode accurate information about object dynamics and physical constraints. This provides intriguing evidence for the emergence of implicit world models.

---

7. Gradient Descent with Large Step Size Restores Symmetry in Deep Linear Networks with Multi-Pathway

arXiv: https://arxiv.org/pdf/2606.05219

This paper studies gradient descent in deep linear networks with multiple pathways. Surprisingly, large learning rates can restore symmetry that would otherwise be broken under gradient flow. The results highlight how finite learning rates can fundamentally alter the implicit bias of optimization.

---

8. Generative Criticality in Large Language Model Temperature Scaling

arXiv: https://arxiv.org/pdf/2606.06238

The authors connect temperature scaling in LLM generation with ideas from statistical physics. As temperature varies, they observe phenomena analogous to phase transitions, including abrupt changes in semantic organization and susceptibility-like quantities. An interesting addition to the growing literature on the Physics of AI.

---

9. Pretraining Recurrent Networks without Recurrence

arXiv: https://arxiv.org/pdf/2606.06479

Can recurrent networks make a comeback?

This paper proposes a novel pretraining strategy that avoids recurrence during training. Instead, memory representations are learned separately before training the recurrent update rule. The approach sidesteps many of the optimization challenges associated with backpropagation through time.

---

10. Balancing Learning Rates Across Layers: Exact Two-Step Dynamics and Optimal Scaling in Linear Neural Networks

arXiv: https://arxiv.org/pdf/2606.00340

The authors derive exact solutions for two-step gradient descent dynamics in linear neural networks and use them to analyze optimal layer-wise learning rate scaling. The work provides theoretical insights relevant to μP, scaling laws, and deep network optimization.

---

11. Massive Spikes in LLMs are Bias Vectors: Mechanistic Uncovering and Spike-Free Quantization

arXiv: https://arxiv.org/pdf/2606.02288

Large activation spikes are a well-known obstacle for LLM quantization. This paper argues that these spikes correspond to structured bias vectors rather than random outliers. Leveraging this insight, the authors propose a new spike-free quantization method that improves compression while preserving performance.

Top 3 Picks of the Week

🥇 **The Invisible Hand of Physics**
A fascinating result suggesting that video diffusion models may internally possess richer physical world models than their outputs reveal.

🥈 **An Asymptotic Theory of Chain-of-Thought in In-Context Learning**
One of the first attempts to establish a rigorous theory of reasoning-depth scaling.

🥉 **Why Muon Outperforms Adam**
A timely theoretical explanation for one of the most discussed optimizers in large-scale AI training.

Keywords

Mechanistic Interpretability · Chain-of-Thought · Optimization · Edge of Stability · World Models · Physics of AI · Recurrent Networks · Scaling Laws