# 每周 arXiv 好文推荐｜2026-06-06

本周主题：**Mechanistic Interpretability、Chain-of-Thought 理论、优化动力学、Edge of Stability、Physics of AI、World Models**。

---

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

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

LLM 为什么会在简单加法上犯错？这篇工作发现，模型内部并非在执行离散的符号计算，而是在残差流（Residual Stream）中形成一种连续几何结构。作者提出 *Iso-Raw-Sum Trajectory (IRST)*，将加法错误解释为连续表征跨越量化边界导致的"几何滑移（Geometric Slippage）"。这一视角不仅解释了算术错误，也解释了为什么简单 probe 能从同一个激活向量中同时解码出正确答案和幻觉信息。

---

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

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

这篇论文首次系统研究了 CoT 深度（Reasoning Depth）的 Scaling Law。作者构造了一个可精确求解的理论模型，把 CoT 看成对参数估计的迭代修正过程，并推导出了泛化误差关于推理深度的闭式表达。结果显示存在明显的相变现象：随着推理步数增加，性能会经历指数提升、饱和、甚至"过度思考（Overthinking）"退化等不同阶段。

---

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

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

研究神经网络 Hessian 谱结构与曲率指数（Curvature Exponent）的关系。作者提出一种精确分解方法，将 Loss Landscape 的复杂谱行为拆解成多个可解释组成部分，为理解训练动力学和二阶优化提供新的理论工具。

---

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

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

Edge of Stability（EoS）通常被视为整体训练现象，但这篇工作发现它实际上会选择性地影响不同样本子群的学习速度。模型并不是同时学习所有数据，而是在 EoS 机制下优先学习某些方向和子分布。

## 5. Why Muon Outperforms Adam: A Curvature Perspective

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

Muon 是近一年受到广泛关注的新优化器。这篇工作从曲率视角解释其成功原因：Muon 的优势并不主要来自更大的更新步长，而来自对高曲率方向的更好控制，从而能够在保持稳定性的同时提高收敛速度。

---

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

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

本周最有意思的论文之一。

作者发现视频 Diffusion Model 生成的视频虽然经常违反物理规律，但模型内部其实已经学会了许多正确的物理知识。换句话说，模型"知道"的比它最终"表现出来"的更多。这为 World Model 和 Physical Intelligence 提供了新的证据。

---

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

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

研究大步长梯度下降在多路径深线性网络中的行为。与传统 Gradient Flow 的预测不同，大学习率会重新恢复网络中的对称结构，从而改变模型最终学到的表示。

---

## 8. Generative Criticality in Large Language Model Temperature Scaling

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

将 Temperature Scaling 与统计物理中的临界现象联系起来。作者观察到当温度变化时，LLM 的生成行为会出现类似相变（Phase Transition）的现象，为 Physics of AI 提供了新的研究案例。

---

## 9. Pretraining Recurrent Networks without Recurrence

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

这篇工作尝试回答一个经典问题：RNN 是否还能卷土重来？

作者提出一种新的预训练框架，在训练阶段不需要真正展开时间递归，而是先学习记忆表示，再训练递归更新规则，从而避免 BPTT 带来的训练困难。

---

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

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

研究不同层学习率应如何缩放。作者在可解析线性网络中精确求解两步梯度下降动力学，为 Layer-wise Learning Rate、μP 等理论提供了新的分析工具。

---

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

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

LLM 中常见的大幅激活 Spike 一直是量化的难点。这篇工作发现这些 Spike 本质上对应于特殊的 Bias Vector，而非随机异常值。基于这一发现，作者提出新的 Spike-Free Quantization 方法，在保持性能的同时降低量化难度。

---

# 本周 Top 3 推荐

🥇 **The Invisible Hand of Physics**
物理世界模型方向最值得关注的一篇。

🥈 **An Asymptotic Theory of Chain-of-Thought in In-Context Learning**
首次系统建立 CoT 深度的理论框架。

🥉 **Why Muon Outperforms Adam**
理解新一代优化器的重要工作。

# 本周关键词

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