每周 arXiv 好文推荐｜2026-06-13

本周主题：**Mechanistic Interpretability、训练动力学、相变理论、世界模型、RLHF、多模态学习与新型架构。**

---

1. Detecting Functional Memorization in Code Language Models

arXiv: https://arxiv.org/abs/2606.12764

代码模型可能不仅记住了训练集中的代码字符串，还记住了程序的功能行为（Functional Memorization）。作者提出新的检测框架，能够区分模型是在真正泛化，还是仅仅记住了等价程序的功能映射。这对于代码生成模型的评测与数据污染分析具有重要意义。

---

2. Where Computation Lives Inside TabPFN: Causal Localisation of Attention Head Function

arXiv: https://arxiv.org/abs/2606.12917

TabPFN 是近年来非常成功的表格基础模型。这篇工作通过因果干预方法定位不同 Attention Head 的功能，揭示了模型内部哪些头负责统计推断、哪些头负责信息路由，是 Mechanistic Interpretability 在非语言模型上的优秀案例。

---

3. The Geometry of Phase Transitions in Generative Dynamics via Projection Caustics

arXiv: https://arxiv.org/abs/2606.13191

作者提出一种几何学视角理解生成模型中的相变现象。论文将生成过程中的突变行为与光学中的 Caustics（焦散）联系起来，建立了生成动力学、奇异性理论与相变之间的统一框架。

---

4. Different Layers, Different Manifolds: Module-Wise Weight-Space Geometry in Transformer Optimization

arXiv: https://arxiv.org/abs/2606.13276

Transformer 不同模块的参数空间几何结构是否相同？作者发现 Attention、MLP 和 Embedding 层实际上位于不同的低维流形上，并呈现出显著不同的优化行为。这为模块化优化器和结构化训练策略提供了理论依据。

---

5. Phase Transitions in Attention: A Bayesian Theory of Copy Head Emergence

arXiv: https://arxiv.org/abs/2606.12058

Copy Head 是 Transformer 中最经典的涌现现象之一。这篇论文首次给出了贝叶斯理论解释，证明当数据量、模型规模或信噪比跨越某个临界点时，Copy Head 会经历类似物理相变的突然出现。

---

6. ICA Lens: Interpreting Language Models Without Training Another Dictionary

arXiv: https://arxiv.org/abs/2606.11722

目前许多解释性工作依赖 Sparse Autoencoder (SAE)。这篇论文提出直接利用 ICA（Independent Component Analysis）分析神经元表示，无需额外训练字典即可获得可解释特征，为轻量级 Mechanistic Interpretability 提供了新方向。

---

7. The Standard Interpretable Model

arXiv: https://arxiv.org/abs/2606.12289

作者试图回答一个根本问题：什么样的模型才算“可解释”？论文提出类似统计学中 Standard Model 的统一框架，希望将不同解释性方法纳入同一个理论体系中，是 Interpretability 理论化的重要尝试。

---

8. What Fits (Into Few Tokens) Doesn't Overfit: Compression and Generalization in ML Research Agents

arXiv: https://arxiv.org/abs/2606.11045

研究 Agent 为什么能够泛化。作者发现，能够被压缩成短提示（Few Tokens）的规律更容易泛化，而复杂冗长的策略更容易过拟合。这为 Research Agent、自动科学发现和推理压缩提供了新的理论视角。

---

9. Between Amnesia and Chaos: A Memory–Stability–Expressivity Trilemma for Trainable Dissipative Oscillator Networks

arXiv: https://arxiv.org/abs/2606.09929

论文提出一种新的“三难困境”：记忆能力、动态稳定性和表达能力无法同时最大化。作者在可训练振荡器网络中严格分析了这一权衡关系，与 Reservoir Computing 和连续时间神经网络密切相关。

---

10. Rank Collapse, Fixed Points, and the Renormalization Group Structure of MLP Residual Networks

arXiv: https://arxiv.org/abs/2606.10324

这是 Physics of AI 味道非常浓的一篇工作。作者将残差网络中的 Rank Collapse 现象解释为重整化群（Renormalization Group）流中的固定点行为，建立了深度网络动力学与统计物理之间的联系。

---

11. When to Align, When to Predict: A Phase Diagram for Multimodal Learning

arXiv: https://arxiv.org/abs/2606.11190

多模态学习究竟应该做对齐（Alignment）还是预测（Prediction）？作者给出了一个相图（Phase Diagram），指出不同数据规模和模态相关性下最优训练策略的变化规律。

---

12. Skip a Layer or Loop It? Learning Program-of-Layers in LLMs

arXiv: https://arxiv.org/abs/2606.06574

当前 Transformer 的层结构是固定的。作者提出让模型动态决定哪些层执行、哪些层跳过、哪些层重复调用，相当于学习一个 Layer Program。该思路有望提高推理效率和模型适应性。

---

13. Do Video Foundation Models Understand Intuitive Physics? A Layerwise Probing Analysis

arXiv: https://arxiv.org/abs/2606.09646

视频基础模型是否真的学会了物理规律？作者逐层探测模型内部表示，分析物理知识在不同层中的形成过程。与上周的《Invisible Hand of Physics》形成有趣呼应。

---

14. A Unifying View of Attention Sinks: Two Algorithms, Two Solutions

arXiv: https://arxiv.org/abs/2606.08105

Attention Sink 是长上下文 Transformer 中的重要现象。作者指出目前存在两类本质不同的机制，并给出了统一理论框架，解释为什么不同修复方案在不同场景下表现差异巨大。

---

15. Explaining Data Mixing Scaling Laws

arXiv: https://arxiv.org/abs/2606.08167

大模型训练中，不同数据源应该如何混合？论文给出了 Data Mixing Scaling Law 的理论解释，说明为什么最佳数据配比会随着模型规模变化而变化。

---

16. Rethinking the Divergence Regularization in LLM RL

arXiv: https://arxiv.org/abs/2606.09821

RLHF 与 RLVR 通常依赖 KL 正则化约束策略漂移。作者指出传统 KL 目标存在理论缺陷，并提出新的散度正则化框架，为后训练优化提供新的思路。

---

17. EinSort: Sorting is All We Need for Tensorizing LLM

arXiv: https://arxiv.org/abs/2606.08565

一篇非常有创意的系统论文。作者发现排序操作可以作为张量化 Transformer 的核心构件，从而显著降低计算和存储开销，为高效 LLM 推理提供新路径。

---

18. The Spectral Dynamics and Noise Geometry of Muon

arXiv: https://arxiv.org/abs/2606.08388

继上周解释 Muon 为什么优于 Adam 之后，这篇论文进一步分析 Muon 的谱动力学与噪声几何结构。作者揭示了 Muon 如何改变梯度噪声的传播方式，并影响训练轨迹的稳定性。

本周 Top 5 推荐

🥇 **Phase Transitions in Attention: A Bayesian Theory of Copy Head Emergence**

🥈 **Rank Collapse, Fixed Points, and the Renormalization Group Structure of MLP Residual Networks**

🥉 **The Geometry of Phase Transitions in Generative Dynamics via Projection Caustics**

🏅 **What Fits (Into Few Tokens) Doesn't Overfit**

🏅 **Do Video Foundation Models Understand Intuitive Physics?**

本周关键词

Mechanistic Interpretability · Phase Transition · Physics of AI · World Models · RLHF · Multimodal Learning · Transformer Dynamics · Optimization Geometry