Weekly arXiv Picks | Week of June 13, 2026

This week's papers span **Mechanistic Interpretability, Phase Transitions, Physics of AI, Optimization Dynamics, World Models, RL for LLMs, and Multimodal Learning**.

---

1. Detecting Functional Memorization in Code Language Models

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

Code models may memorize more than exact training examples—they may memorize program functionality itself. This paper introduces a framework for detecting *functional memorization*, distinguishing true generalization from memorized behavior over semantically equivalent programs. The work has important implications for code model evaluation and data contamination studies.

---

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

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

TabPFN has emerged as one of the most successful foundation models for tabular learning. Using causal interventions, the authors identify which attention heads implement statistical inference, information routing, and other computational primitives. A compelling example of mechanistic interpretability beyond language models.

---

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

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

This paper develops a geometric theory of phase transitions in generative models. The key insight is that abrupt changes in generation dynamics can be understood through *projection caustics*, mathematical structures familiar from optics and singularity theory. The result is a new bridge between geometry, dynamical systems, and generative AI.

---

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

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

Do all Transformer layers live in the same optimization landscape? Surprisingly, the answer appears to be no. The authors show that attention layers, MLP blocks, and embeddings occupy distinct low-

dimensional manifolds and exhibit fundamentally different optimization geometries. The findings may inform future optimizer and architecture design.

---

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

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

One of the most intriguing papers this week.

Copy heads are among the most iconic emergent structures in Transformers. This work develops a Bayesian theory explaining why copy heads suddenly appear once data size, model size, or signal-to-noise ratio crosses a critical threshold. A rare example of a rigorous theory for emergence in neural networks.

---

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

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

Much of modern mechanistic interpretability relies on Sparse Autoencoders (SAEs). This paper proposes a simpler alternative: directly applying Independent Component Analysis (ICA) to model activations. The resulting “ICA Lens” extracts interpretable features without training additional dictionaries, offering a lightweight interpretability pipeline.

---

7. The Standard Interpretable Model

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

Interpretability research currently consists of many disconnected tools and viewpoints. This paper attempts to define a unified framework—a “Standard Interpretable Model”—that could serve as a common foundation for understanding and comparing interpretability methods.

---

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

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

Why do some AI research agents generalize while others overfit? The authors argue that compressibility is the key: hypotheses that can be represented in a small number of tokens tend to generalize well, while verbose strategies are more prone to overfitting. The work connects information compression, scientific discovery, and agentic AI.

---

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

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

This paper identifies a fundamental trilemma between memory, stability, and expressivity in trainable oscillator networks. Improving one dimension inevitably degrades another. The results provide theoretical insights relevant to continuous-time neural networks, recurrent systems, and reservoir computing.

---

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

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

A particularly exciting paper from a Physics of AI perspective.

The authors reinterpret rank collapse in residual networks as a renormalization group (RG) flow toward fixed points. By importing tools from statistical physics, they uncover deep structural parallels between neural network dynamics and critical phenomena.

---

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

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

Should multimodal systems focus on alignment objectives (e.g., CLIP) or predictive objectives (e.g., autoregressive modeling)? This paper develops a phase diagram identifying when each paradigm is optimal, depending on data scale and cross-modal correlation strength.

---

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

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

Rather than executing every Transformer layer exactly once, why not learn a computation program over layers? The authors propose a framework where the model dynamically chooses to skip, repeat, or reuse layers. This introduces adaptive computation while preserving the Transformer backbone.

---

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

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

Do video foundation models genuinely understand physics? Using layer-wise probing, the authors investigate where and how physical concepts emerge inside video models. The findings complement recent work suggesting that video generators may contain latent physical world models.

---

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

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

Attention sinks are a major phenomenon in long-context Transformers. This paper argues that there are actually two distinct mechanisms behind attention sinks and develops a unified theoretical framework explaining when different mitigation strategies succeed or fail.

---

15. Explaining Data Mixing Scaling Laws

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

Modern foundation models are trained on mixtures of heterogeneous datasets. This paper provides a theoretical explanation for data-mixing scaling laws and explains why the optimal dataset composition changes as model size grows.

---

16. Rethinking the Divergence Regularization in LLM RL

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

KL regularization has become a standard component of RLHF and RLVR pipelines. This work argues that commonly used divergence penalties are suboptimal and proposes a new framework for controlling policy drift during reinforcement learning for language models.

---

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

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

An unexpectedly creative systems paper. The authors show that sorting operations can serve as a core primitive for tensorizing large language models, significantly reducing computational and memory costs while maintaining competitive performance.

---

18. The Spectral Dynamics and Noise Geometry of Muon

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

Following last week's theoretical explanation of Muon, this paper investigates its spectral dynamics and noise geometry. The analysis reveals how Muon reshapes gradient noise and alters optimization trajectories, offering deeper insight into why it performs so well in large-scale training.

Top 5 Picks of the Week

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

A rare theoretical explanation for an iconic emergent behavior in Transformers.

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

Perhaps the most "Physics of AI" paper of the week.

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

A beautiful connection between geometry, singularity theory, and generative models.

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

A thought-provoking perspective on compression, scientific reasoning, and AI agents.

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

An important step toward understanding whether video models contain genuine world models.

Keywords

Mechanistic Interpretability · Emergence · Phase Transitions · Physics of AI · World Models · Optimization Dynamics · Multimodal Learning · RLHF · Foundation Models