See Ideas for open research questions and recent papers I've enjoyed thinking about

Ideas:

These are open questions across reasoning/interpretability that I find interesting and important:

1. How does in-context learning update internal representations? Does implicit gradient descent occur on the weights during ICL?
2. "Data Drives Behavior" -> How does training on certain types of data increase/decrease types of model behavior?
3. How can we steer fine-tuning so that models acquire desired internal circuits while suppressing the emergence of unwanted or unsafe mechanisms?
4. What mechanisms underlie OOD reasoning failures, and can we build models whose internal representations remain stable and trustworthy under distribution shift?
5. How can we design models capable of continual, in-context learning that refine their internal representations through interaction with their environment?
6. What are methods to move beyond the next-token prediction paradigm and deepen a model's creative output?
7. How do a model's internal representations and implicit world model give rise to its goals, preferences, and emergent personality-like behaviors?
8. As we scale test-time compute for advanced models (scratchpads, chain-of-thought, tree search), what new internal mechanisms emerge, and how can we interpret or control them?

Recent Papers I've enjoyed reading!

Reasoning:

1. Recursive Language Models (Zhang et al., 2025)
2. Continuous Autoregressive Language Models(Shao et al., 2025)
3. The Surprising Effectiveness of Negative Reinforcement in LLM Reasoning(Zhu et al., 2025)
4. Parallel trade-offs in human cognition and neural networks: The dynamic interplay between in-context and in-weight learning(Russin et al., 2025)

Interpretability:

1. On the Biology of a Large Language Model(Lindsey et al., 2025)
2. Measuring Progress in Dictionary Learning for Language Model Interpretability with Board Game Models(Karvonen et al. 2024)
3. The Platonic Representation Hypothesis(Huh et. al 2024)
4. Are Sparse Autoencoders Useful? A Case Study in Sparse Probing(Kantamneni et al. 2025)

Alignment/Safety:

1. Subliminal Learning(Cloud et al., 2025)
2. Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples(Souly et al., 2025)
3. Eliciting Secret Knowledge From Language Models(Cywinski et. al 2025)
4. Persona Vectors: Monitoring and Controlling Character Traits in Language Models(Chen et al., 2025)