
Exploration track: Interpreting LRMs¹

Suvajit Majumder
Gobubble

Aviral Kaintura
Indian Institute of
Technology Madras

Subramanyam Sahoo
BASIS, UC Berkeley

Johnson Arokiadoss
Affiliation

Jedrzej Konrad Kolbert

With
PIBBSS, Timaeus, & Apart Research

Abstract

We probe the reason behind the recent success of reasoning models on complex tasks like Maths and coding. Focusing on reflection, backtracking and other analogous exploratory tokens in the response, we differentiate between behavior of base and reasoning finetuned checkpoints of Qwen2.5-Math-1.5b on Math500 dataset. We probe the internal representation differences between these, which can help steer model-training and inference time corrections in production.

Keywords: Reasoning model, thinking trace, interpretability

1. Introduction

Reasoning models have been at the forefront of model performance across closed and open-source language modelling, leading all popular benchmarks over their non-reasoning counterparts. However, it is quite unclear what exactly leads to this performance improvement, and how we can efficiently steer model training and inference towards this behavior selectively in a task-specific manner. There have been numerous recent efforts to understand the implications of these longer thinking traces, with mixed results regarding their faithfulness and relevance in reaching the final answer. Fine-grained understanding on the evolution of the quality of the thinking traces over the course of the training is still lacking, esp. in settings like Deepseek-r1 where only format-rewarding and outcome-based rewarding is used for post-training RL. This, in turn, leads to the emergence of uncontrolled behaviors like sycophancy and deterioration of misaligned tasks (not part of the RL training dataset).

¹ Research conducted at the [AI Safety x Physics Grand Challenge](#) | [Apart Research](#), 2025

Our goal in this project is to accurately map the behavior of a (small) reasoning model on Math-based prompts, during the course of its journey to get to the final answer. We use various monitoring and visualisation techniques borrowed from existing interpretability literature and toolkits such as Logit Lens and Circuit Tracer. In addition, we define some of our own heuristic metrics inspired by physics and statistics, and information theory.

2. Motivations

Reasoning models are being heavily used for day-to-day coding and planning tasks for developers, researchers and the casual explorers. Understanding and quantifying the statistics of generated rationale behind any response, both internally and in output tokens is paramount to ensure alignment and safety in wide-scale usage. On another side of the coin, this can also help with guiding training with new objectives reducing verbosity of the models on simpler tasks.

One key feature common to the thinking traces in reasoning model responses is the presence of trigger words/phrases like “but”, “so”, “Let’s verify”, etc. But, what leads to the model deciding on using them: low confidence on answer, pre-training bias, etc.? And how do we reach these trigger words during inference, i.e. what are the alternatives to using these trigger words at this point of time in the inference step. Another interesting question one can ask is what are the internal model representations leading to these output tokens: are there any universality of representations? Also, what’s the confidence score on the answer token at this point, and along the duration of token generation?

Coarse-graining, scaling and emergence are well-studied phenomena across various branches of physics from condensed matter systems, to thermodynamics, to fluid dynamics, to biophysics. Observables inspired from these, and other mature fields of study can and should be used to guide our understanding of behaviors in large language modelling. Bottom-up approaches of AI modelling theories have been far outpaced by the lightning speed of deployment of larger and more sophisticated models, necessitating a phenomenological approach. In such a situation, it’s even more critical that we use heuristics to guide us in order to make practical gains. In this project, we have tried to follow this ethos, and developed monitoring tools and metrics guided by renormalisation group theory and information theory in practically useful models of size 1-14B parameters.

3. Proposed Research

Theoretical background:

Interpretability toolkits such as TransformerLens, logit lens and circuit tracer have been developed over the past several years to better understand the internal reasoning of the model preceding any response. This involves looking at the

activations during the forward pass of the input tokens through the model. We leverage such existing tools initially, along with developing our own ones (especially for test-time compute scaling) during the course of the project for initial investigations into the model’s internal features.

Coarse-graining in the feature space can be a useful technique to gain more robust signals from the model, such as whether it’s reaching a phase transition, performing in a low-confidence regime (hallucinations), etc. Renormalisation groups provide useful methods to identify such features from the model internal activations. On the output layer, we monitor spread of confidence on target tokens during the course of generation using dimensional reduction techniques.

Method description:

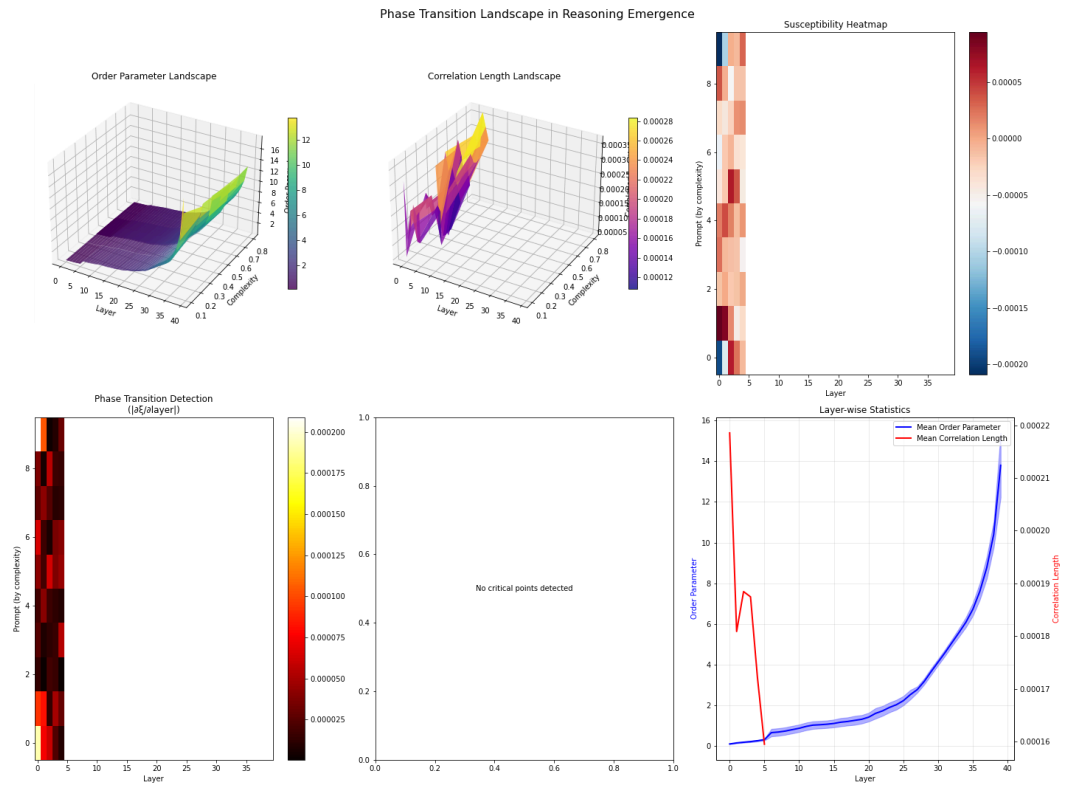
We start with reasoning models like “Deepseek-R1-distill-1.5B” and its non-reasoning counterpart “qwen2.5-Math-1.5B”. We test them on the Math500 dataset where the first model outperforms the second one and compare their thinking traces using the toolkits mentioned above. We look for the trigger tokens that lead to further generation before the statement of the answer, and monitor the activation space in such scenarios. We also contrast the confidence in the answer logits in the reasoning and non-reasoning models and how they evolve over the course of generation of output tokens. Furthermore, we do a parallel set of experiments on a larger model “qwen-3-14B” in order to identify scaling related differences.

Steps for research project:

1. Train base models of varying sizes using GRPO to incorporate reasoning in niche tasks. Identify phase transitions, the so-called “aha moment”, and map activation and weight space statistics preceding/succeeding it.
2. Scale the analysis for larger models.
3. Compute RG-based computation metrics that show positive early signals of phase transitions.
4. Monitor confidence, entropy, etc. for output token across training checkpoints and model scales.
5. Detect observable statistics and heuristic metrics that detect early phase transitions, model hallucination, etc.

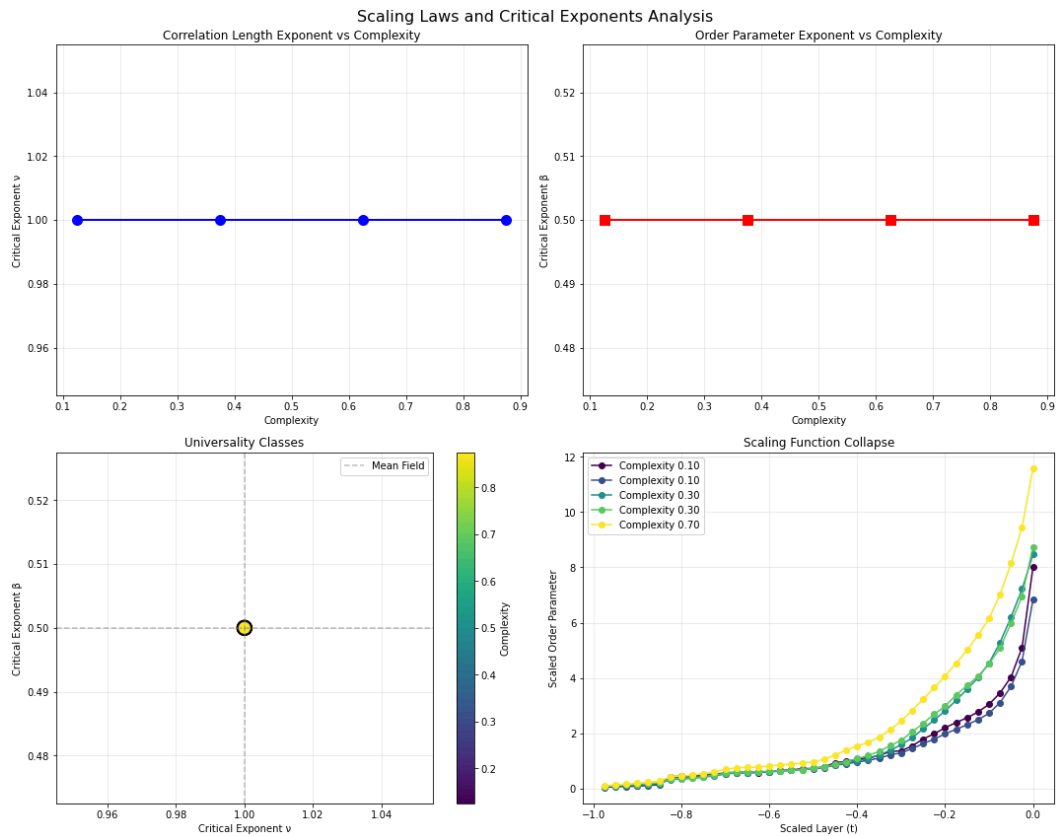
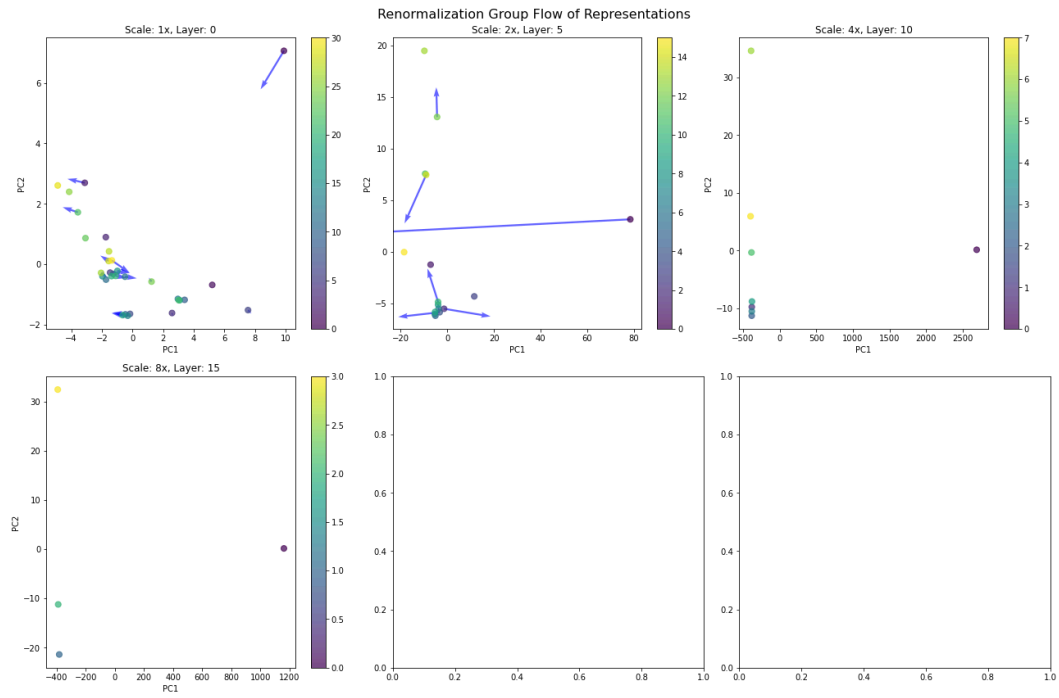
Relevant results from prelim experiments:

Fig 1-7: Qwen-3-14B RG flow toy reasoning dataset



Analysis Summary:

- Analyzed 10 prompts across 40 layers
- Detected 0 critical points
- Mean order parameter: 2.7791

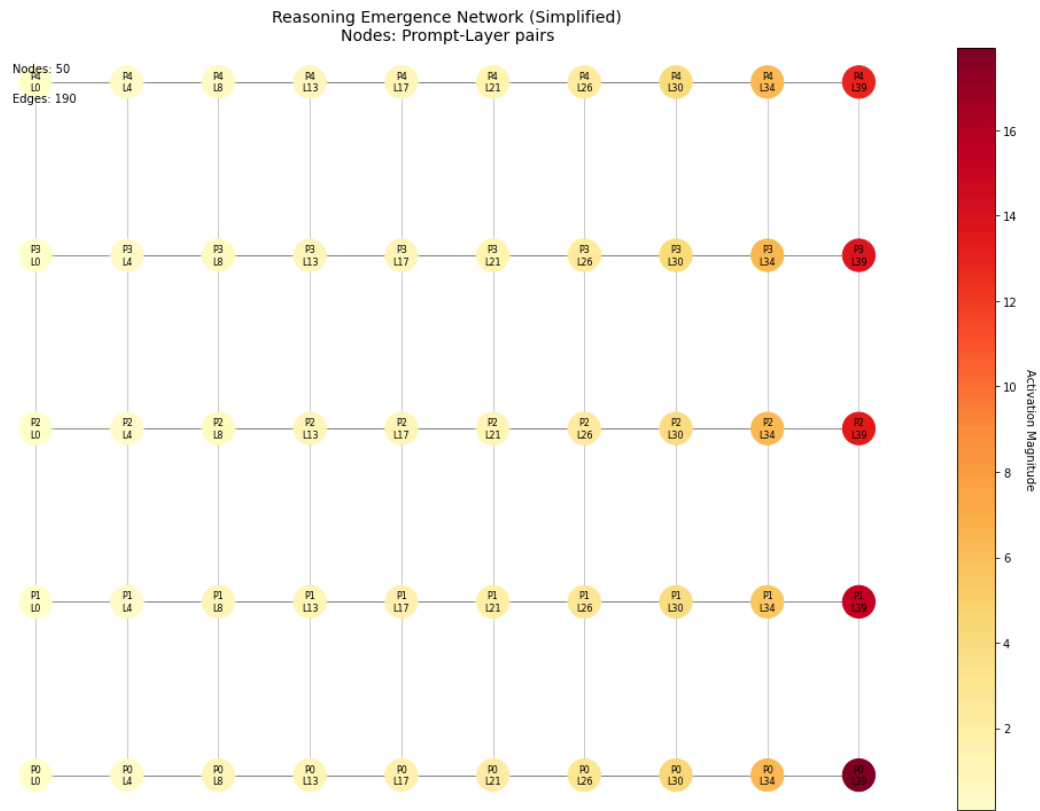


Scaling Analysis Summary:

Complexity bins analyzed: 4

Mean ν exponent: 1.000

Mean β exponent: 0.500



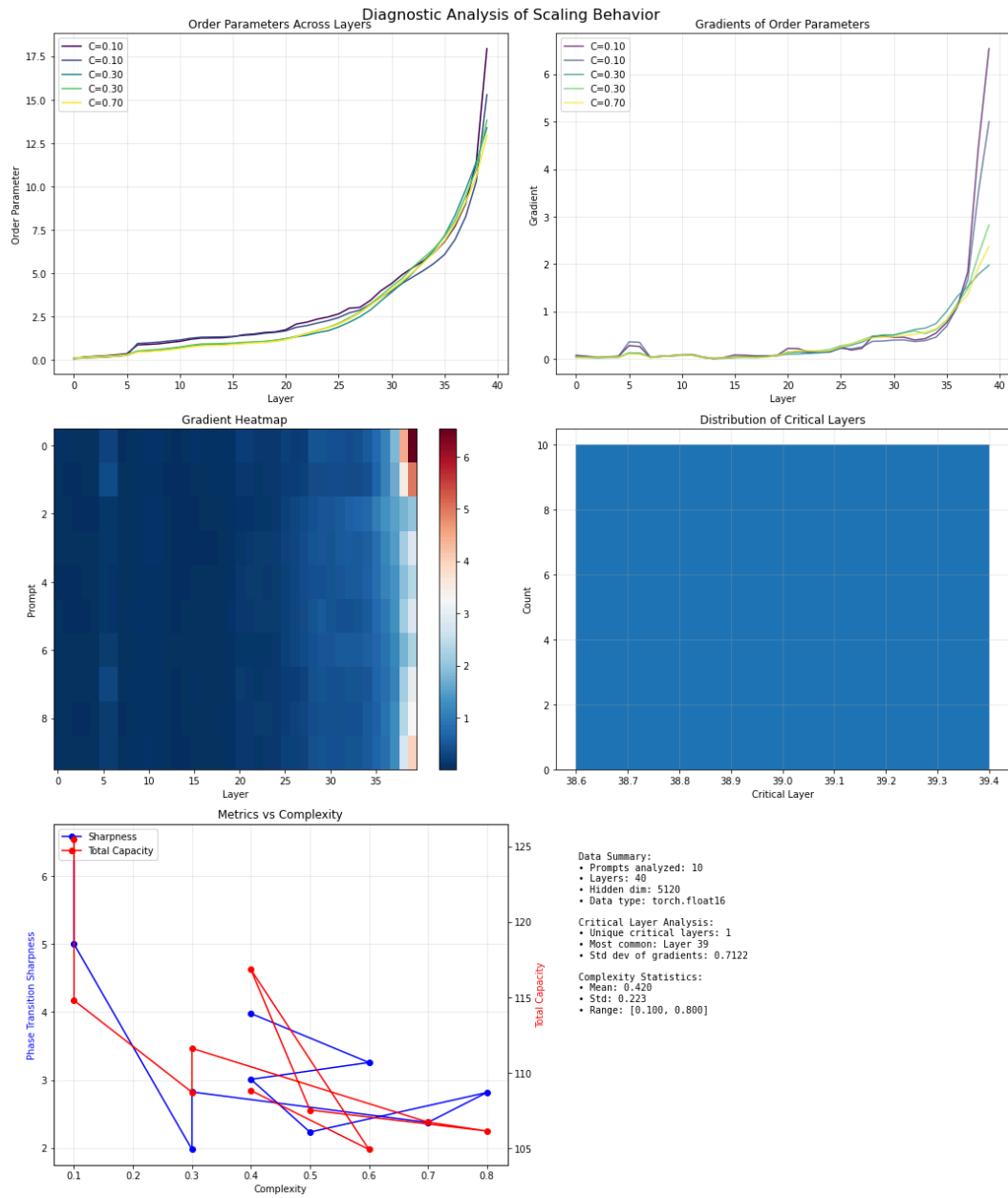
```
=====
INVESTIGATING LAYER 39:
=====

Total layers in model: 40
Layer 39 is 97.5% through the model

This suggests the model has a consistent phase transition point,
regardless of input complexity. This could indicate:
1. A structural property of the model architecture
2. All prompts trigger similar processing patterns
3. The complexity metric doesn't capture processing difficulty

=====
ANALYZING LAYER 39 NEIGHBORHOOD:
=====

Examining layers [36, 37, 38, 39]
Layer 36: mean activation = 7.7871, std = 0.4661
Layer 37: mean activation = 9.1300, std = 0.5161
Layer 38: mean activation = 10.9361, std = 0.4451
Layer 39: mean activation = 14.6781, std = 1.8133
```



DIAGNOSING SCALING LAW ISSUES

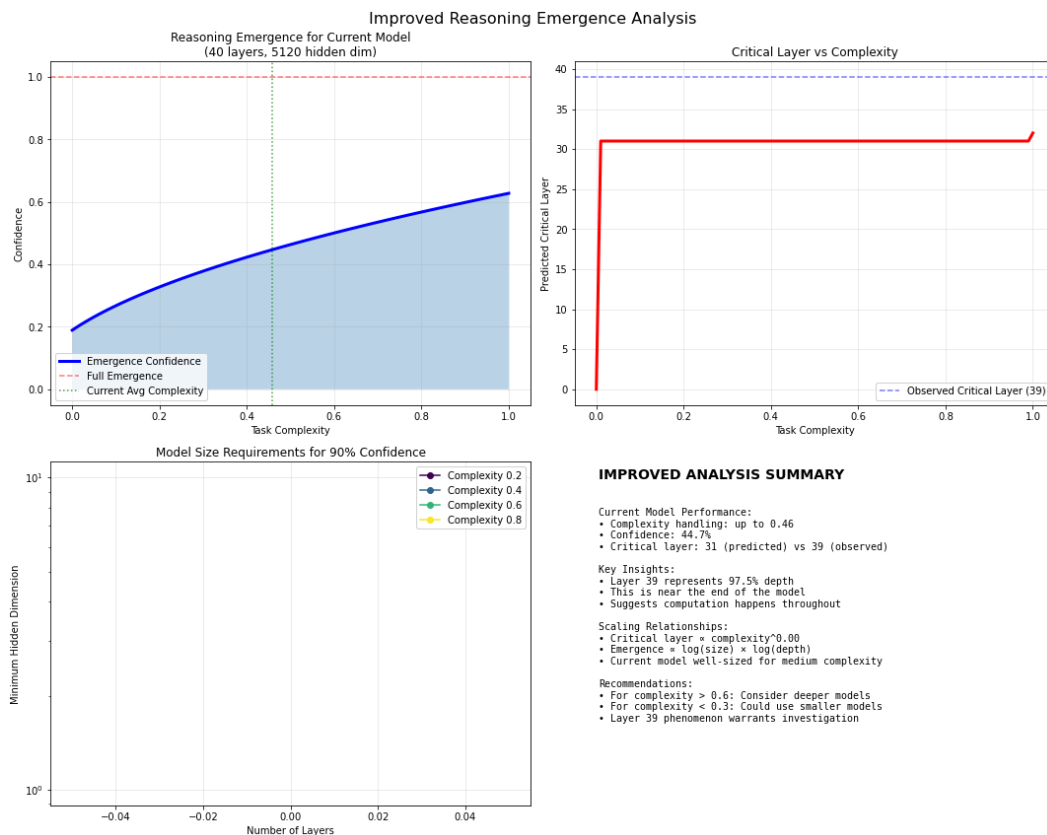
Current scaling laws:

critical_layer_exponent: 3.963339682329999e-09
correlation_length_exponent: -0.1985609914101891
emergence_threshold: inf

IMPROVED REASONING EMERGENCE ANALYSIS

Current Model Analysis (Improved):

- Model: 40 layers, 5120 hidden dim
- Average task complexity: 0.458
- Reasoning emergence prediction: NO
- Confidence: 44.70%
- Emergence score: 4.47
- Threshold: 10.00
- Expected critical layer: 31 / 40



FINAL CONCLUSIONS:

1. The model shows a universal phase transition at layer 39 (97.5% depth)
2. This is independent of input complexity, suggesting architectural constraints
3. Current model has 45% confidence for avg complexity 0.46
4. The late-layer transition suggests distributed computation throughout the model
5. Consider architectural modifications to enable earlier phase transitions for efficiency

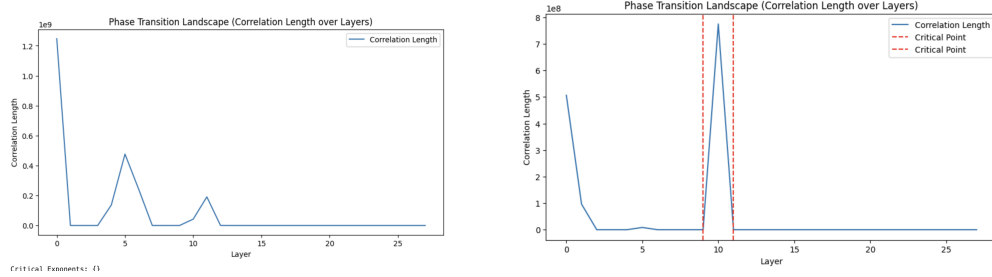


Fig 8: Phase Transition Landscapes for (left) base model Deepseek-R1-distill-1.5B, (right) model finetuned on single task example. We observe that the peak at layer 5 disappears as a result of fine tuning while the peak at layer 10 remains.

4. Discussion and Conclusion

In this work, we show promising initial signs of using RG based metrics for detecting phase transitions and monitoring test-time compute behavior. In future work, we plan to expand on this by scaling to larger models, and using more contrastive performances to detect key distinguishing features relevant for alignment and safety.

5. References

Nanda, N., & Bloom, J. (2022). *TransformerLens*. GitHub.
<https://github.com/TransformerLensOrg/TransformerLens>