Scaling Laws for Computational Mechanics

Group: Katarzyna Łyczek
Area: Computational mechanics
Research question:

How does belief state geometry sharpen (or fails to sharpen) as the scale of the model increases?

Why it matters: How would an answer help alignment, or our understanding of the area?

Computational mechanics makes falsifiable predictions about what internal representations of the models should look like, assuming optimal prediction (rather than reverse-engineering the structure post-hoc, which is what most mechanistic interpretability attempts to do). Targeting an instrumental goal as an intervention target point is scale-free: it will be relevant to both current and future models (in fact, it becomes even more relevant as models get more capable and hence likely closer to the optimal predictors).

Progress in computational mechanics can be broken down into two orthogonal directions:

1. For what classes of data generating processes do we understand optimal prediction?
2. For what scale of neural networks have we identified the structures of optimal prediction?

Our project fits into the second research direction. Currently, one of the main objections to the usefulness of computational mechanics research for AI alignment is that it is limited to small models. To better understand how the existing findings of computational mechanics translate into large scale setups, we want to derive scaling laws for the sharpness of the belief state geometry.

Background: What is already known? Key papers or results you build on.

- Shai et al. (2024) showed that small transformers trained on HMM data linearly represent belief state geometry.
- Piotrowski et al. (2025) characterised how attention implements constrained belief updates.
- Shai et al. (2026) extended the account to factored processes.
- Dai et al. (2025) showed that pretrained LLMs predict HMM data near-optimally in context, with performance organised by entropy and mixing rate.
- (arXiv 2609.17376) found that in the in-context setup the belief states are linearly decodable across six open LLMs and 40 HMMs. It also showed through steering and patching that the belief subspace is causally upstream of log-NTP and of the model's predictions.

Approach: Theory, experiments, or both. What would you actually do in the first weeks?

We quantify the sharpness of the belief state geometry in terms of R^2 (and other metrics if we come up with other metrics later). We want to analyze how it scales along two axes: model size and HMM complexity (i.e. task difficulty)

We will study model scale in two complementary setups: A) the toy setup where we can precisely control the number of parameters and size of the training dataset; and B) real-world setup with LLMs and HMM learning happening in-context.

Setup A: trained-from-scratch toy transformers. Following Shai et al. (2024), we train on HMM data over a grid of depth L ∈ {1, 2, 4, 8, 16} and width d ∈ {32 … 1024}, with matched data and compute. This regime lets us separate depth from width cleanly, as well as control the size of the dataset used for training, which open model families cannot.

Setup B: in-context, pretrained models. We run the existing pipeline across the widest available size ladder within single model families, for example all sizes of Llama 3.x and of Qwen3.5. Keeping the family fixed holds data and architecture roughly constant. The pipeline covers KL versus k-HMM crossover, belief probes with shuffled and random controls, the surplus over NTP and 1-HMM, and steering and patching.

In each of these setups, we can scale HMM complexity by:

- increasing the number of hidden states in the HMM
- considering non-ergodic HMMs (https://www.lesswrong.com/posts/JfJ4WTRHmooPBWRFv/the-geometry-of-nonergodic-composition)
- considering factorizable HMMs (https://arxiv.org/abs/2602.02385)

What success looks like: A result, a proof, a plot, a negative finding?

We obtain heatmaps showing R^2 where on one axis we have model size, on the other we have HMM complexity.

For toy models and for a fixed HMM, we can also plot a depth x width heatmap of R^2.

A negative finding would be that there is no clear relationship to be observed.

Risks and open questions: Where could this fail? What are you unsure about?

- **Saturation.** The paper's HMMs were selected so that LLMs already do well, and R^2 is often around 0.99. There may be little room to observe sharpening, so harder processes are probably necessary.
    - As we increase the number of hidden states in the HMM above 3, it is no longer straightforward to visualize the belief state geometry in 3 dimensions – we might have to rely on metrics like R^2 to quantify the sharpness of the structure.
    - We can increase difficulty by extending the setup to non-ergodic HMMs (https://simplex.pub/nonergodic-geometry/).
- **What counts as "scale."** Models of different sizes are usually trained on different token counts, and some are MoE or use "effective parameter" counts, as Gemma does. We have to be careful about what we mean about model scale. Parameters, training compute and data are entangled. Regime B is our main defence here – in this setup we can control the size of the training dataset exactly.
- **Probe capacity grows with width.** With about 1,000 training positions and d in the thousands, OLS probes are close to underdetermined. A bigger model could look "sharper" simply because the probe has more features to fit with. We'll need regularisation, projection to a fixed number of principal components, or control baselines that scale with d.

Feedback (mentors, with name):