Tracking Training Phases in Compositional Learning with Task-Agnostic Measures

Niclas Dern* · **Selma Mazioud*** · **Jakob Heiss** · **Avrajit Gosh** · **Curtis James McDonald** · **Gabriel Clara** · **Bin Yu**

*Equal contribution · UC Berkeley · Mechanistic Interpretability Workshop, ICML 2026

Problem

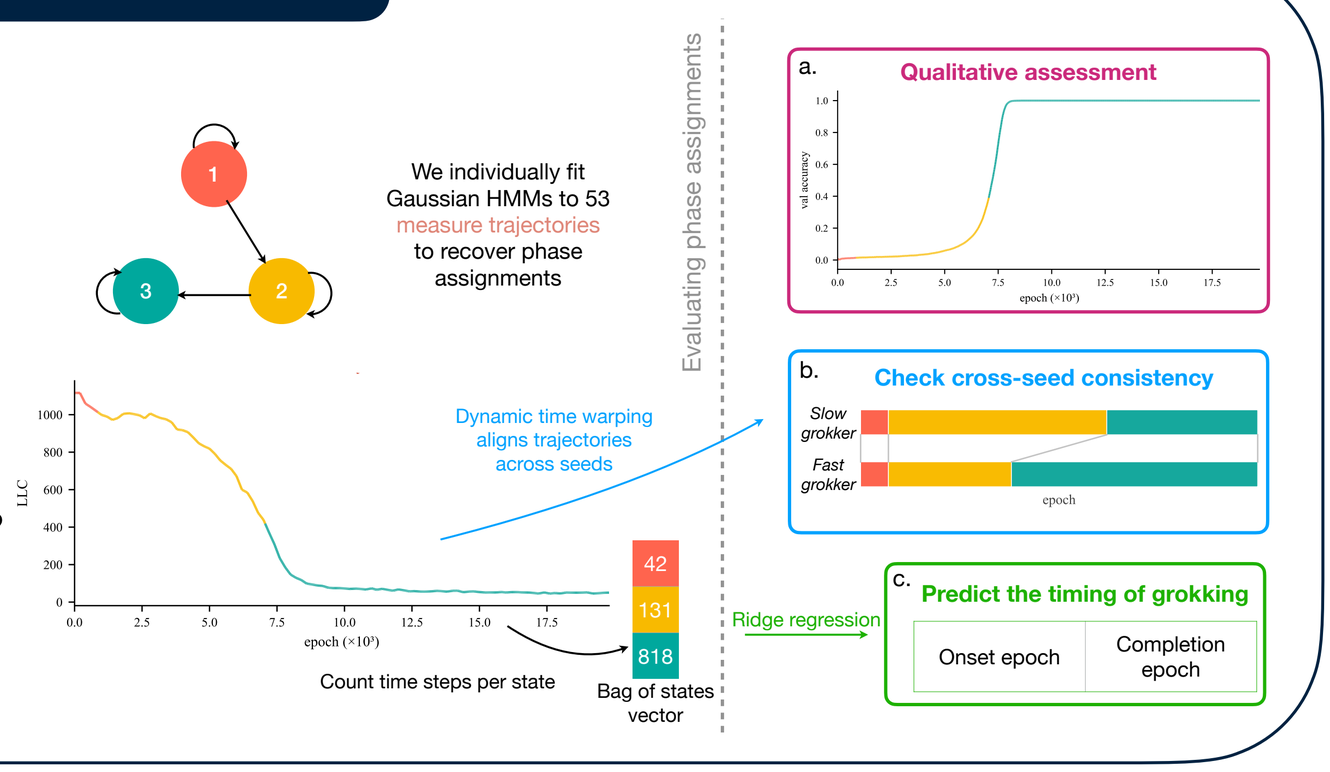
- Most reliable phase analyses need a mechanistic target chosen ahead of time and cannot flag unanticipated phases.
- Alternative: task-agnostic scalar measures, computed from a model's parameters, representations, or outputs with no reference to any target capability.
- **We systematically compare 53 measures for training-phase segmentation on a common compositional testbed.**

Our measures

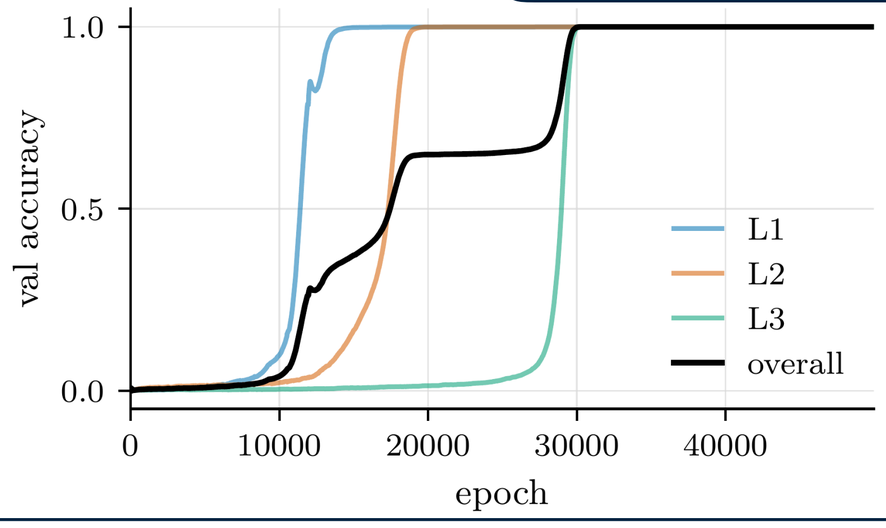
Class	#	Drawn from	Examples
Parameter-space	24	parameter values	param. norm, weight norms, spectral entropy, path norm
Loss-landscape	11	loss-surface geometry	LLC, Hessian sharpness, PAC-Bayes flatness
Representation-space	4	hidden activations	PCA / TwoNN intrinsic dim., neural-collapse stats
Output-space	14	model outputs	margin-based, prediction churn, prediction entropy

Pipeline

1. **Featurize:** per measure and run, z-score the trajectory and its first differences; pool across runs.
2. **Fit and decode:** one Gaussian HMM per measure (Baum-Welch, pooled); Viterbi-decode each run into phases.
3. **Evaluate:**
   A. qualitative match to known transitions
   B. retrospective R^2: does time-in-phase predict grokking onset and completion on held-out seeds?
   C. cross-seed consistency: ARI between phase sequences aligned across seeds by dynamic time warping.



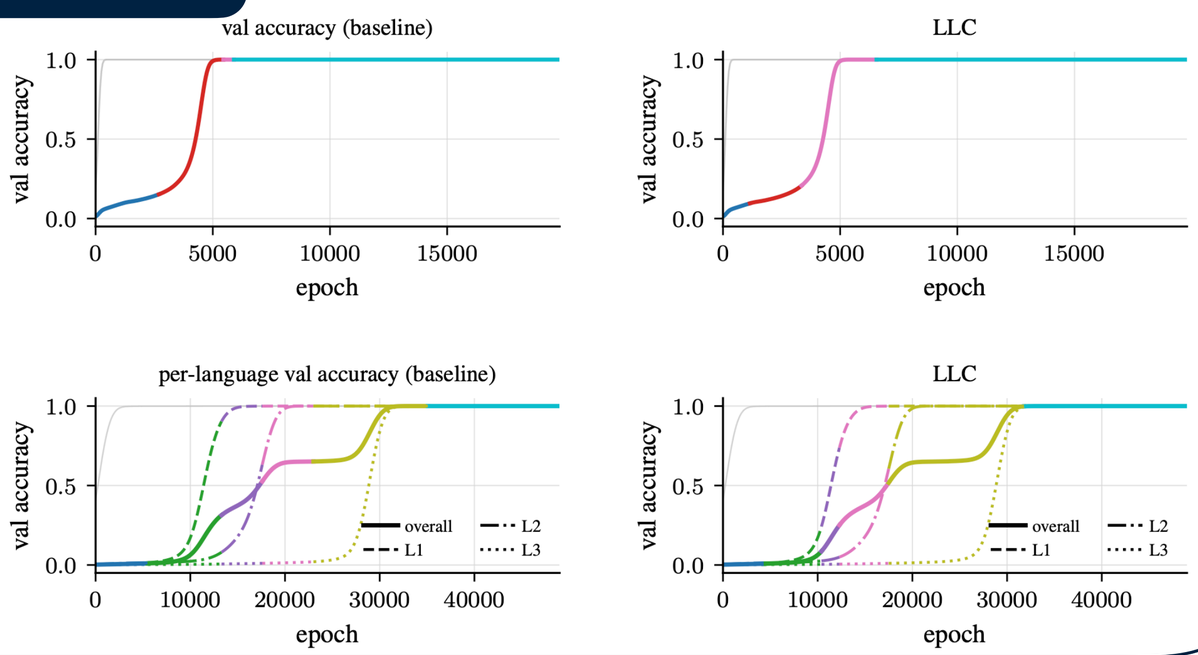
Staggered grokking as a testbed



- **Modular addition (baseline):** predict (a+b) mod 113 with a one-layer transformer, 40 seeds. *Grokking* = generalization long after memorization.
- **Multilingual (new):** same task in disjoint token vocabularies ("languages"). Smaller per-language fraction → later grokking, so languages grok in sequence (30%, 27%, 21%). The fraction spacing tunes overlap; a two-language variant (30%, 21%) stays separated.

Results

1. **When transitions are separated, task-agnostic measures recover the known phases.** Overlap breaks recovery across the board.
2. A small group of measures is consistently strongest: **prediction entropy, the LLC, and PCA effective dimension.**
3. The bottleneck is the HMM framework, not only the measures. Performance is poor even when we fit the HMM directly to the per-language validation-accuracy reference.



Berkeley UNIVERSITY OF CALIFORNIA · AISI · NSF · Swiss National Science Foundation