

Geometric Precursors to Emergent Capability

NeurIPS-style research report • chronology: v1 → v6 → v8 → v9 → belief/domain extension

Abstract. We implement an outcome-gated framework for testing whether internal developmental transitions can provide reproducible lead time before a subsequently emerging capability. The work deliberately separates (i) instrument calibration, (ii) capability emergence, (iii) temporal precursor testing, (iv) predictive validation, and (v) causal/mechanistic follow-up. The latest executed controlled run validates the instrumentation on modular-arithmetic grokking and across three calibration seeds, but the five-seed planning experiment did not cross its independently defined behavioral-onset threshold. Consequently, no positive precursor, lead-time, causal, or warning-monitor claim is made. A separate belief/domain fine-tuning extension demonstrates a working real-LLC pipeline in a smoke configuration, but is explicitly exploratory and does not establish universal transfer.

1. Research question and scientific contract

The central hypothesis is narrow: can measurable internal transitions precede a later, independently measured capability onset? For metric m , lead time is defined as $\tau_m = t_{\text{behavior}} - t_m$. Positive τ is observed temporal precedence, not causality. The pipeline is outcome-gated: later analyses are disabled when earlier empirical gates fail. This prevents retrospective selection of attractive internal curves.

2. Chronological development

v1. Established the source-grounded experimental contract: real modular-arithmetic calibration, multiple seeds, a real multi-step planning task, outcome-gated precursor analysis, and explicit non-claims. **v6.** Added validated-LLC and architecture-specific causal gates, including chain convergence/sensitivity requirements and intervention controls. **v8.** Integrated the full end-to-end orchestration and Hugging Face external-validation track, with frozen transfer rules. **v9.** Hardened T4/HPC execution, fixed plotting/runtime issues, and added publication-quality diagnostics. **Belief/domain.** Extended the framework to fine-tuning robustness questions using real LLC, adversarial prompting, and domain overreach measures.

3. Experimental design

Calibration uses modular addition $(a+b) \bmod p$ with a train/test split and a 2-layer, 128-dimensional transformer. Developmental instrumentation includes loss/accuracy, representation CKA/drift, participation-ratio effective dimension, parameter norms, checkpoint persistence, and change detection. The controlled capability task is a genuine supervised multi-step sequence-learning/planning problem with hidden intermediate structure and held-out compositions; it is not replaced by a proxy or a hand-authored risk score. Five independent capability seeds (11, 22, 33, 44, 55) were evaluated.

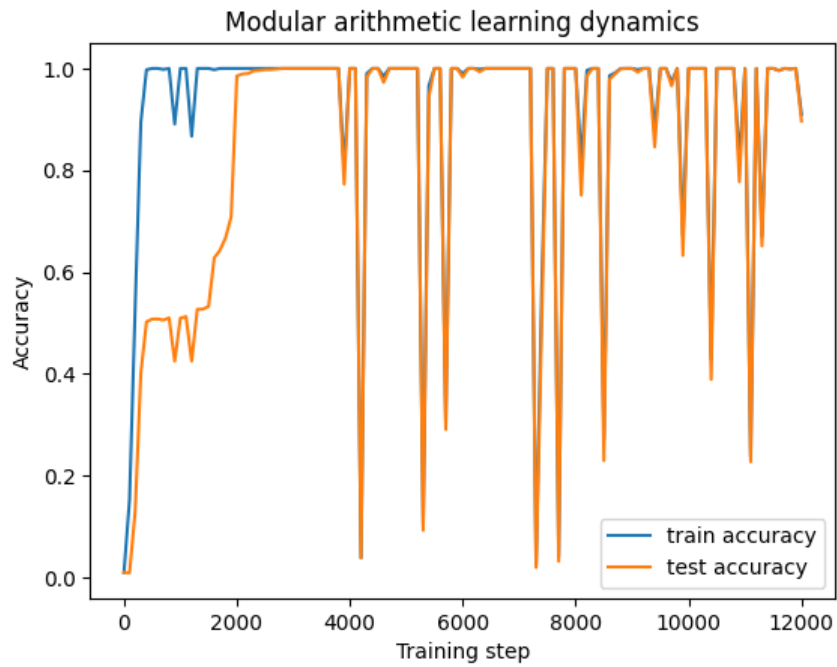


Figure 1. Modular-arithmetic calibration learning dynamics. This validates the developmental instrumentation on a known learning transition; it is not itself evidence of a warning signal.

4. Engineering system

The implementation targets a 16-GB NVIDIA T4. It uses FP16 AMP where appropriate, *torch.inference_mode()*, fixed-shape compilation, pinned/asynchronous transfers, bounded activation caches, checkpoint artifacts, and selective expensive computation. Hugging Face developmental analysis uses FP16 weights without scientific quantization; external inference may use supported 4/8-bit screening. Attention uses SDPA/memory-efficient backends where available. The design keeps LLC and causal intervention analyses gated rather than running them indiscriminately.

5. Results: the controlled developmental test

Calibration. Phase A passed; Phase B completed three independent calibration runs with finite metrics and non-trivial learning. In the displayed grokking trajectory, test accuracy reaches 1.0 around step 3,000 and remains high while representation drift and effective dimension continue to fluctuate. The point is instrumentation validity, not an inference about risk.

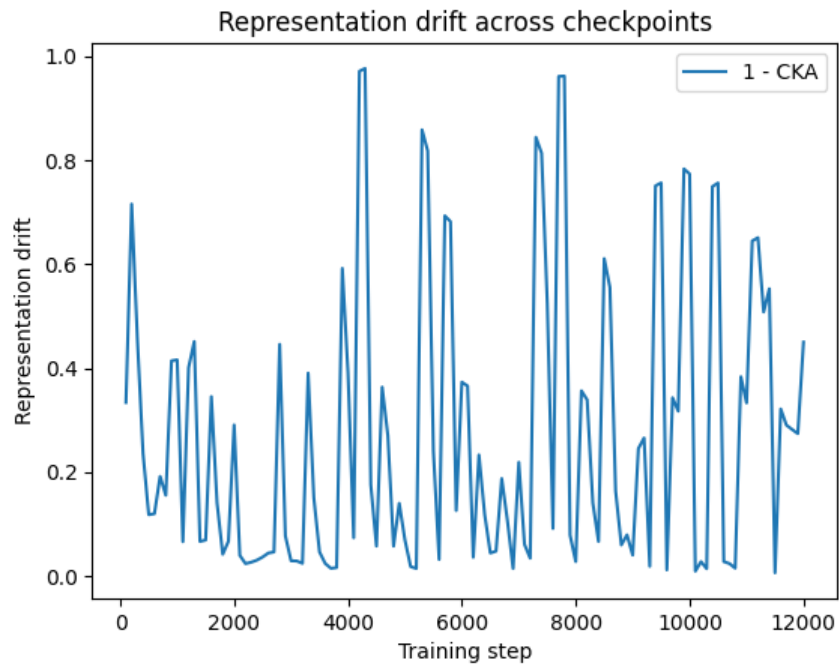


Figure 2. Representation drift across calibration checkpoints.

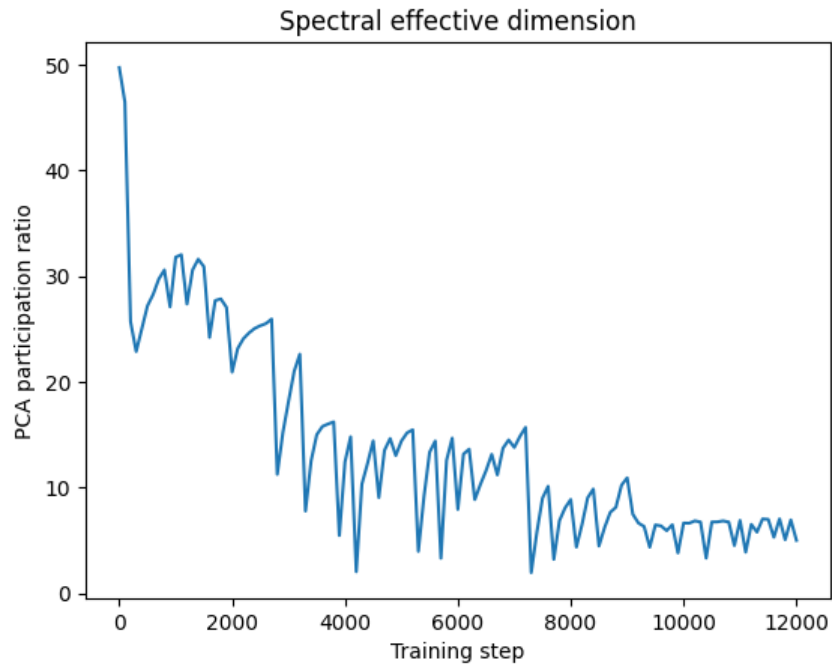


Figure 3. Spectral participation-ratio effective dimension across the developmental trajectory.

Capability outcome. Across all five seeds, exact-sequence accuracy = 0.0 and path validity = 0.0 at the evaluated terminal point. All five behavioral onsets are null. The temporal precursor table therefore contains no valid positive lead-time observations. Loss alone had a detectable transition around step 200, but because behavioral onset was absent, it cannot be converted into a warning lead time.

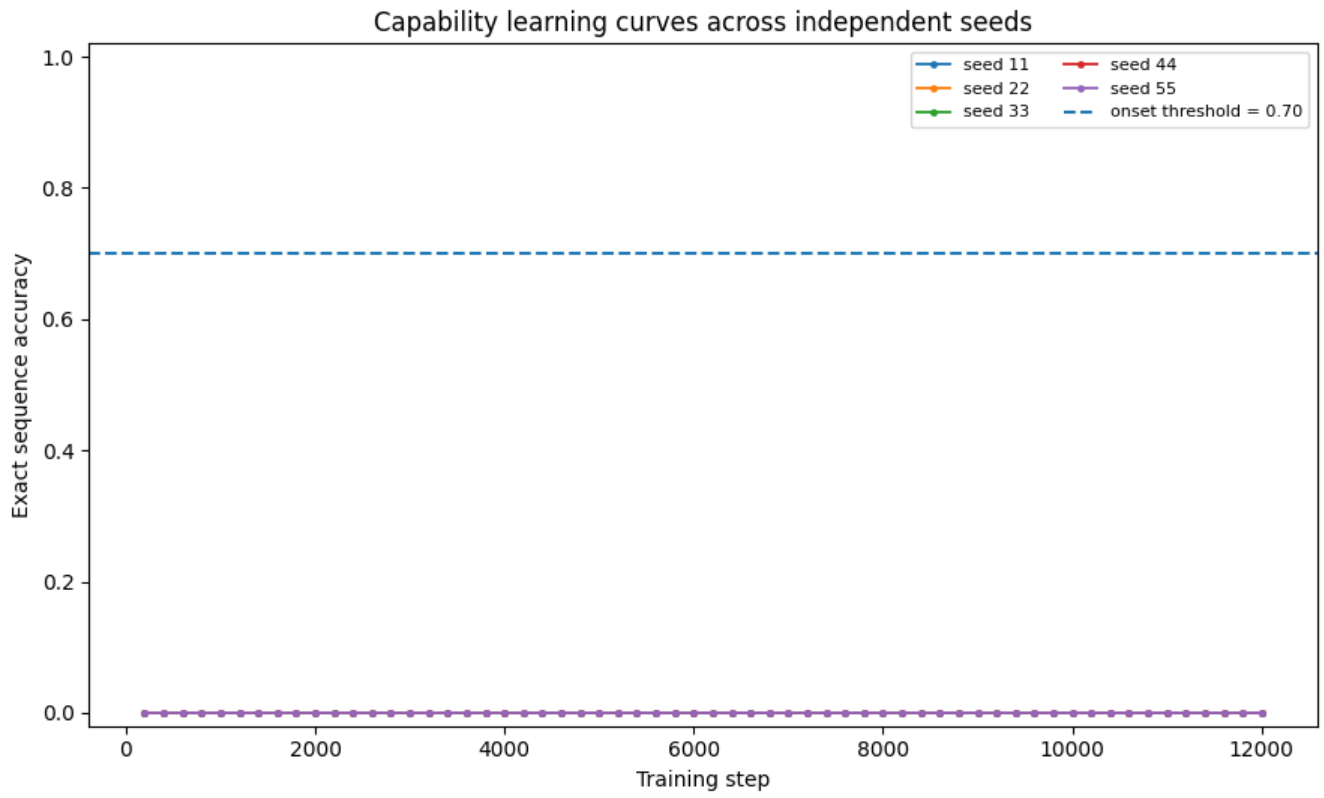


Figure 4. Capability learning curves across the five independent seeds. No seed crosses the pre-specified behavioral-onset criterion.

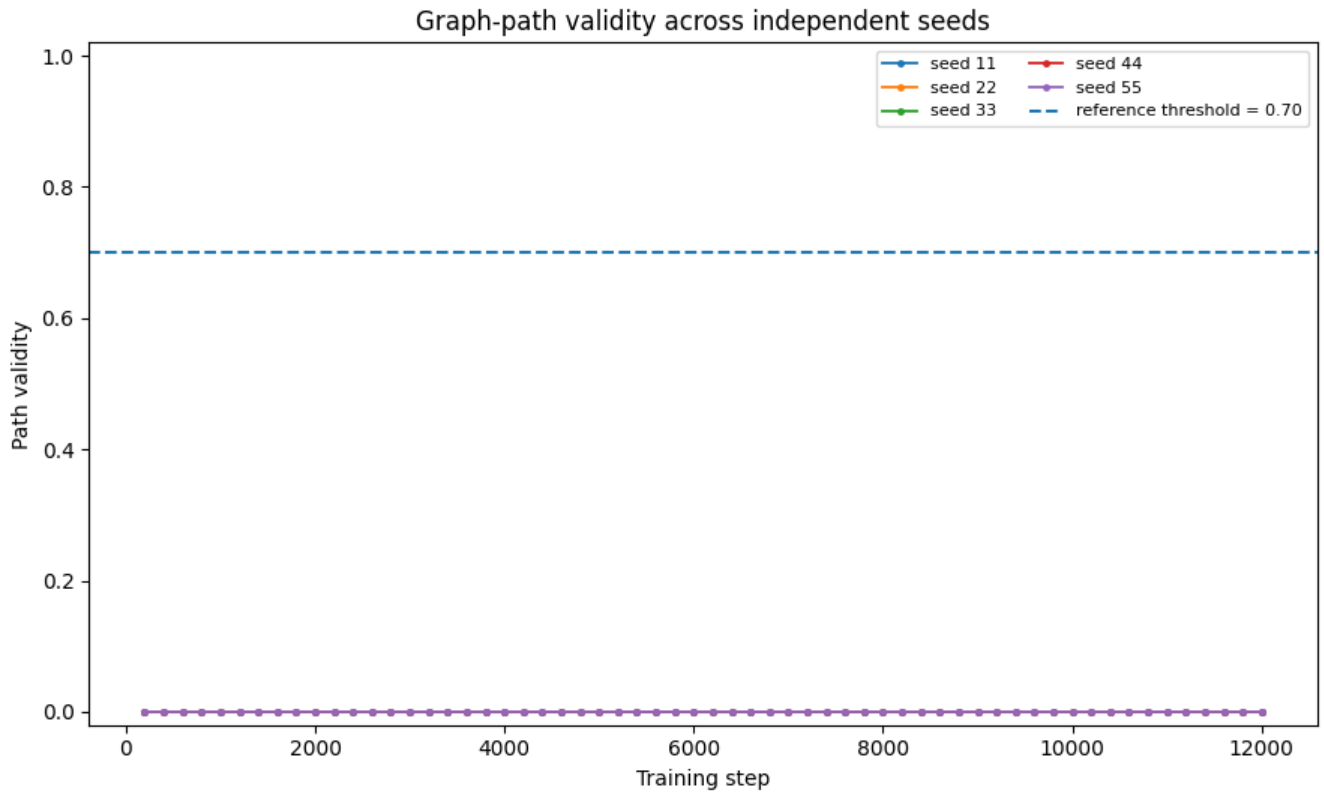


Figure 5. Graph-path validity across independent capability seeds; terminal validity remains zero.

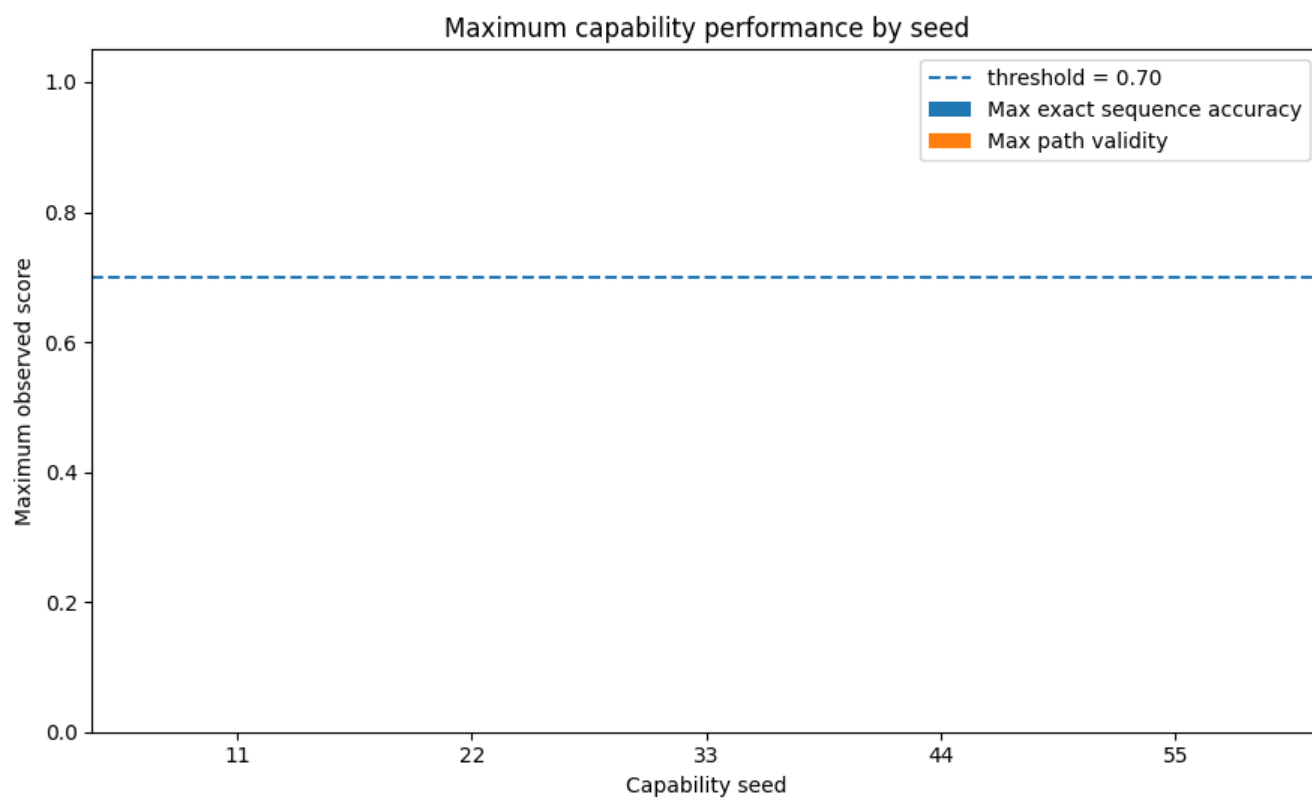


Figure 6. Maximum capability performance by seed; the evaluated runs do not exhibit the required onset.

6. Calibration reproducibility and developmental diagnostics

The additional calibration panels expose cross-seed behavior rather than relying on a single illustrative trajectory. They are retained in the main document at 0.75× native dimensions so that the original diagnostic detail remains inspectable.

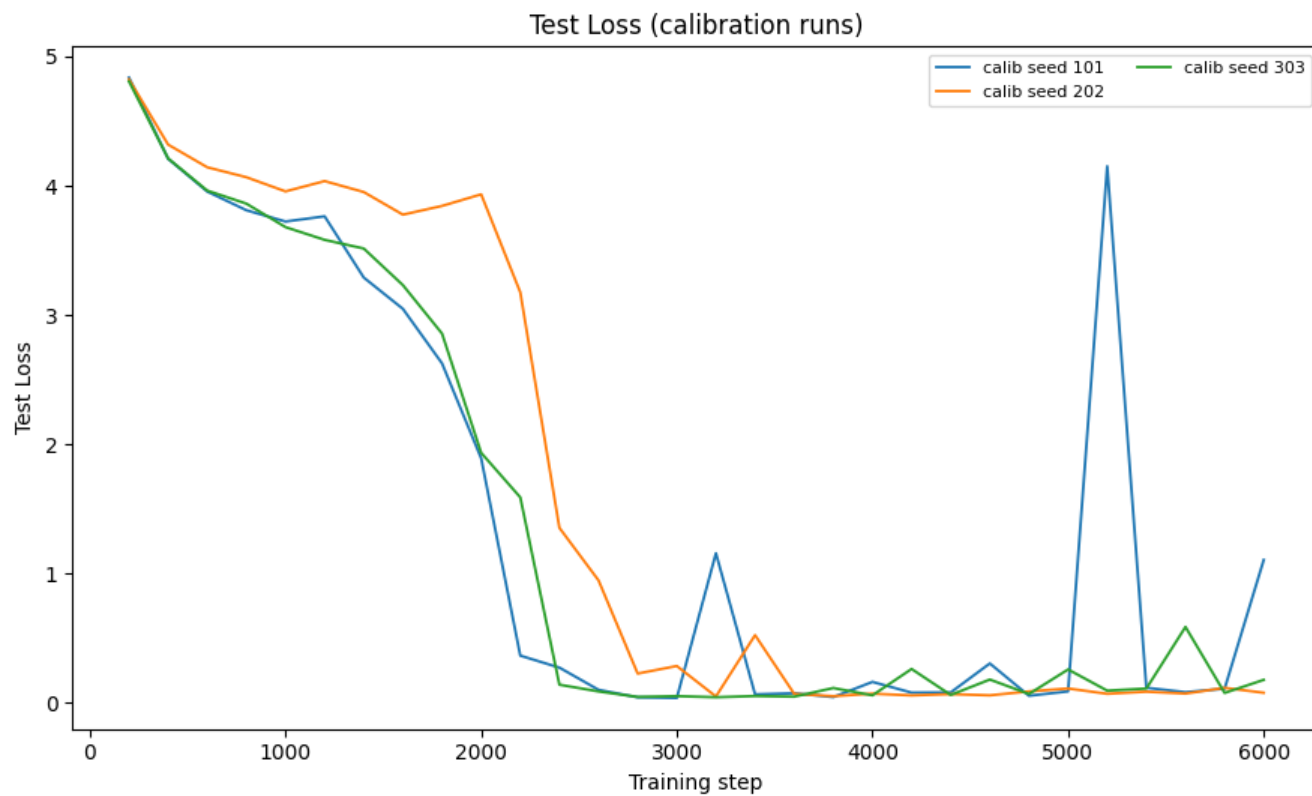


Figure 7. Test-loss calibration runs across independent seeds.

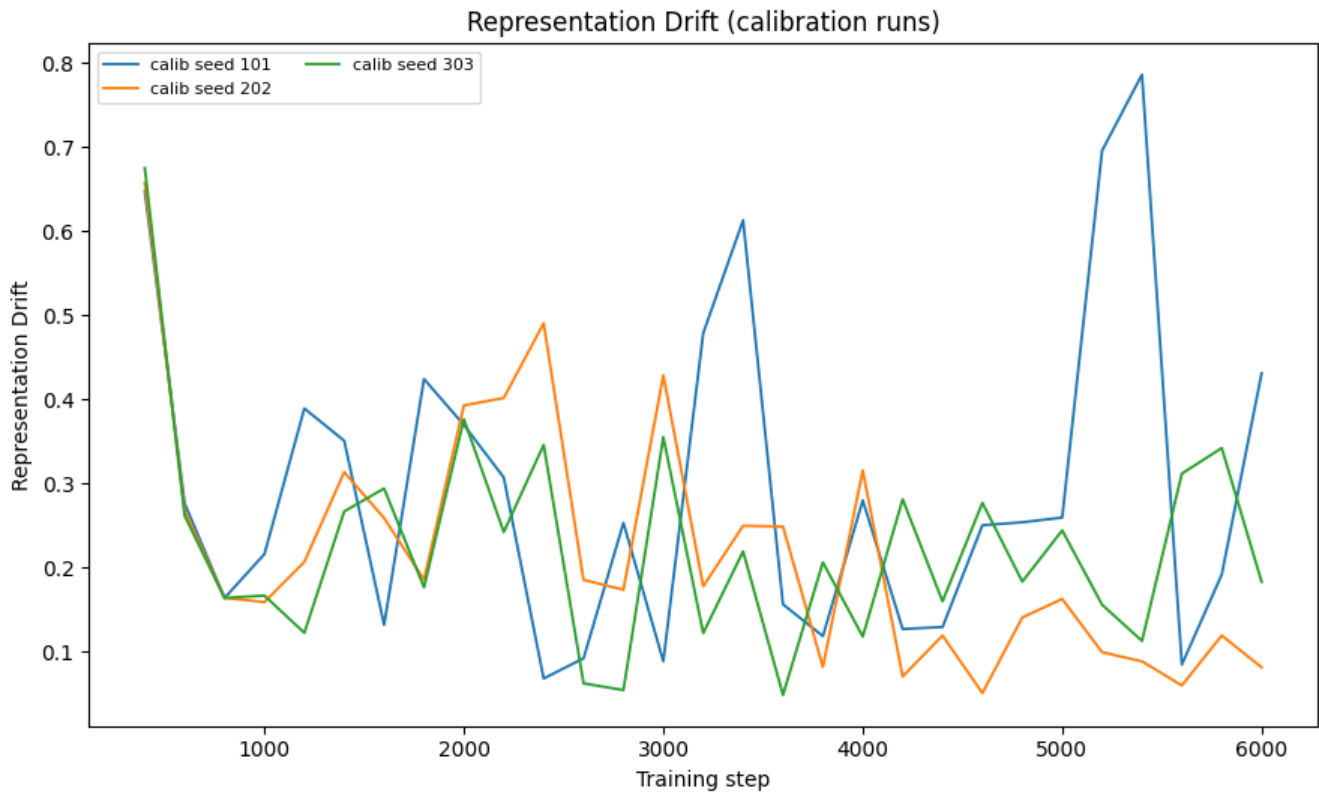


Figure 8. Representation-drift calibration runs across independent seeds.

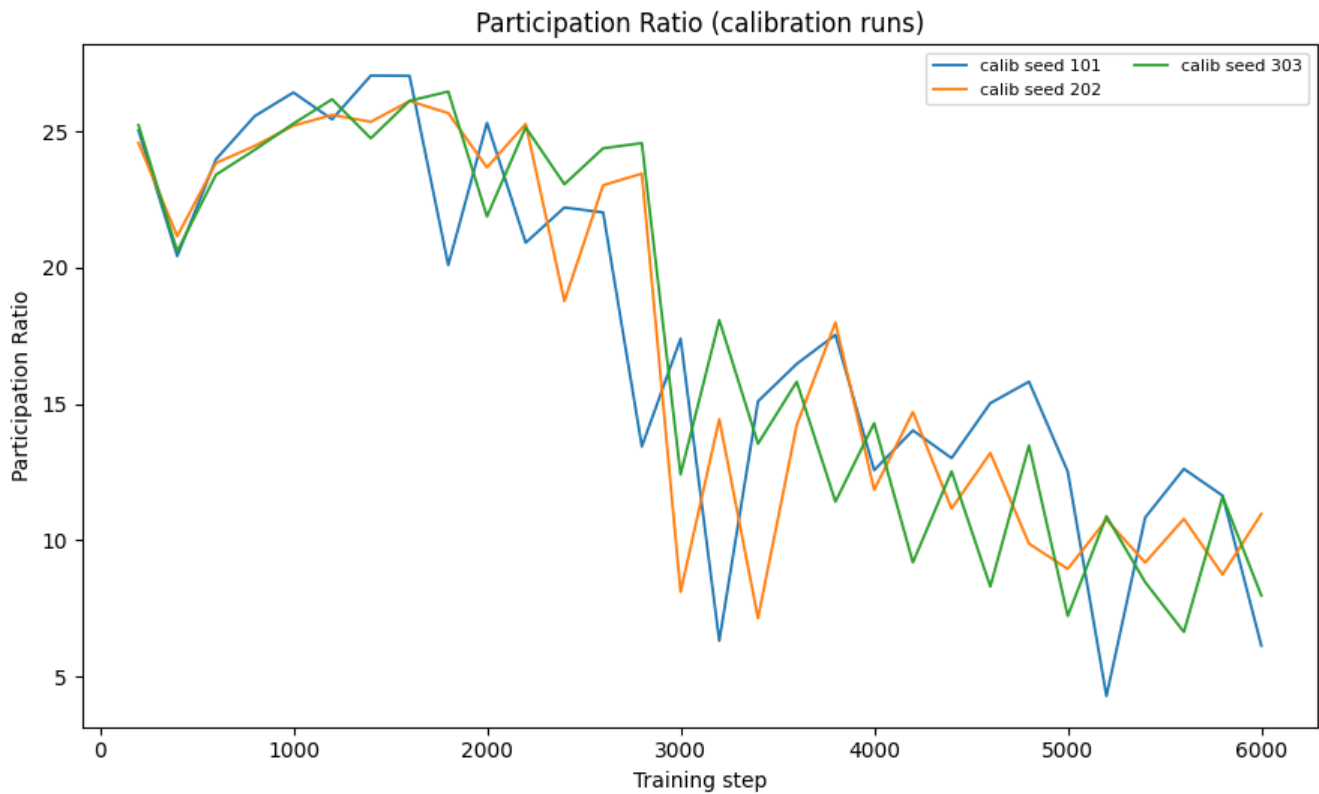


Figure 9. Participation-ratio calibration runs across independent seeds.

7. Predictive and mechanistic gates

Phase D failed because the protocol requires observed capability emergence plus positive lead time. Phase E consequently did not fit a warning model, and the warning monitor was not constructed. Mechanistic machinery—component discovery, ablation, activation patching, and backward checkpoint tracing—is present in the implementation but execution is gated on predictive evidence. This is a deliberate scientific result rather than a software failure.

8. LLC: measurement discipline

The v9 notebook successfully imports the authoritative *devinterp.slt.llc* backend and defines a fail-closed adapter. Configured validation includes independent chains, burn-in/thinning, chain agreement, effective-sample/autocorrelation diagnostics when available, budget sensitivity, temperature/localization sensitivity, and uncertainty. Critically, the controlled warning-shot run did not substitute a raw SGLD loss statistic for LLC. Where LLC is unavailable or fails validation, the field remains unavailable.

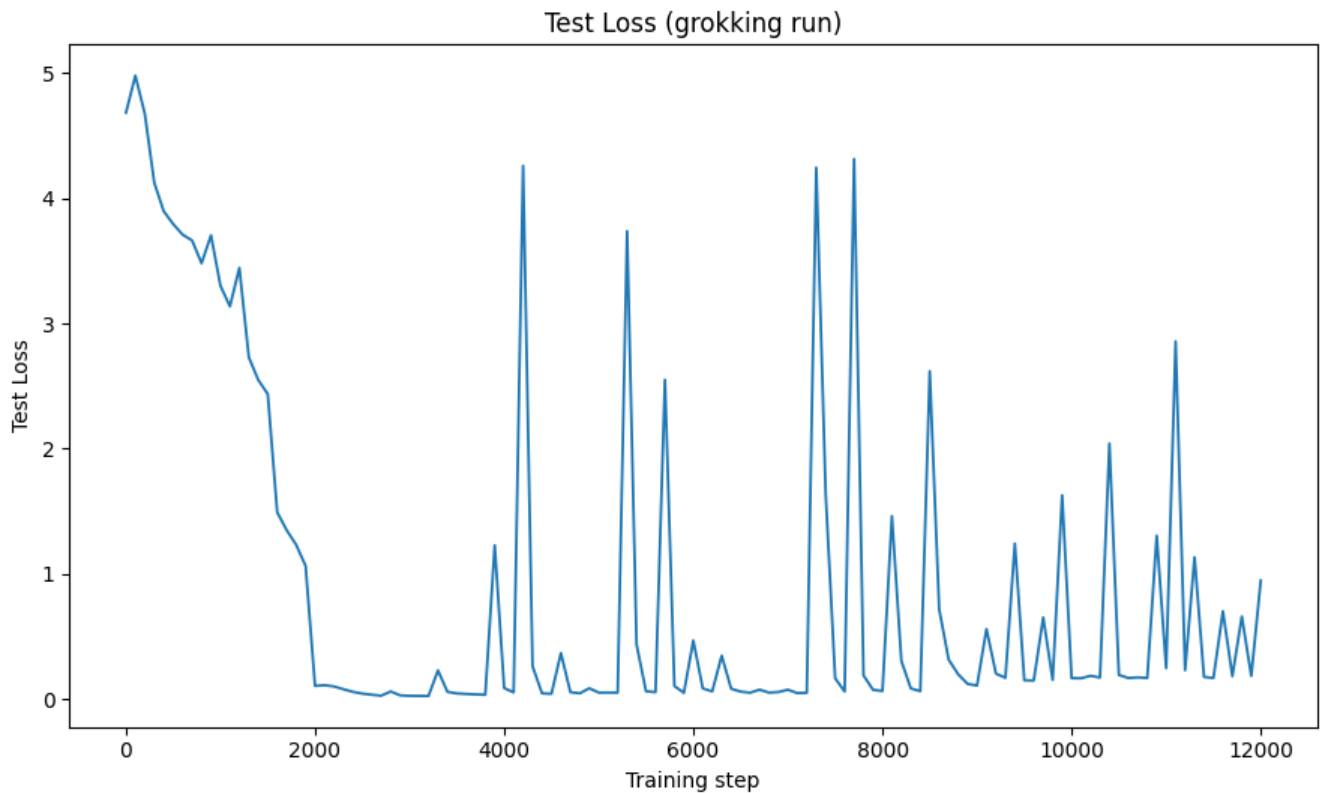


Figure 10. Test-loss grokking diagnostic retained as an execution trace.

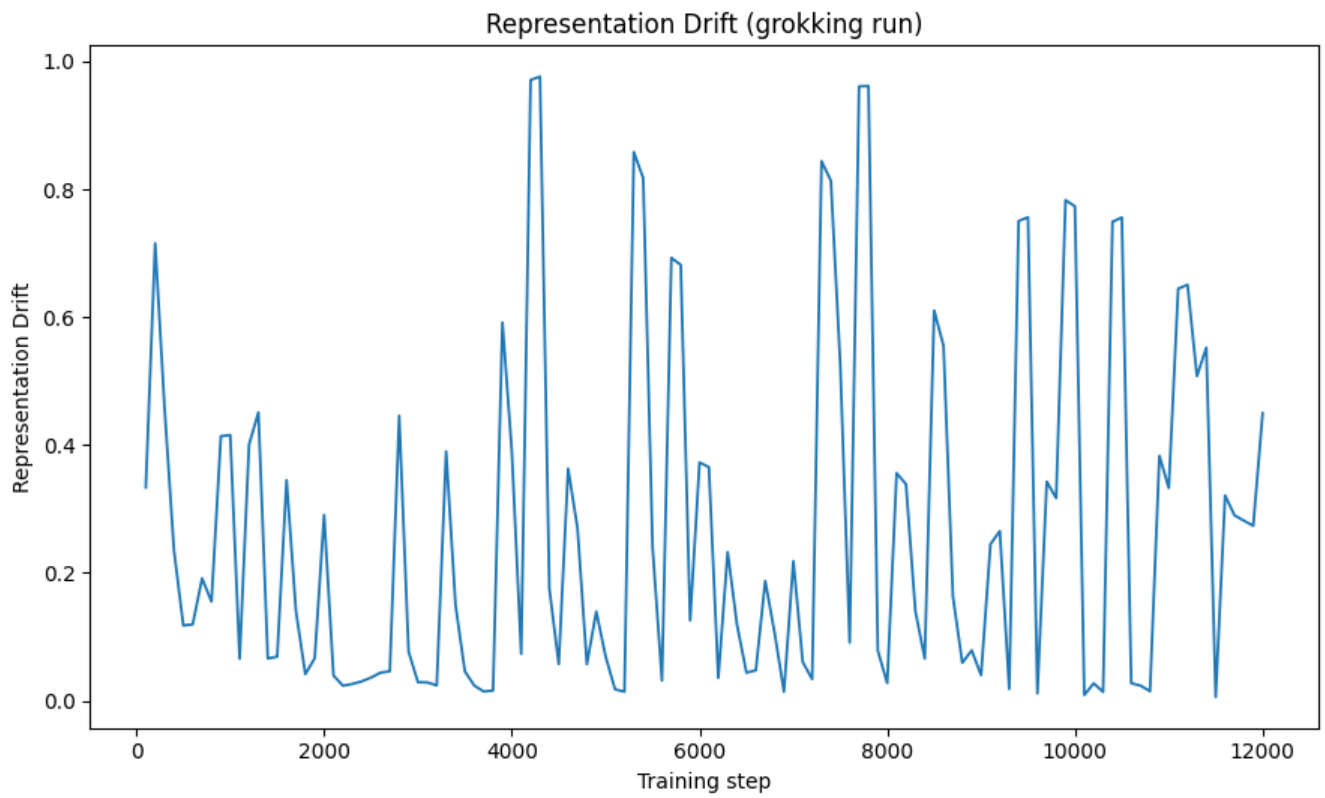


Figure 11. Representation-drift grokking diagnostic retained as an execution trace.

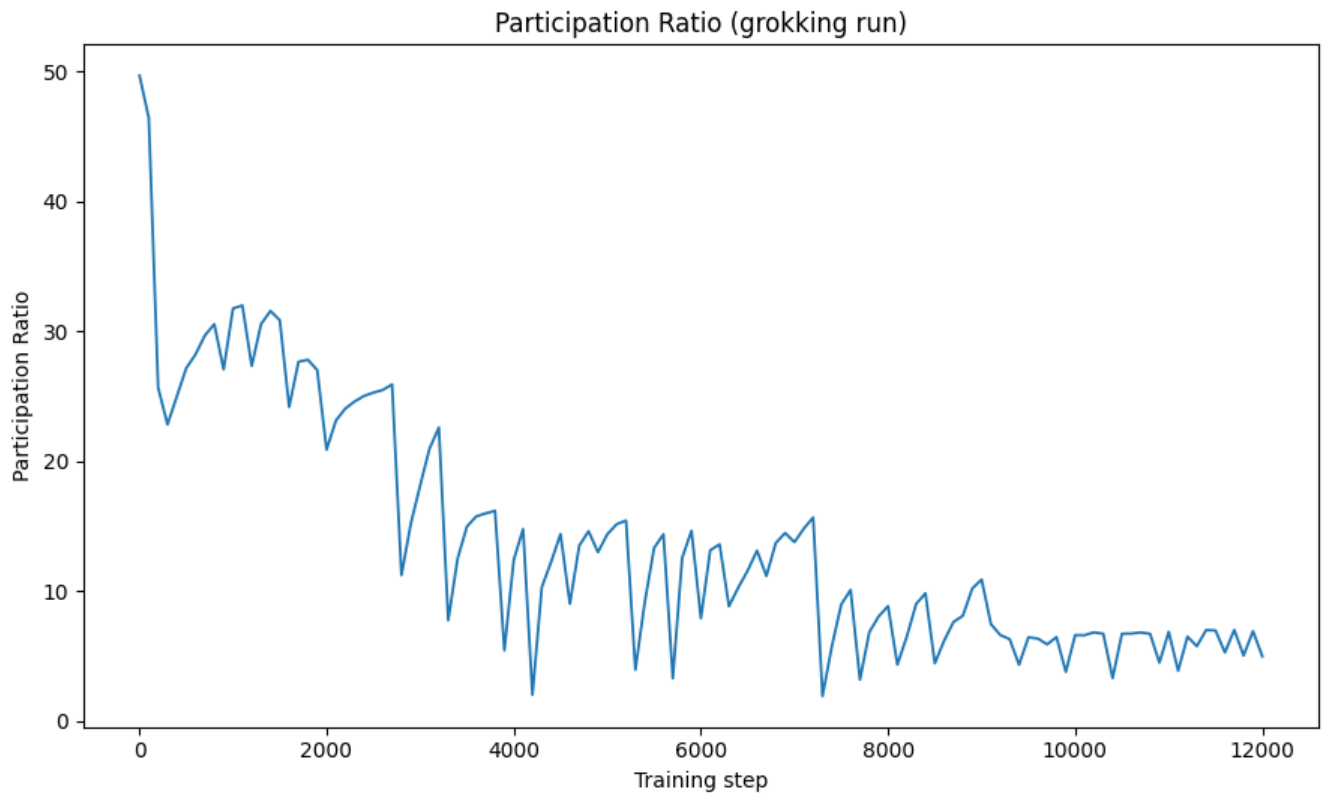


Figure 12. Participation-ratio grokking diagnostic retained as an execution trace.

9. External-validation architecture

The Hugging Face track is designed as external validity rather than a replacement for the controlled trajectory. Public developmental checkpoints (especially Pythia) provide the preferred temporal external track; final checkpoints from other families test geometric generality. Frozen warning definitions are calibrated only on development data, then applied without retuning to held-out families/scales/capabilities. The executed v9 artifact reports that no HF developmental trajectory or frozen external transfer result was successfully evaluated, so no transfer claim is made.

10. What the negative result means

The current evidence supports a narrower conclusion: the research and measurement infrastructure is operational enough to reproduce a known developmental transition and to run the intended controlled capability experiment, but the tested planning task did not emerge under the executed protocol. Therefore the central predictive claim remains untested/unsupported in this run. A negative result is scientifically informative: it prevents an internal metric from being promoted to a warning signal without the event it was supposed to precede.

11. Belief and domain-expertise robustness extension

The final extension asks a related but distinct question: during fine-tuning, do internal-complexity changes accompany belief robustness or domain overreach under adversarial prompting? Track A defines 10 belief types with 14 training statements and 20 adversarial prompts per belief. Track B defines five domain-pair settings, with domain statements, 50 out-of-domain concept statements, and 20 overreach prompts per pair. The executed notebook is explicitly a smoke configuration: two belief types and one domain pair, using GPT-2 smoke, two fine-tuning epochs and small LLC sampling budgets.

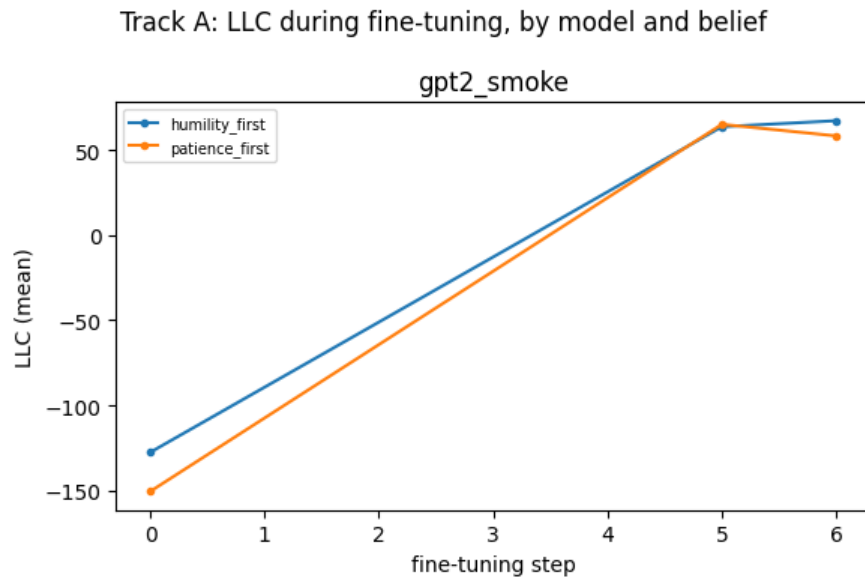


Figure 13. Belief-robustness smoke diagnostic: local LLC trajectory for the executed belief subset.

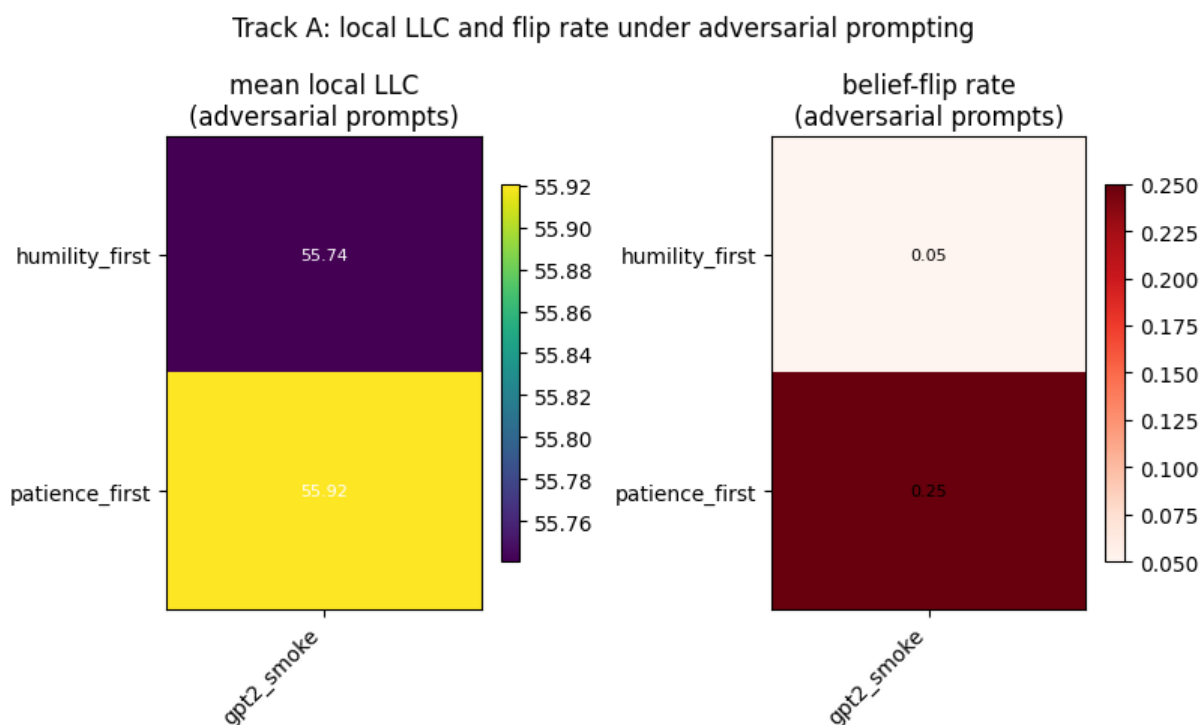


Figure 14. Belief robustness/adversarial prompting summary from the executed smoke configuration.

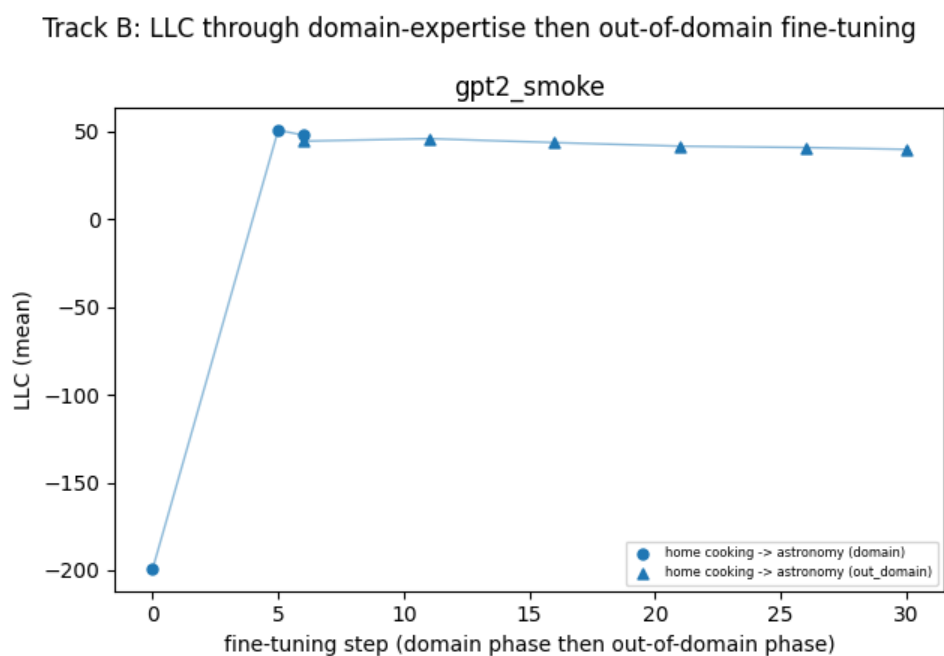


Figure 15. Domain-overreach smoke diagnostic for the executed domain pair.

Track B: local LLC and overreach rate under adversarial prompting

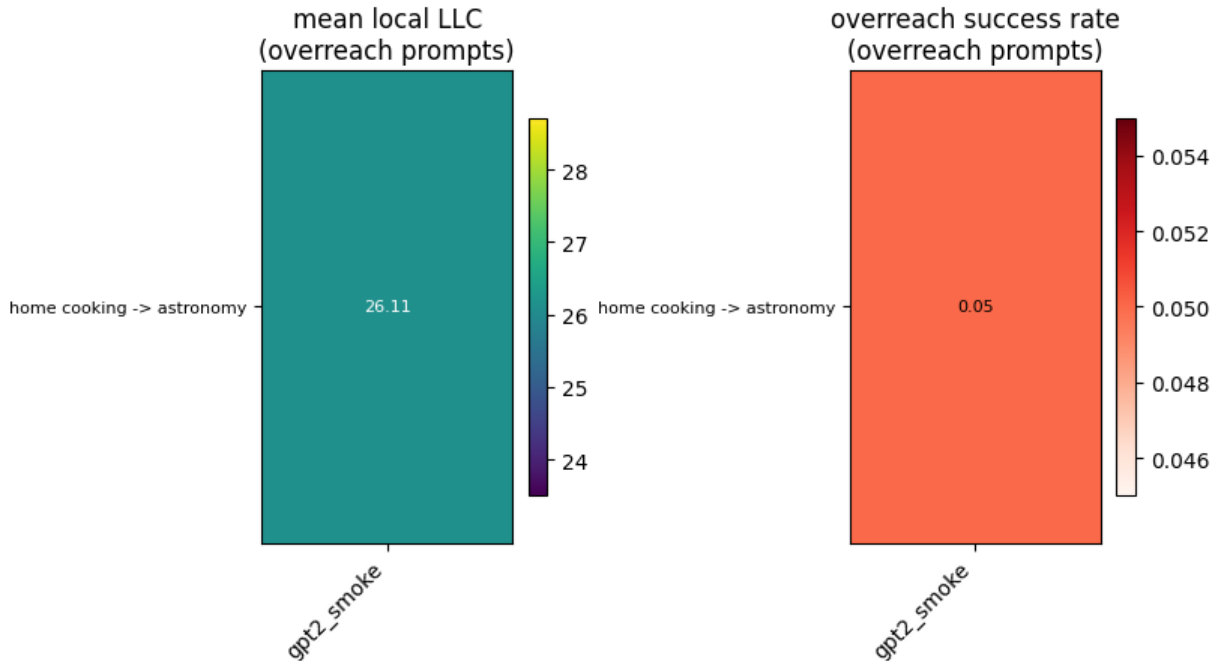


Figure 16. Local-LLC/domain robustness summary for the executed smoke configuration.

Executed smoke observations are descriptive only: 20 valid local-LLC points per evaluated belief/model; belief flip rates of 0.05 for *humility_first* and 0.25 for *patience_first*; and, for the home-cooking → astronomy pair, 20 valid local-LLC points with mean local LLC ≈ 26.11 and overreach rate 0.05. These observations do not support general claims over the declared matrix.

12. Limitations and threats to validity

(1) No capability onset. The central warning-shot hypothesis is not demonstrated because planning capability remained at zero. **(2) Task scope.** Multi-step planning is a controlled capability analogue, not evidence about real-world strategic or harmful behavior. **(3) Statistical power.** Five capability seeds and the current calibration design are insufficient for broad developmental-law claims. **(4) LLC coverage.** Although the real estimator is implemented and validated in principle, the main warning-shot run did not generate an accepted LLC trajectory. **(5) External validity.** The HF pipeline is designed for transfer, but the executed artifact records no completed developmental HF transfer. **(6) Smoke extension.** The belief/domain notebook runs only a small subset of the declared matrix. **(7) Causality.** No causal circuit claim is warranted without a behavioral onset, predictive evidence, controlled intervention, and backward temporal trace.

13. Reproducibility and next empirical step

The next scientifically clean step is not to loosen the onset threshold until a positive result appears. Instead: repeat the controlled planning experiment with a pre-specified training/task budget that is demonstrably sufficient for capability emergence; retain all five seeds and add independent seeds; freeze all internal-metric definitions before outcome inspection; then test whether any metric has positive lead time against loss-only and behavioral baselines. If onset occurs, run the pre-specified causal and backward-trace gates. Separately, expand the belief/domain experiment from smoke mode to its declared matrix, with fixed prompts, independent model seeds, and uncertainty intervals.

14. Conclusion

This project is best characterized as a scientifically gated measurement framework with a negative controlled capability outcome, not as evidence that a particular internal metric detects emergent danger. Its strongest contribution in the present state is methodological: it makes positive claims contingent on independent behavioral onset, temporal precedence, cross-seed validation, and—where warranted—causal intervention. The implementation also demonstrates substantial T4/HPC engineering discipline and a path toward real-LLM external validation. The correct scientific status is therefore: **calibration validated; controlled capability emergence not observed; warning signal not established; belief/domain extension exploratory.**

Figure handling. All 16 executed plots are embedded in the main report rather than moved to an appendix. Each is displayed at approximately 0.75× its native pixel dimensions, subject only to page-width constraints, preserving substantially more visual detail than the earlier merged contact-sheet figures.

References

References added without changing the substantive report content.

1. Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, D. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. arXiv:2201.02177.
<https://arxiv.org/abs/2201.02177>
2. Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. Proceedings of ICML, 3519–3529.
<https://proceedings.mlr.press/v97/kornblith19a.html>
3. Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent abilities of large language models. arXiv:2206.07682.
<https://arxiv.org/abs/2206.07682>
4. Biderman, S., Schoelkopf, H., Anthony, Q. G., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. Proceedings of ICML, 2397–2430.
<https://proceedings.mlr.press/v202/biderman23a.html>
5. D'Ascoli, S., Sagun, L., Biroli, G., & Goldt, S. (2021). Triple descent and the two kinds of overfitting in neural networks. Proceedings of ICML, 108–118.
<https://proceedings.mlr.press/v139/d-ascoli21a.html>
6. Fort, S., Scherlis, A., Rajan, A., et al. (2020). Deep learning versus kernel learning: An empirical study of loss landscape geometry and generalization. arXiv:2010.12972.
<https://arxiv.org/abs/2010.12972>
7. Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. arXiv:2001.08361.
<https://arxiv.org/abs/2001.08361>
8. Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. NeurIPS 35.
<https://arxiv.org/abs/2201.11903>
9. Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2), 157–166.
<https://doi.org/10.1109/72.279181>