From Actions to Understanding: Conformal Interpretability of Temporal Concepts in LLM Agents

T. Padhi*, R. Kaur*, K. Agarwal, A. D. Cobb, D. Elenius, M. Acharya, C. Samplawski, A. M. Berenbeim, N. D. Bastian, S. Jha, U. Kursuncu, A. Roy
Georgia State University · SRI International · University of Florida · USMA West Point

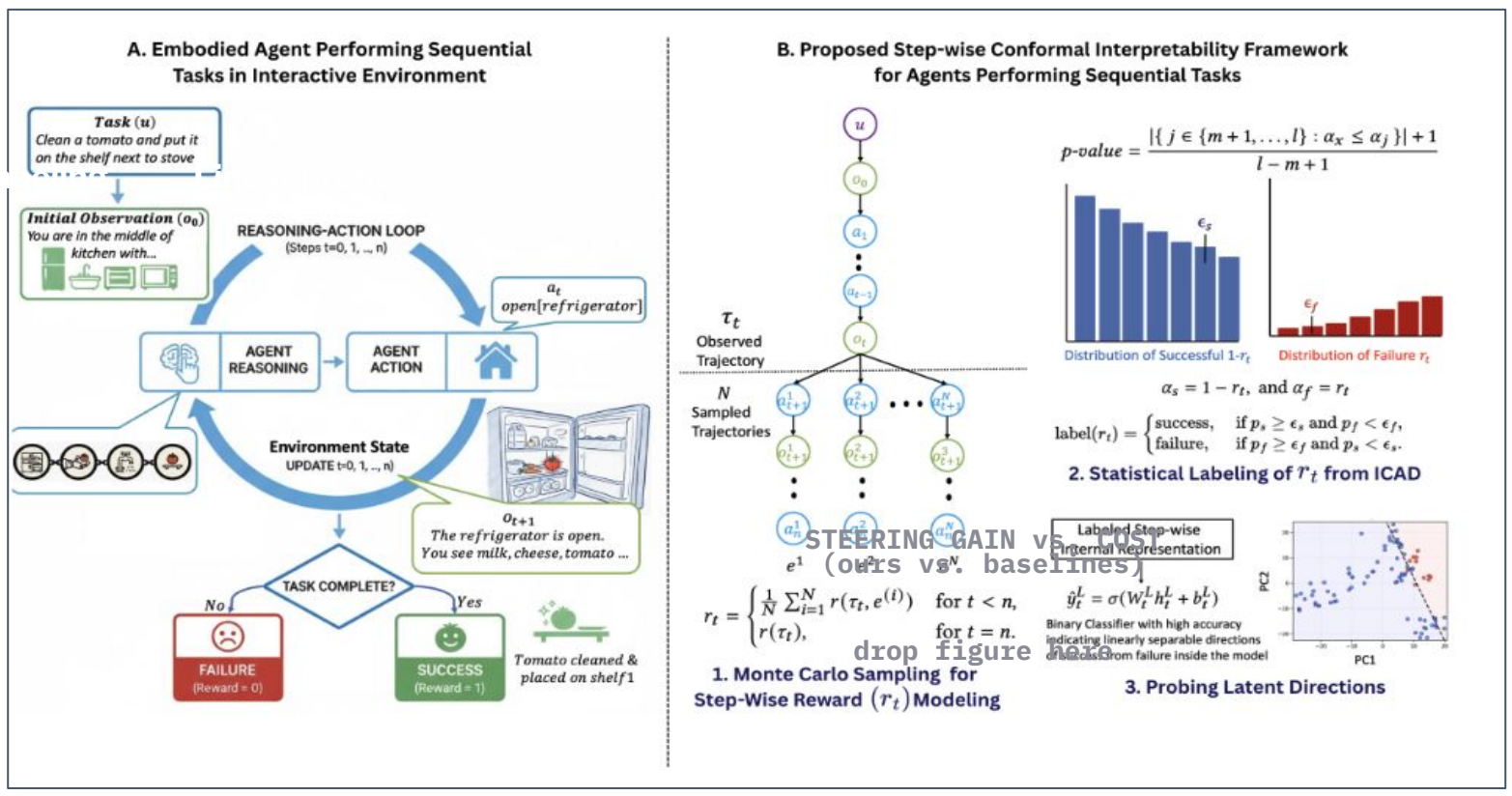
Takeaway: You can read success or failure off an LLM agent's hidden state at every step catching *reasoning drift early and steering it back on track*.

Motivation & Background

LLM agents plan and act over many steps, but they stay black boxes. A task that ultimately fails may contain many correct steps; a success may hide accidental recoveries. Static interpretability looks at single inputs—so it cannot say *when* an agent drifts from success toward failure.

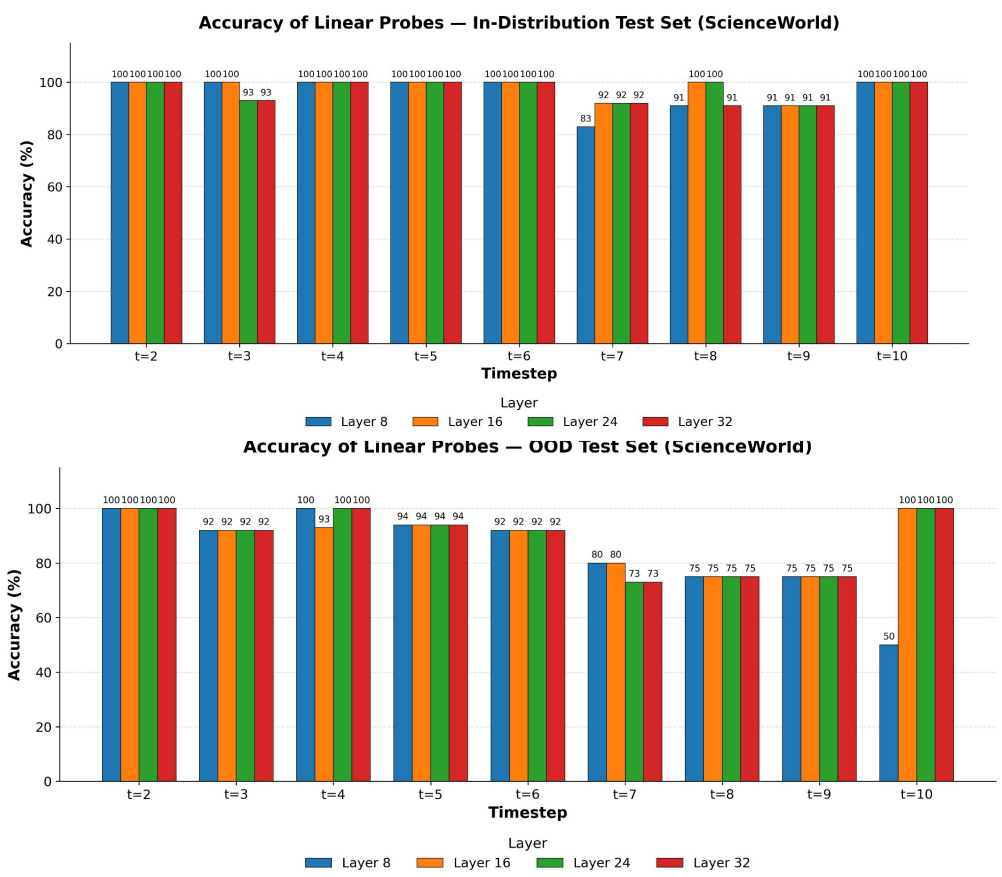
Methodology

We use Monte-Carlo rollouts to convert sparse terminal rewards into dense per-step signals, apply Inductive Conformal Prediction to label each step's success/failure with error bounded at 10%, then train linear probes on Llama-2-7B activations to test concept separability across layers and time.



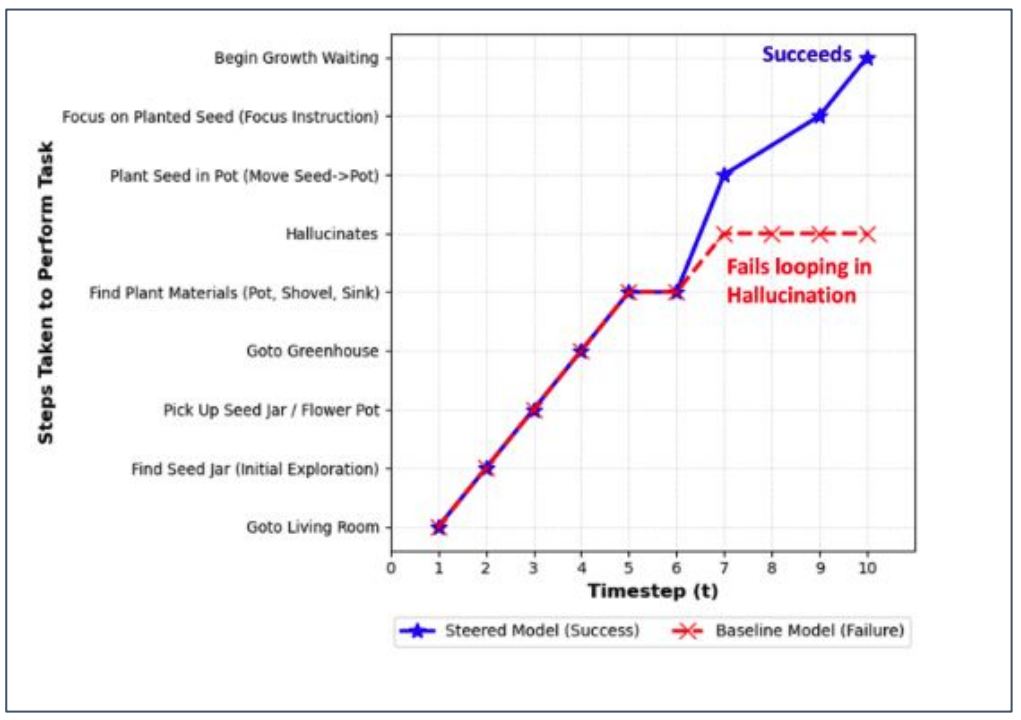
Result 1 - Success & Failure Concepts are Linearly Separable

Linear probes separate successful from failing steps with up to 100% accuracy (ScienceWorld))—across timesteps, layers, and in/out-of-distribution domains.



Result 2 - Can we steer using the identified failures towards success? YES

One early intervention (step 3, RepE, coeff. 0.025) gives +1.1% accuracy—competitive with far costlier Best-of-N (+2.8%), RFT (+4.2%), and DPO (+6.8%).



References

Zou, Andy, et al. "Representation engineering: A top-down approach to ai transparency." arXiv preprint arXiv:2310.01405 (2023).

Acknowledgements

This material is based on work supported by the United States Air Force and Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-23-C-0519, the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR0011-24-9-0424

SCAN ME