On Logical Extrapolation for Mazes with Recurrent and Implicit Networks

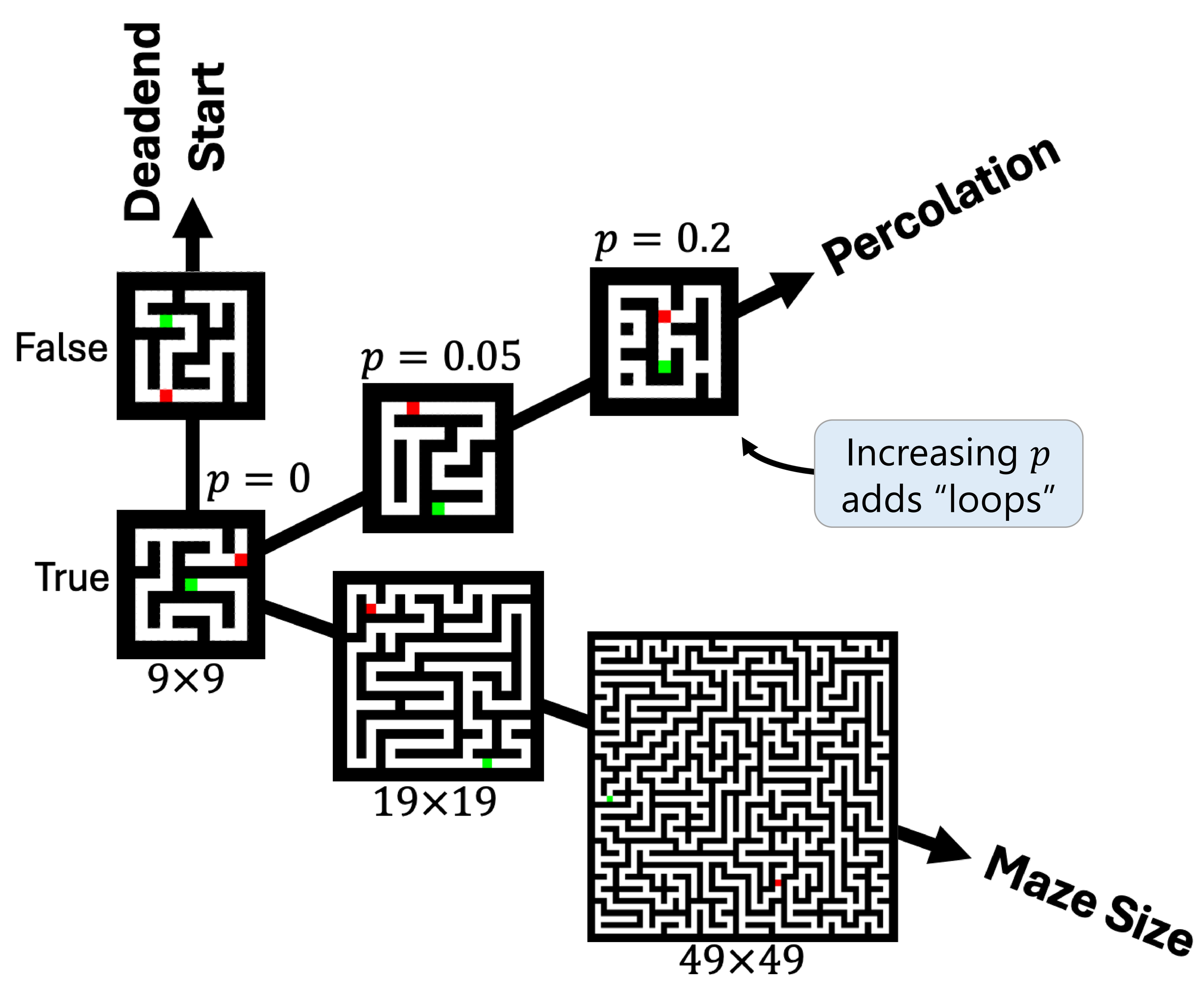
Brandon Knutson, *Amandin Chyba Rabeendran, Michael Ivanitskiy, Jordan Pettyjohn, Cecilia Diniz Behn, Samy Wu Fung, Daniel McKenzie*

Introduction

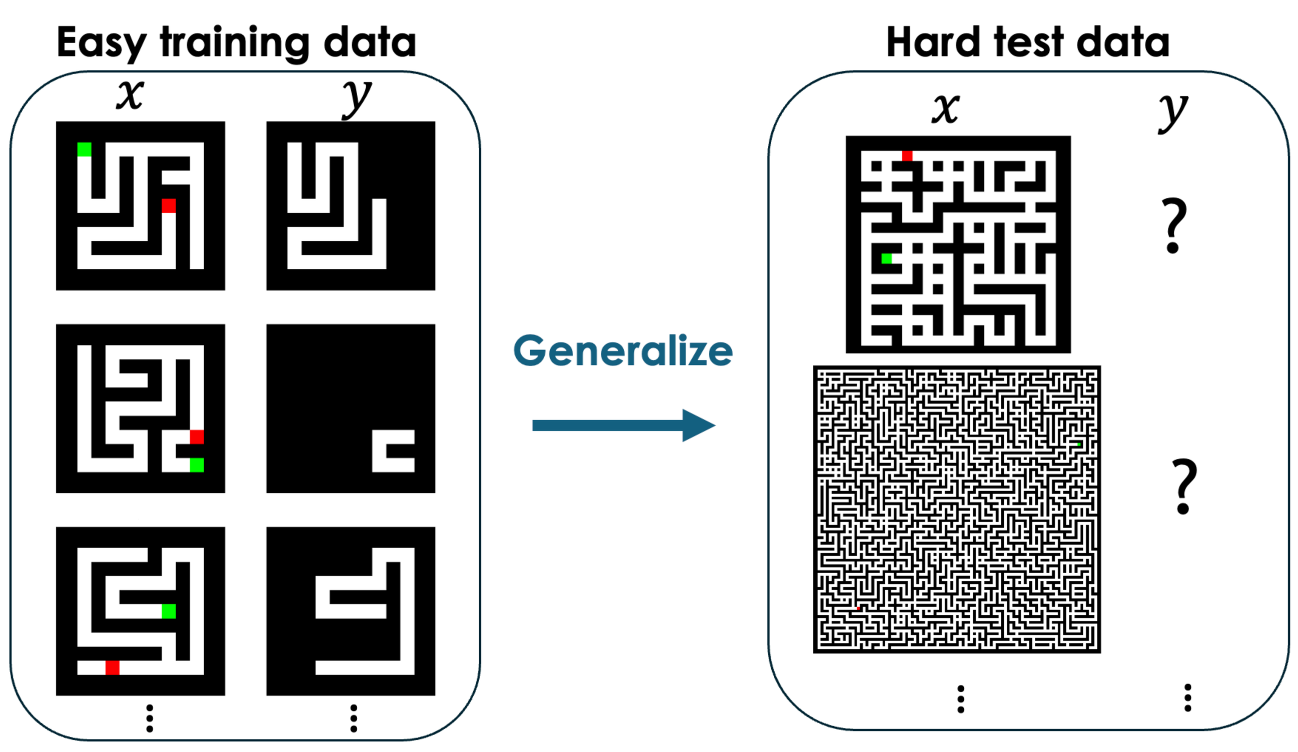
- **Logical extrapolation** is generalization from easy to hard problems in structured tasks, a key feature of human intelligence.
- **Mazes are an ideal testbed.** Diverse generation and solving algorithms let us simulate controlled out-of-distribution (OOD) shifts while retaining ground-truth solutions.
- **Recurrent and implicit neural networks** (RNNs/INNs) exhibit **test-time scaling**: they can "think longer" at inference, making them well-suited for logical extrapolation.

Methods

- We consider three OOD shifts on mazes: **size**, **percolation**, and **deadend start**.



- Using an RNN and INN, we **train on easy data** and **test on hard data** to evaluate logical extrapolation ability.

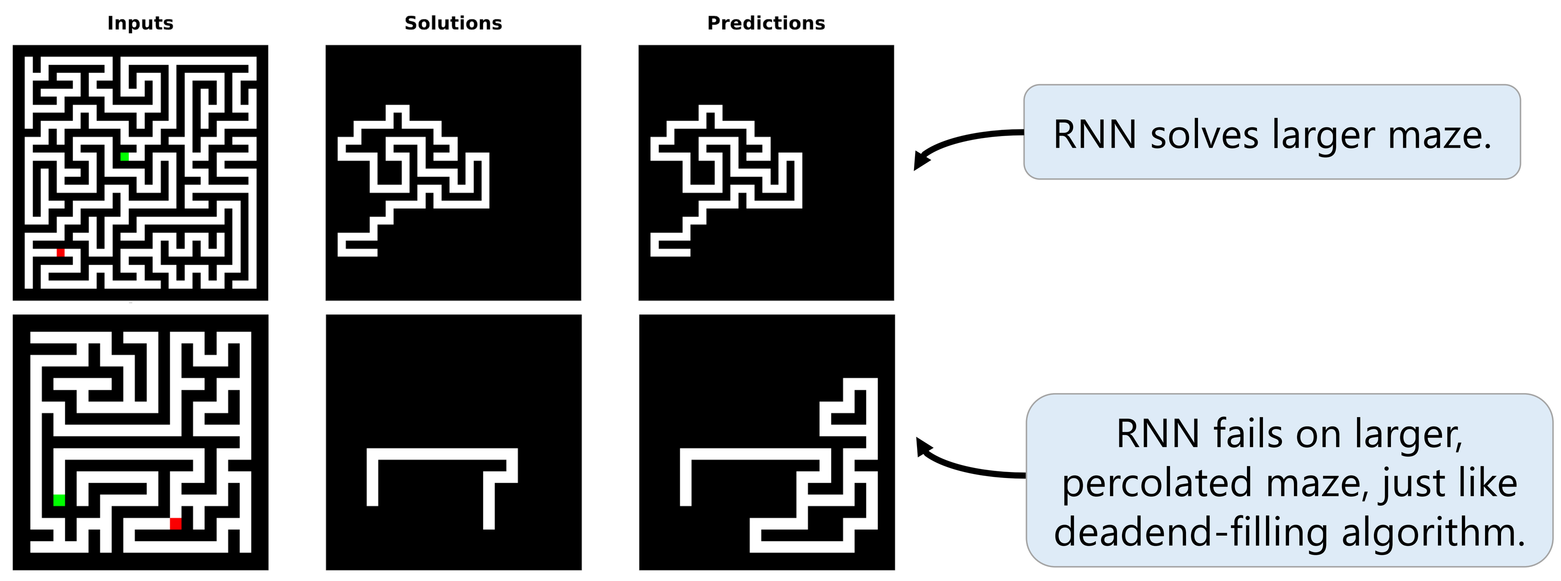


- We further study algorithmic behavior, recurrent convergence, and the effect of training-data diversification.

Results

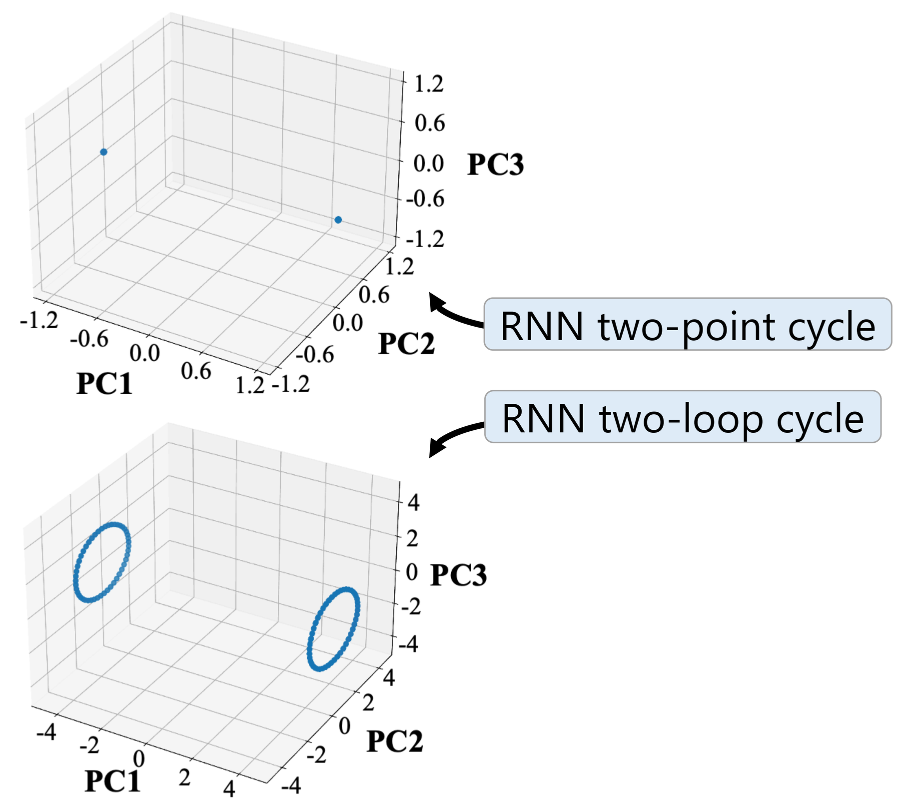
1 Do RNN/INNs learn a known maze-solving algorithm?

The RNN approximately learned the **deadend-filling algorithm**.



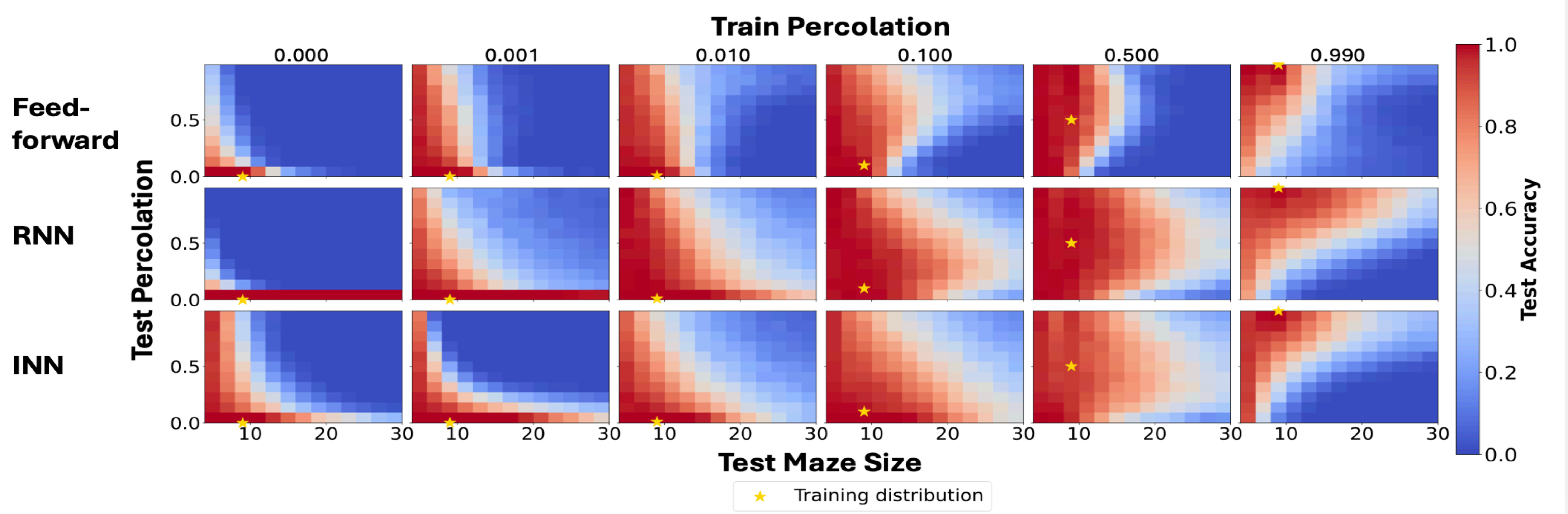
2 Are the recurrent iterations converging? To what?

The INN always converges to a **fixed point**. The RNN often forms **cycles**, showing fixed-point convergence is not necessary for logical extrapolation.



3 How does diversifying the training data affect logical extrapolation?

Slight diversification improves performance on percolated mazes. But resulting models **don't appear algorithmic**, and **size extrapolation is harmed**.



Scan to read the **paper**

COLORADO SCHOOL OF MINES
NSF
Association for the Advancement of Artificial Intelligence