Evaluating the World Model Implicit in a Generative Model

Yanjian Zhang

LIPN

09/12/2024

Catalogue

Motivation

Metrics

Task

Result

Motivation

- LLMs can perform some of task will without having a coherent world model
- Build up a theoretically-grounded metrics evaluating whether Language model capture accurate world models.

Annotation - LM

- Finite token set: $a \in \Sigma$
- Sequence: $s = (a_1, a_2, \ldots)$
- Generative models: $\Sigma^* \rightarrow \Delta(\Sigma)$
  - Σ^* Represents the set of all finite strings (sequences) over the Σ
  - $\Delta(\Sigma)$ Denotes the set of all probability distributions over the Σ

- Sequence with positive probability:

$$L^m(s) = \{a_1 a_2 \ldots a_k : \forall j < k,\ m(a_{j+1} \mid s a_1 \ldots a_j) > 0\}$$

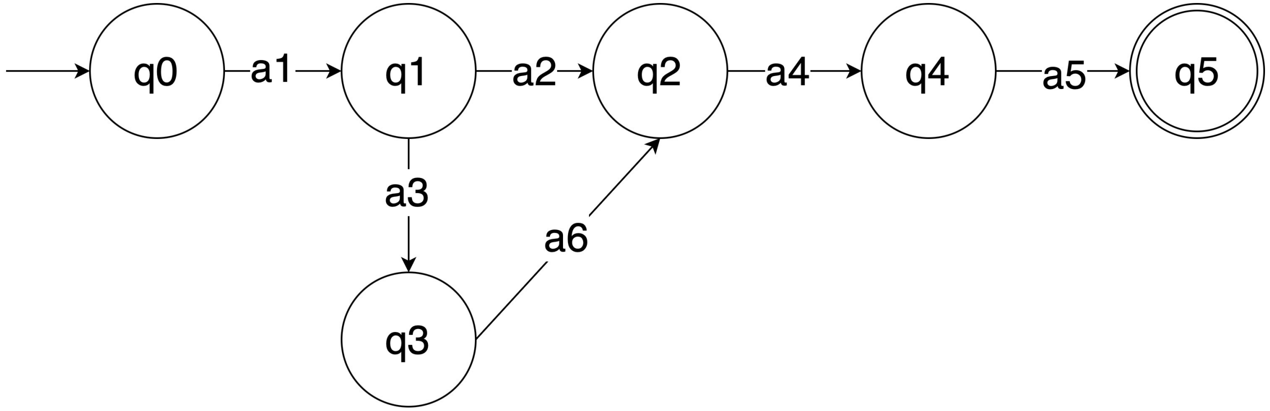
Annotation - DFA

- Deterministic finite automata: $W = (Q, \Sigma, \delta, q_0, F)$

1. Q is a finite set of states,
2. Σ is a finite set of characters,
3. $\delta : Q \times \Sigma \to Q$ is the transition function mapping a state and character to the next state,
4. $q_0 \in Q$ is the start state,
5. $F \subseteq Q$ is the set of accepting states.

- Valid sequence accepted by DFA starting at q : $L^W(q)$
- Collection of sequences leading from state q0 to q in the DFA: $S(q) \subseteq \Sigma^*$

Annotation Example



$$W = (Q, \Sigma, \delta, q_0, F)$$

$Q = \{q_0, q_1, q_2, q_3, q_4, q_5\}$
$\Sigma = \{a_1, a_2, a_3, a_4, a_5, a_6\}$
$\delta: transition\ function$
$q_0: q_0 \ \ \mathrm{q}_{\{\mathrm{reject}\}}: \mathrm{q}_5$
$F: \{q_0, q_1, q_2, q_3, q_4\}$
Assume $q = q_2,\ L^W(q) = \{a_4\}$
$S(q) = \{(a_1, a_2), (a_1, a_3, a_6)\}$

- Deterministic finite automata: $W = (Q, \Sigma, \delta, q_0, F)$
- Valid sequence accepted by DFA starting at q : $L^W(q)$
- Collection of sequences leading from state q0 to q in the DFA: $S(q) \subseteq \Sigma^*$

Two definition of Recovering world models

- Definition 1:

A generative model $m(\cdot)$ **recovers the DFA** W if

$$\forall q \in F, \forall s \in S(q):\ L^{W}(q) = L^{m}(s).$$

- Definition 2:

A generative model $m(\cdot)$ satisfies **exact next-token prediction** under the DFA W if

$$\forall q \in F, \forall s \in S(q), \forall a \in \Sigma:\ m(a \mid s) > 0 \iff \delta(q, a) \neq q_{\text{reject}}.$$

The Myhill-Nerode interior and boundary

The Myhill-Nerode theorem: the sets of sequences accepted by a minimal DFA starting at two distinct states are distinct.

However, while distinct, the two sets may exhibit a great deal of overlap.

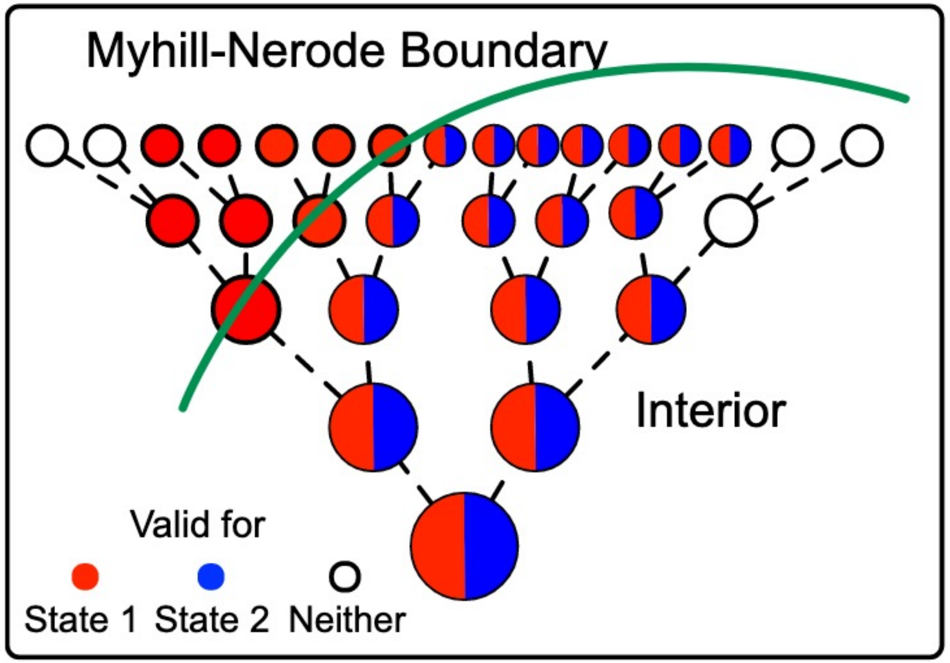
Definition 2.4. Given a DFA W, the **Myhill-Nerode interior** for the pair $q_1, q_2 \in F$ is the set of sequences accepted when starting at both states:

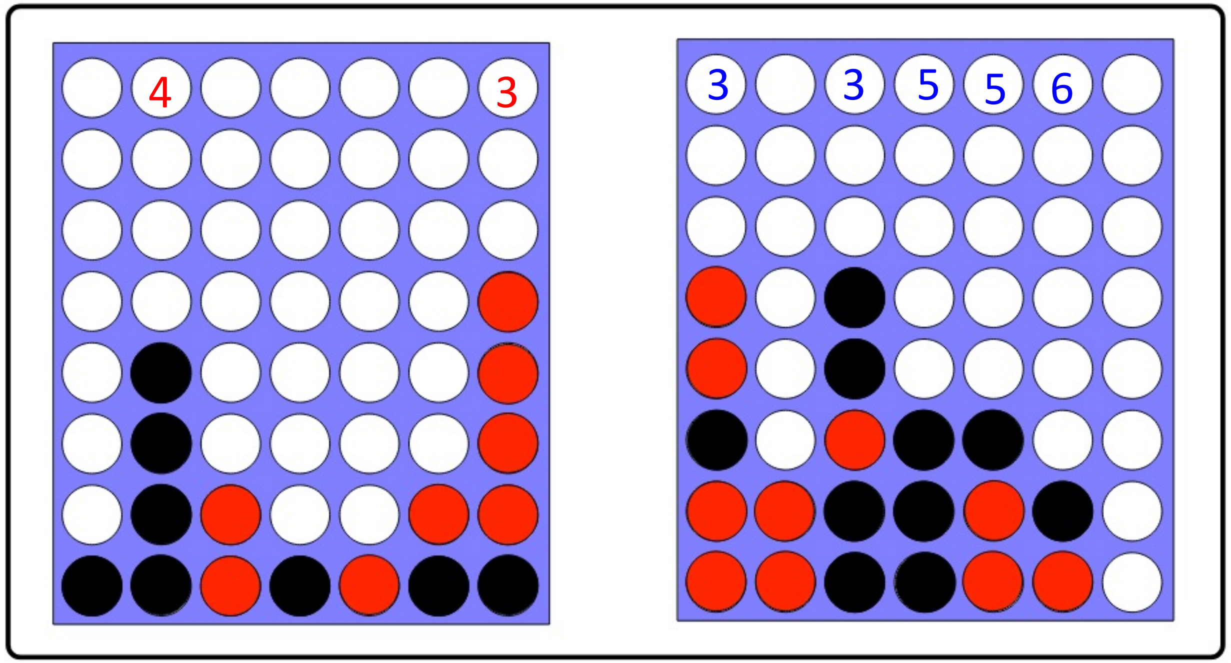
$$\mathrm{MNI}^W(q_1, q_2) = \{s \in \Sigma^* \mid s \in L^W(q_1) \cap L^W(q_2)\}.$$

The **Myhill-Nerode boundary** is the set of minimal suffixes accepted by a DFA at q_1 but not q_2:

$$\mathrm{MNB}^W(q_1, q_2) = \{s = a_1 a_2 \ldots a_k \mid s \in L^W(q_1) \setminus L^W(q_2) \text{ and } \forall j < k : a_1 \ldots a_j \in \mathrm{MNI}^W(q_1, q_2)\}.$$

Figure illustration in Connect-4





The interior contains about 8.8 * 10^27 of lenght 29(3+4+3+5+5+6+3) do not distinguish the two boards.

For model?

Definition 2.5. For two sequences s_1, s_2, the **Myhill-Nerode boundary implied by model** $m(\cdot)$ is

$$\mathrm{MNB}^m(s_1, s_2) = \{x = x_1 \ldots x_k \mid x \in L^m(s_1) \setminus L^m(s_2) \text{ and } \forall j < k : x_1 \ldots x_j \in L^m(s_1) \cap L^m(s_2)\}. \quad (1)$$

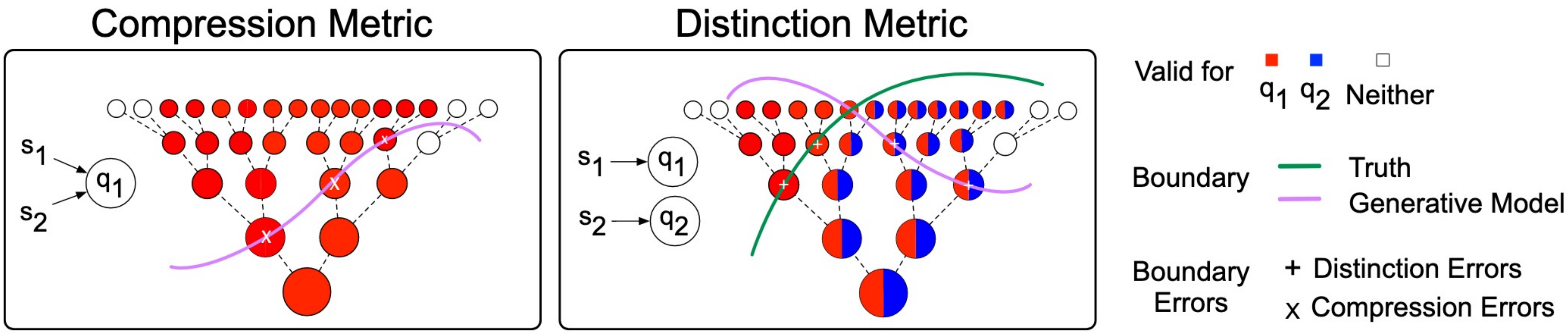
Definition 2.6. The **boundary recall** of generative model $m(\cdot)$ with respect to a DFA W is defined as

$$\frac{|\mathrm{MNB}^W(q_1, q_2) \cap (L^m(s_1) \setminus L^m(s_2))|}{|\mathrm{MNB}^W(q_1, q_2)|}, \quad (2)$$

and the **boundary precision** is defined as

$$\frac{|\mathrm{MNB}^m(s_1, s_2) \cap (L^W(q_1) \setminus L^W(q_2))|}{|\mathrm{MNB}^m(s_1, s_2)|}. \quad (3)$$

Metrics



Compression Metrics: Sample equal state pairs $q_1 = q_2$, summarize whether the generative model correctly compresses sequences that arrive at the same state under the DFA

Distinction Metric: Sample different state pairs $q_1 \neq q_2$, correctly distinguishes sequences that arrive at different states under the DFA

New York City example

- G = (intersection V, street E, distance W) $W : E \rightarrow \mathbb{R}^{+}$
- Edge is label based on cardinal direction: $D : V \times V \rightarrow \{\square, \mathsf{N}, \mathsf{S}, \mathsf{E}, \mathsf{W}, \mathsf{NE}, \mathsf{NW}, \mathsf{SE}, \mathsf{SW}\}$
- Three way of Traversals:
  - Shortest paths
  - Noisy shortest paths(add a gamma noise to approximate traffic condition)
  - Random walks

Model training setting

- Dataset
  - Include direction sequences with length less than 100
  - Train set: 2.9M sequences for shortest path, 31M for noise one, 91M for random works.
  - Test set 1000 sequence
- Model:
  - 117 M and 1.6 B GPT-2

Inference Result

- All models generate valid traversal: 96%-99%

- Calculate compression metrics:
  - Sample states with two distinct traversals and assess whether model correctly admit the same suffix for each prefix.
  - Average over pairs of prefixs for each state then averager over states.

- Calculate distinction metrics:
  - Sample distinct states and traversals, compute boundary recall and precision.

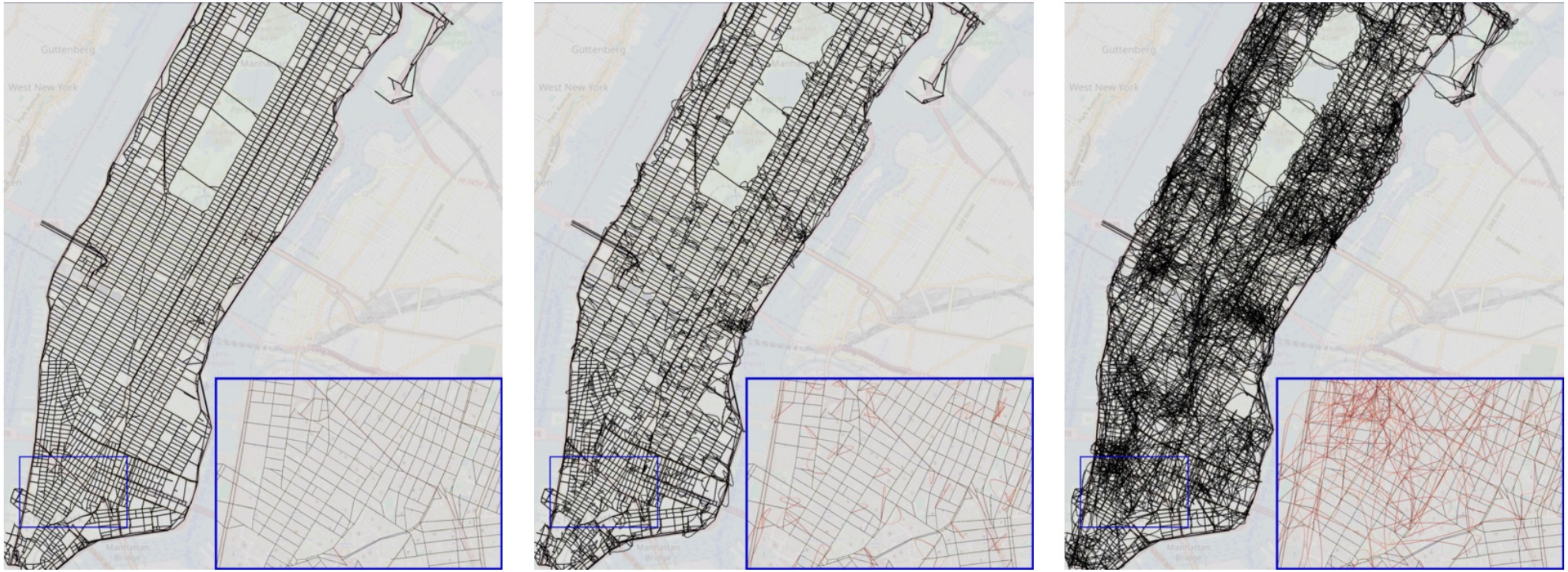
Result

	Existing metrics		Proposed metrics		
	Next-token test	Current state probe	Compression precision	Distinction precision	Distinction recall
Untrained transformer	0.03 (0.00)	0.10 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Shortest paths	1.00 (0.00)	0.91 (0.00)	0.10 (0.01)	0.35 (0.02)	0.20 (0.01)
Noisy shortest paths	1.00 (0.00)	0.92 (0.00)	0.05 (0.01)	0.37 (0.02)	0.24 (0.01)
Random walks	1.00 (0.00)	0.99 (0.00)	0.50 (0.02)	0.99 (0.00)	1.00 (0.00)
True world model	1.00	—	1.00	1.00	1.00

Table 1: Sequence compression and distinction metrics for world models compared to existing metrics (standard errors in parentheses). Models that do well on existing metrics can perform poorly on ours.

Baseline 2: Current state probe is to predict the current intersection with a trained linear probe

Reconstructing implicit map



(a) World model **(b)** World model with noise **(c)** Transformer

Sample 6400 origin-destination pairs, and plot the traversal

Detour fragility

		Probability of detour				
		0%	1%	10%	50%	75%
Random detours	Shortest paths	0.99 (0.01)	0.69 (0.05)	0.08 (0.03)	0.00 (0.00)	0.00 (0.00)
	Noisy shortest paths	0.96 (0.02)	0.52 (0.05)	0.03 (0.02)	0.00 (0.00)	0.00 (0.00)
	Random walks	0.99 (0.01)	0.99 (0.01)	1.00 (0.00)	0.97 (0.02)	0.74 (0.04)
	True world model	1.00	1.00	1.00	1.00	1.00
Adversarial detours	Shortest paths	0.99 (0.01)	0.66 (0.05)	0.06 (0.02)	0.00 (0.00)	0.00 (0.00)
	Noisy shortest paths	0.96 (0.02)	0.64 (0.05)	0.04 (0.02)	0.00 (0.00)	0.00 (0.00)
	Random walks	0.99 (0.01)	1.00 (0.00)	1.00 (0.00)	0.93 (0.03)	0.51 (0.05)
	True world model	1.00	1.00	1.00	1.00	1.00

Table 2: The fraction of traversals that are valid when detours are introduced (standard errors in parentheses).

- Random detour: the model's proposed token is replaced with a randomly chosen (true) valid token
- Adversarial detour: it is replaced with the model's lowest ranked valid token.

Logic puzzles

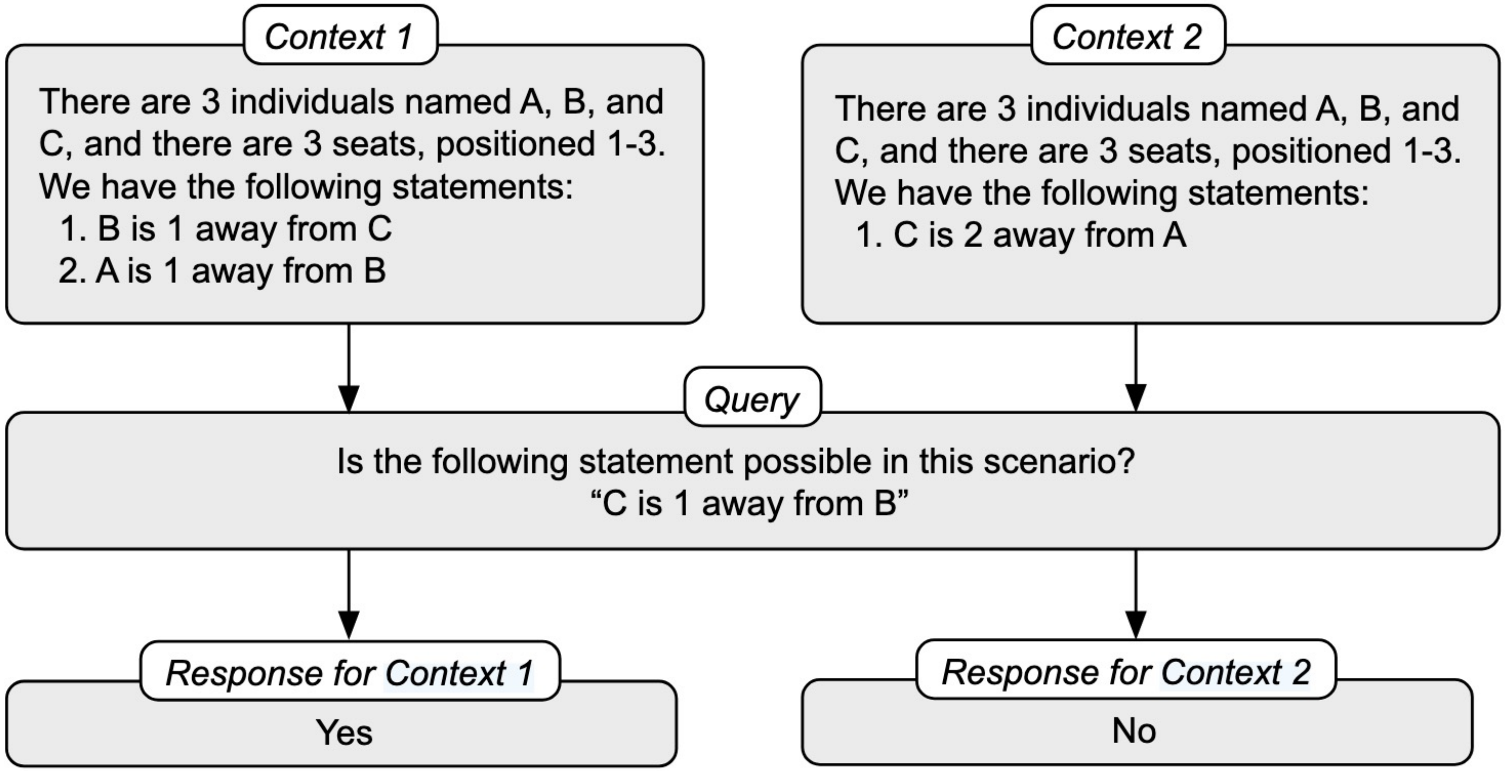


Figure 8: An example of a compression error for GPT-4 on the logic puzzle test. The model is prompted with statements that correspond to the same underlying state and a sample continuation. It assesses that the continuation is valid for one state yet invalid for the other.

LLMs performance in logic puzzle

Example task prompt

There are 3 individuals named A, B, and C, and there are 3 seats, positioned 1-3. We have the following statements:

1. B is in seat 3
2. B is 1 seat away from A

Based on this information, where is C seated? You can use chain-of-thought reasoning.

	Capabilities	Proposed metrics	
	Task accuracy	Compression precision	Distinction recall
Llama-2 (70B)	0.77 (0.03)	0.08 (0.03)	0.42 (0.04)
Llama-3 (8B)	0.85 (0.02)	0.18 (0.04)	0.23 (0.03)
Llama-3 (70B)	0.98 (0.00)	0.25 (0.04)	0.57 (0.04)
Mixtral-8x22B	0.88 (0.01)	0.35 (0.05)	0.57 (0.05)
Qwen 1.5 (72B)	0.88 (0.02)	0.21 (0.04)	0.56 (0.03)
Qwen 1.5 (110B)	0.98 (0.00)	0.53 (0.05)	0.53 (0.04)
GPT-3.5 (turbo)	0.83 (0.02)	0.33 (0.05)	0.18 (0.03)
GPT-4	1.00 (0.00)	0.21 (0.04)	0.56 (0.03)
True world model	1.00	1.00	1.00

Conclusion and Limitation

- Generative model could performan impressively without a coherent model
- But incoherence make them fragile for other tasks involving detours.
- This work only focus on DFA, richer setting could be explored.

Thank you