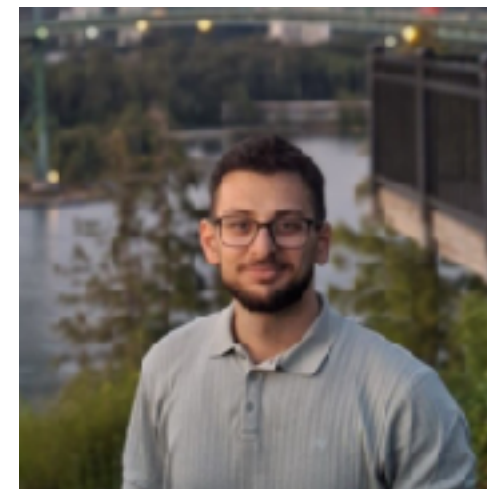


# Dynamical System Perspectives of LLMs

Guannan Qu  
Associate Professor, CMU

Joint work with Muhammed Ustaömeroğlu (CMU)



European Control Conference  
July 10, 2026

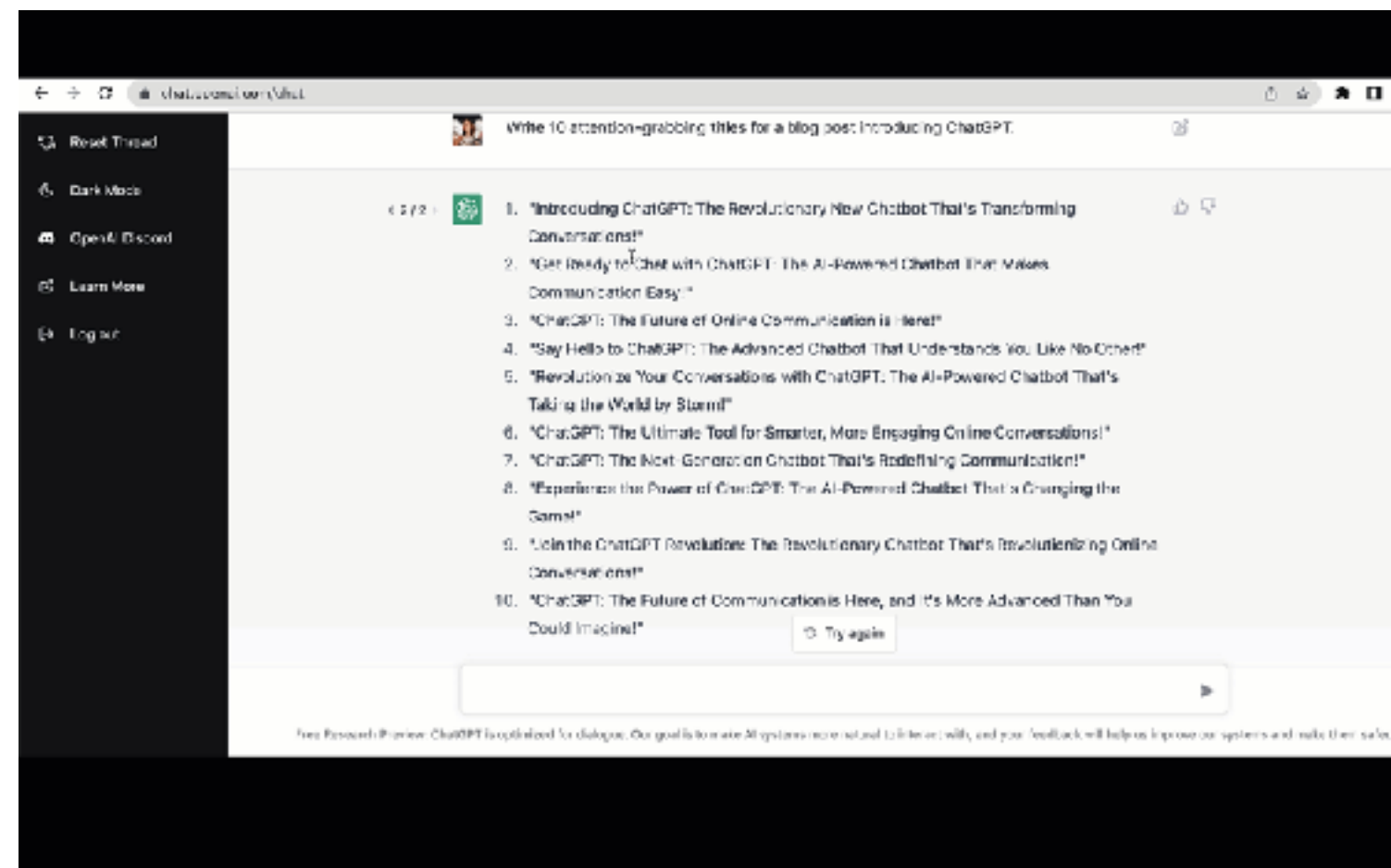
# LLMs are incredibly capable



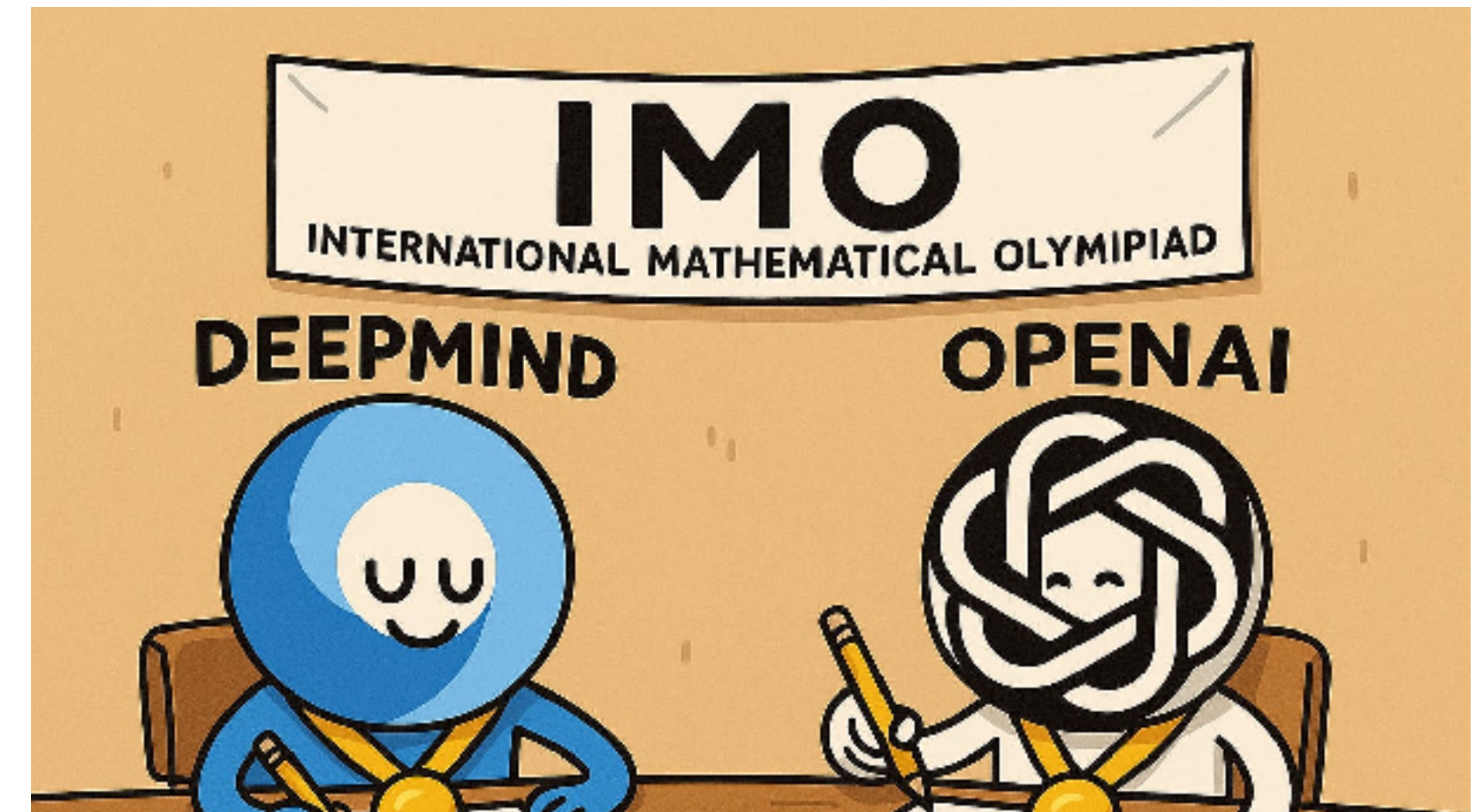


# LLMs are incredibly capable

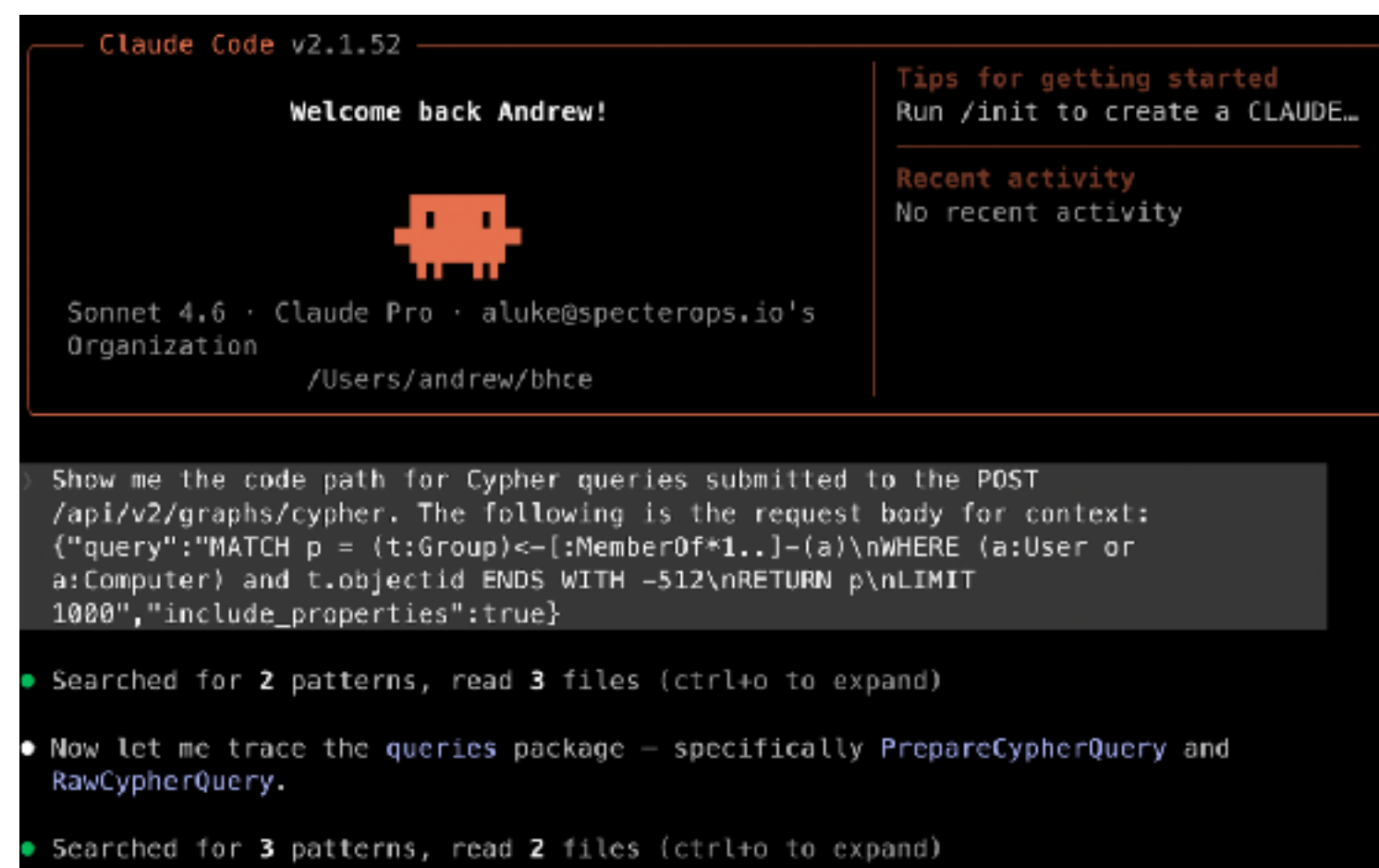
Chat



Reason



Code



Agentic



# Yet we don't understand LLMs well

**We can inspect every parameter, neuron**

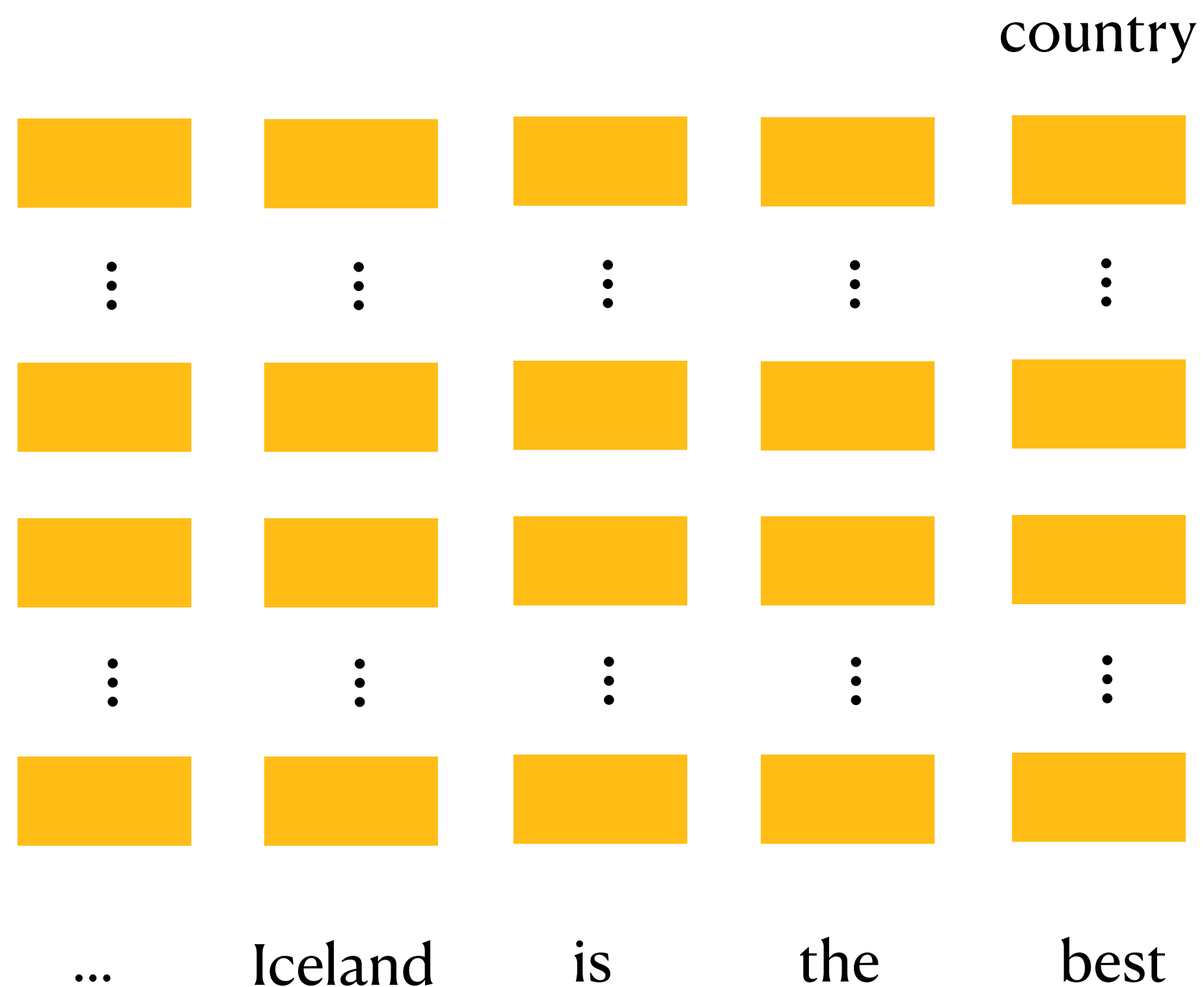
**But we struggle to understand its behavior..**

We don't know

- How LLMs process grammar
- How it stores and retrieves facts and knowledge
- How it chooses what to say

**Such lack of understanding underlies safety issues**

- Hallucinations
- Deceptive/unsafe behaviors
- Jail breaks





# Mechanistic Interpretability Literature

- **Transformer circuits:** reverse-engineering transformers (*Elhage et al., 2021*).
- **Probing:** train probes to test whether concepts are accessible in activations (*Belinkov, 2022*).
- **Causal tracing / activation patching** — intervene on activations to localize components responsible for factual recall or behavior (*Meng et al., 2022*).
- **Sparse autoencoders / feature discovery** — decompose activations into more interpretable latent features (*Bricken et al., 2023*).
- **Model editing** — modify localized internal representations or weights to change specific facts/behaviors (*Meng et al., 2022*)

Most approaches explain static mechanisms in a single forward pass

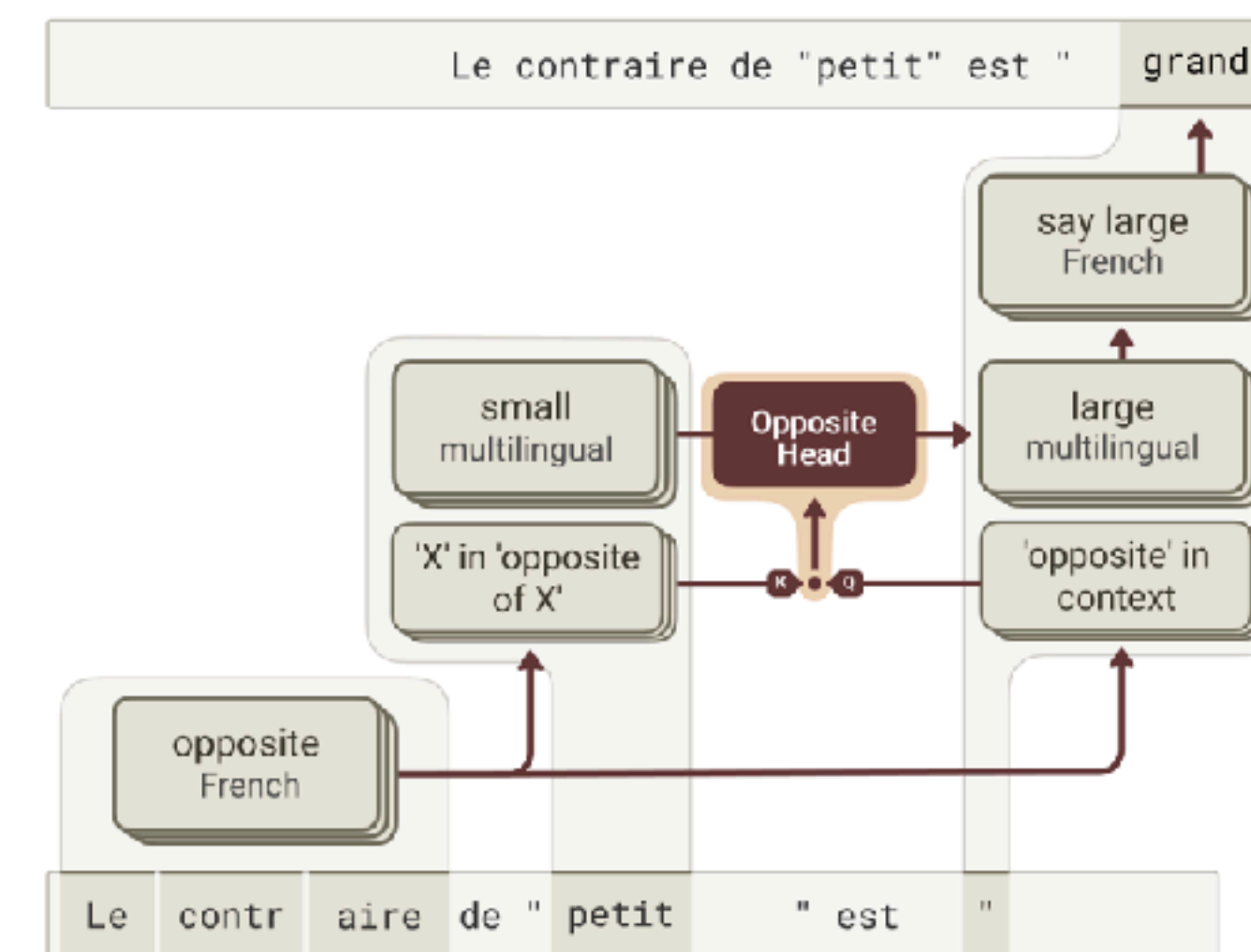
**This Talk: Dynamical System Perspective**

## Anthropic's Interpretability Research

A surprising fact about modern large language models is that nobody really knows how they work internally. The Interpretability team strives to change that — to understand these models to better plan for a future of safe AI.

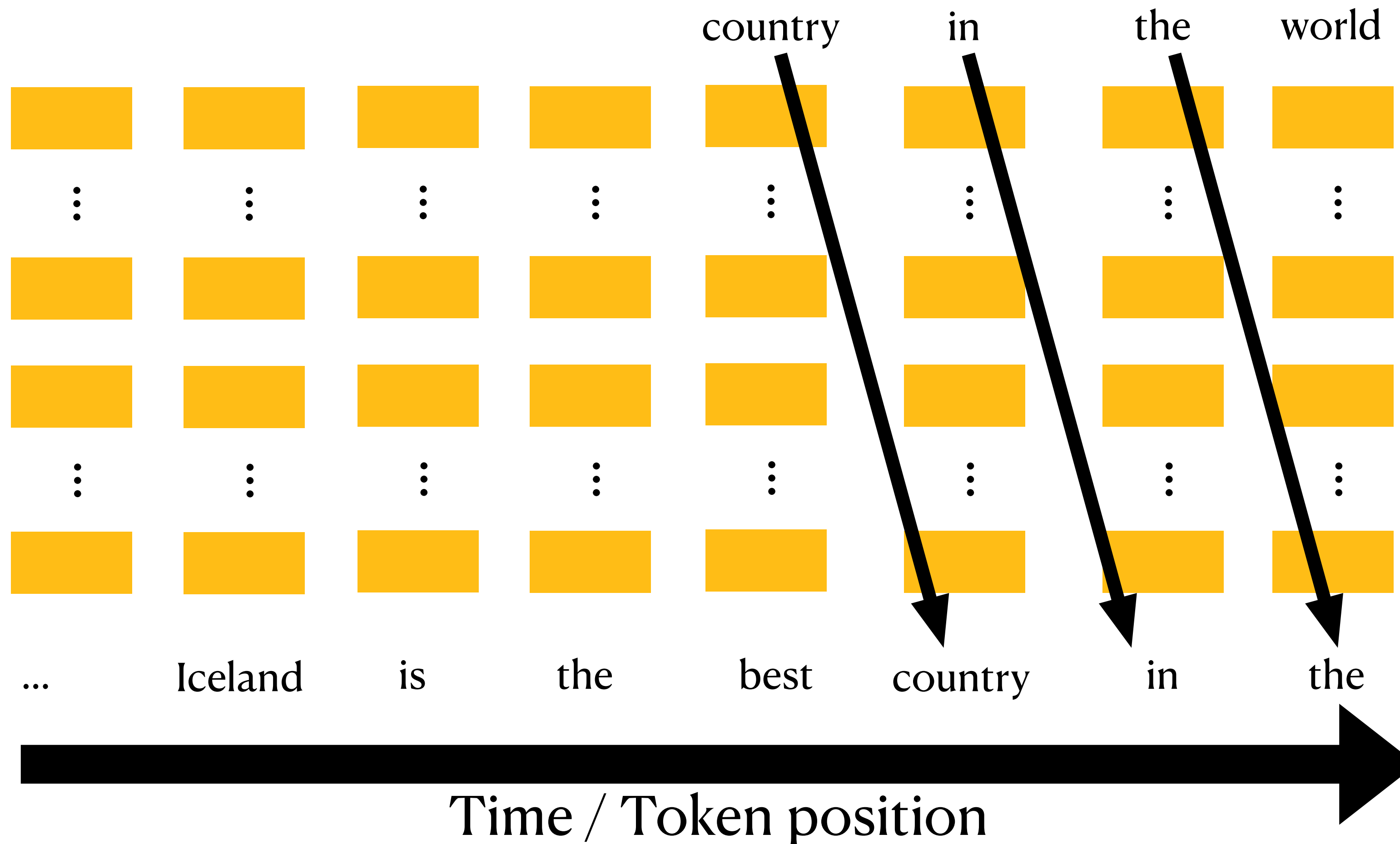
### Tracing Attention Computation Through Feature Interactions

We describe and apply a method to explain attention patterns in terms of feature interactions, and integrate this information into attribution graphs.



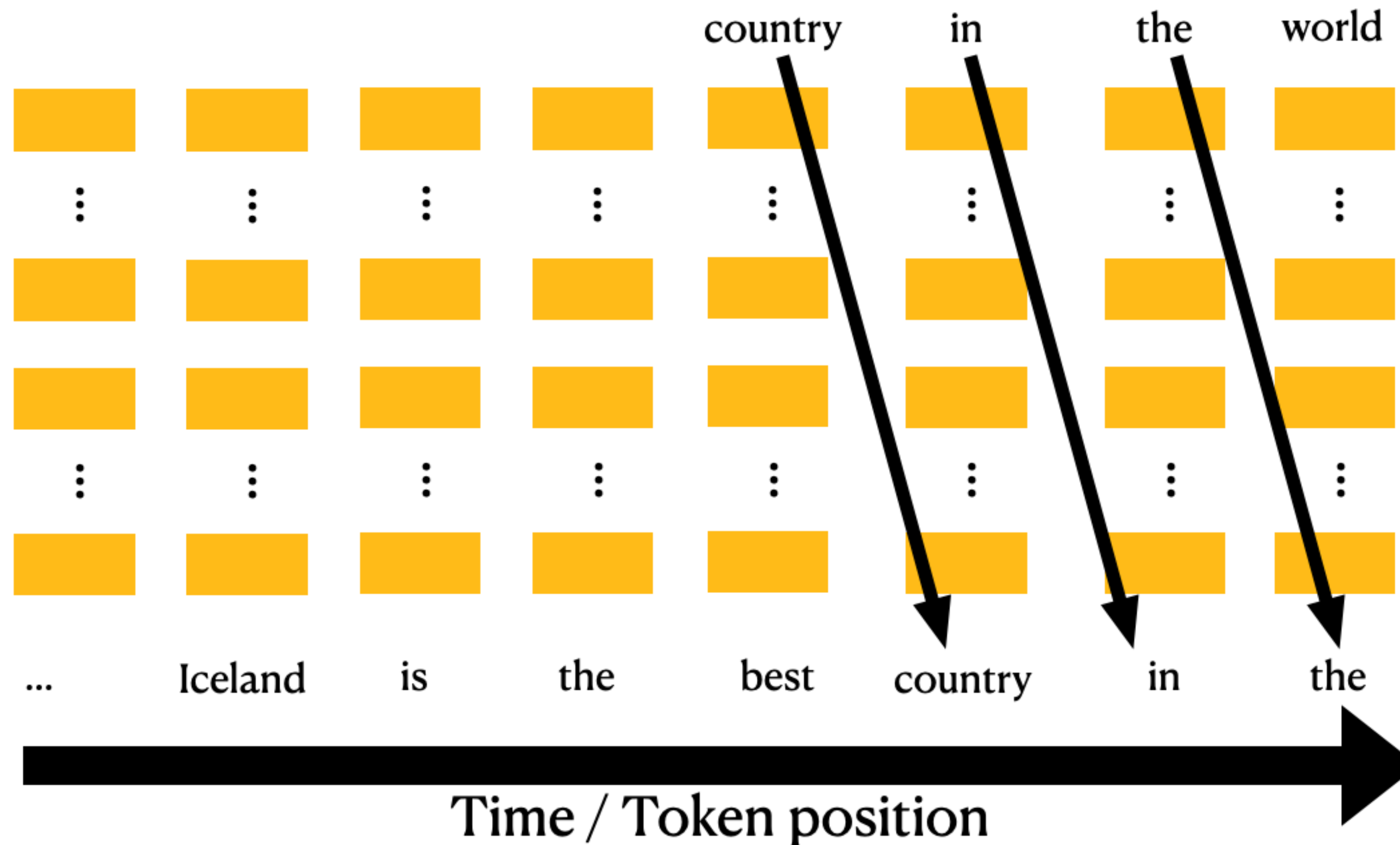
# This Talk: Dynamical System Perspective

**LLM generation is autoregressive**



# This Talk: Dynamical System Perspective

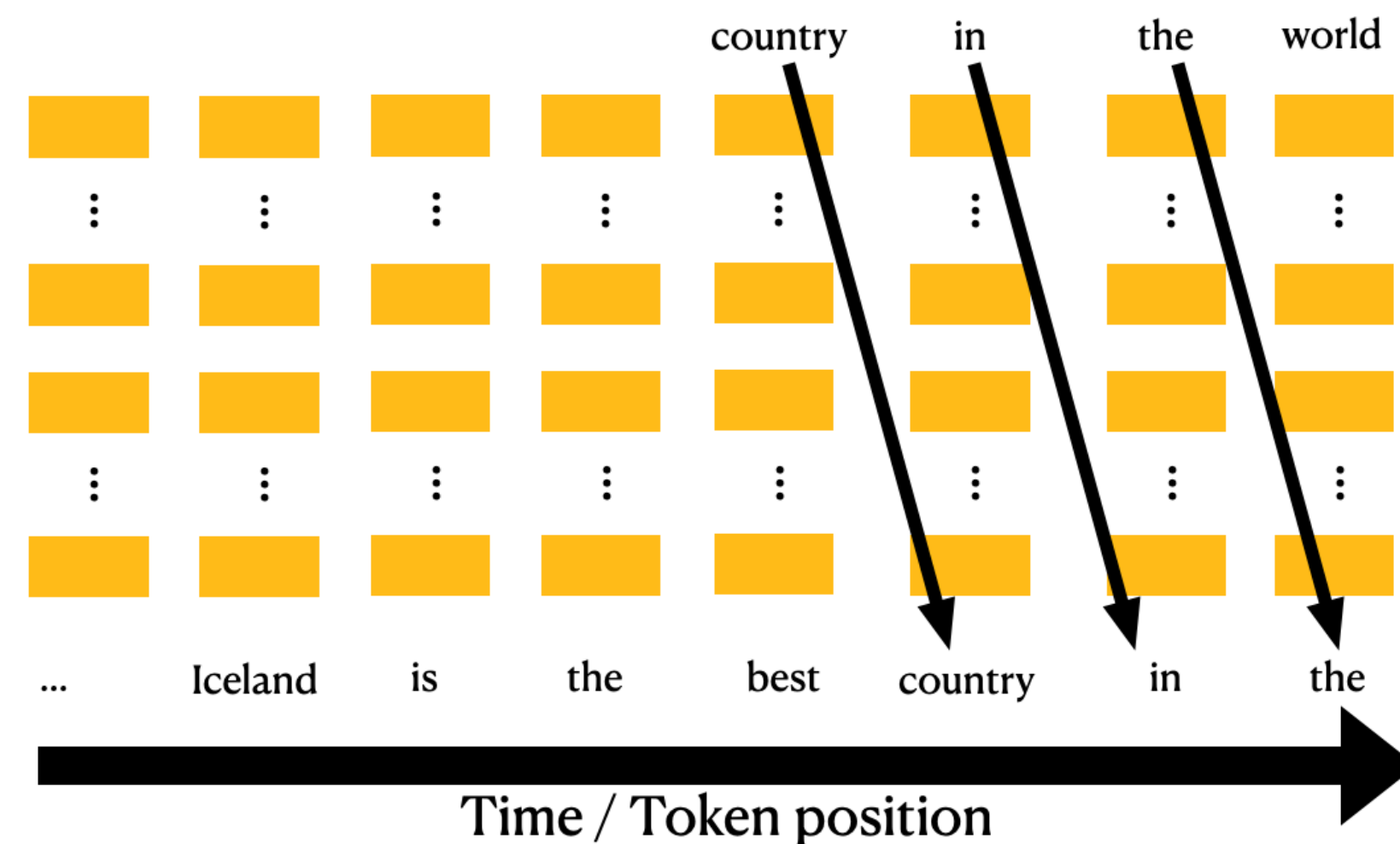
LLM generation is autoregressive



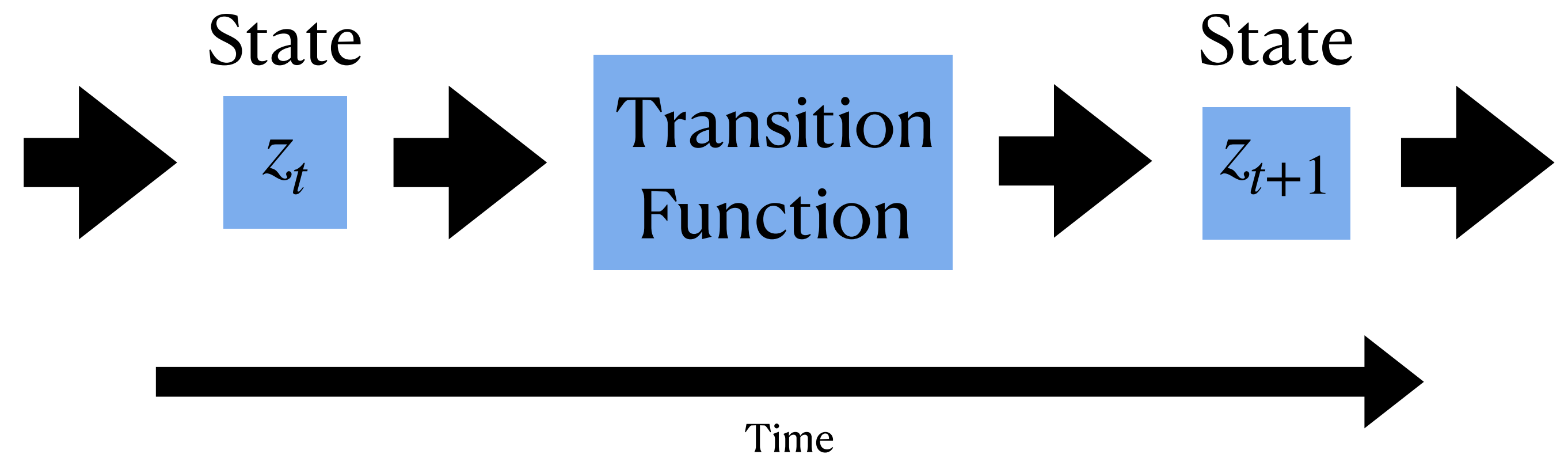


# This Talk: Dynamical System Perspective

LLM generation is autoregressive



This is reminiscent of dynamical systems



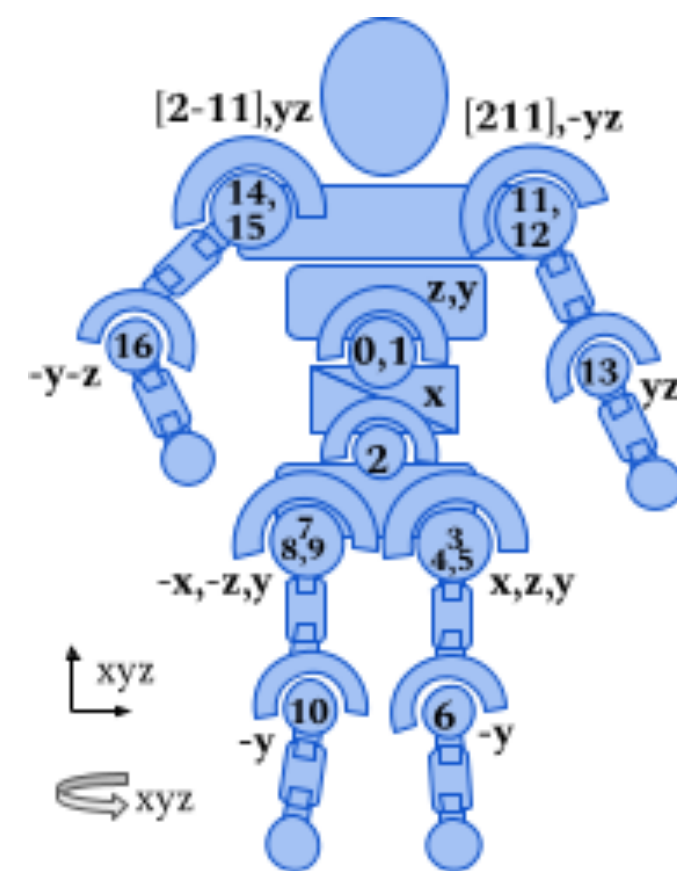
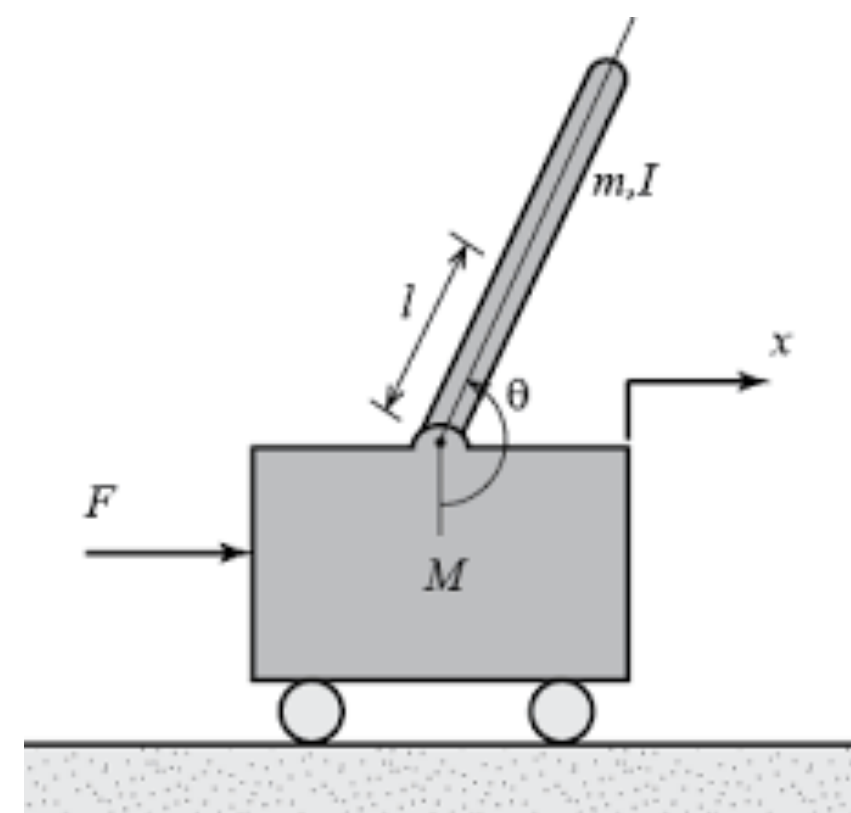


# This Talk: Dynamical System Perspective

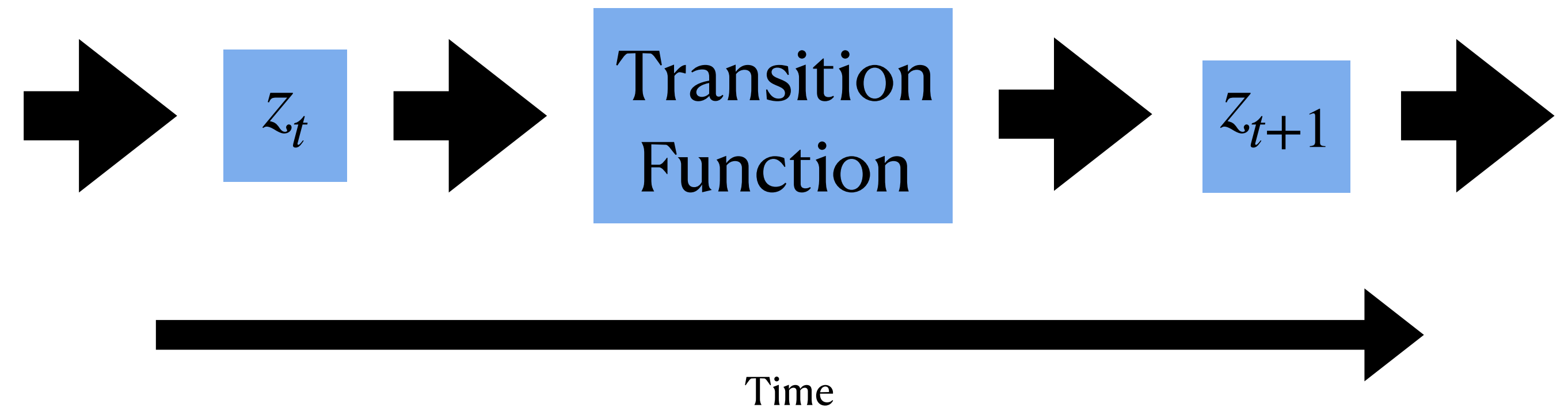
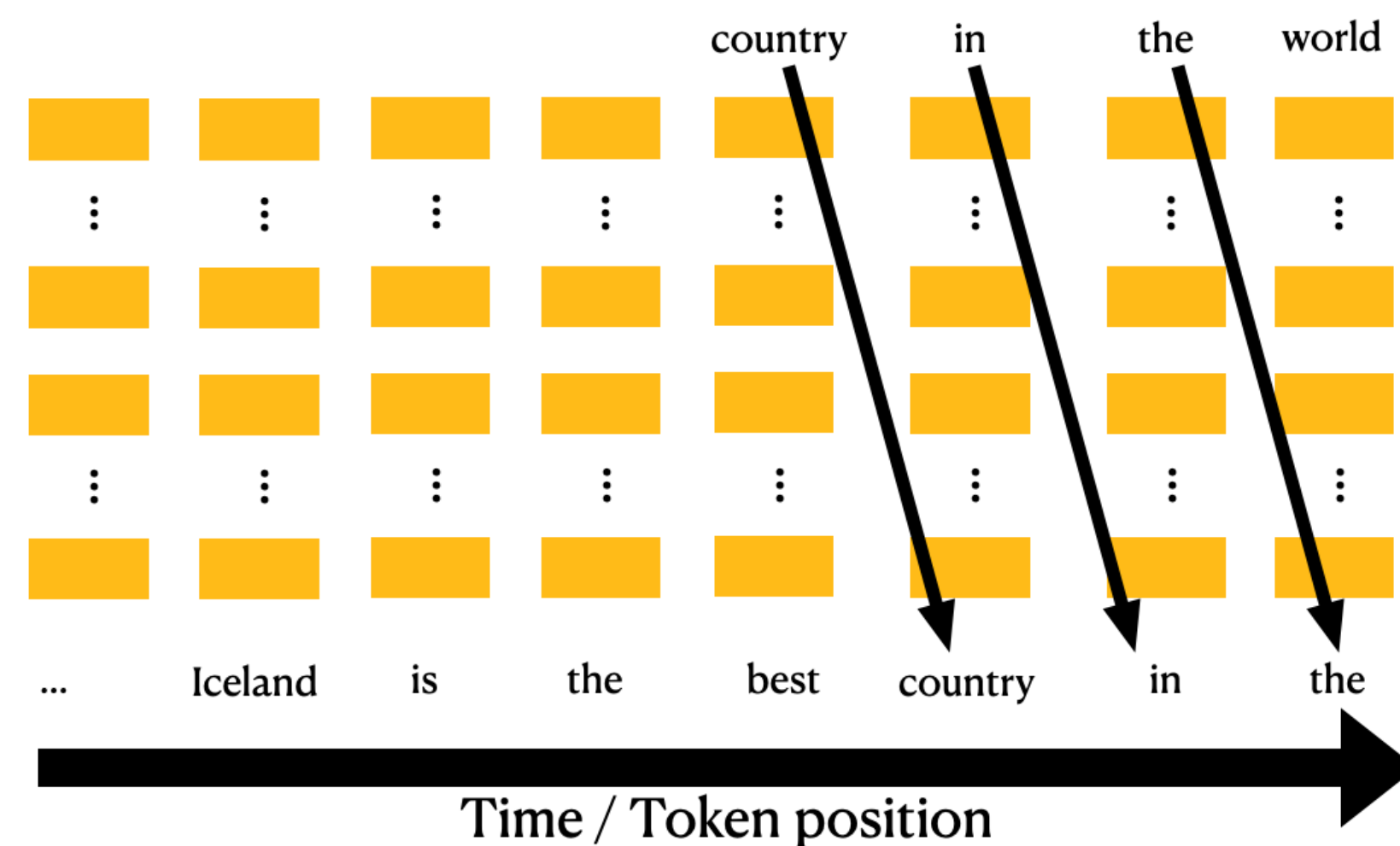


**State is a summary of the system that is**

(a) predictive of its future (b) can easily extract practically relevant information



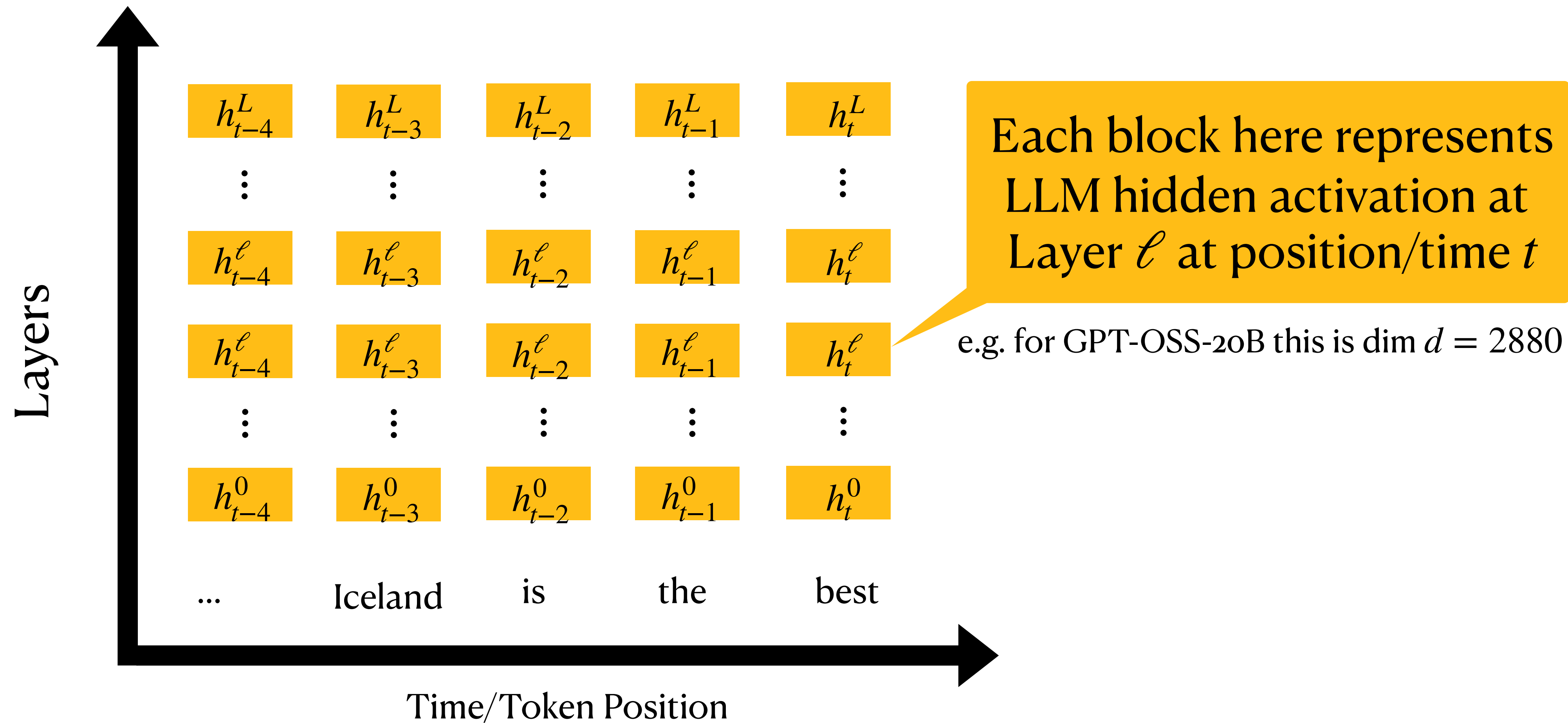
# This Talk: Dynamical System Perspective



**What is the state of LLMs in the autoregressive generation process?**

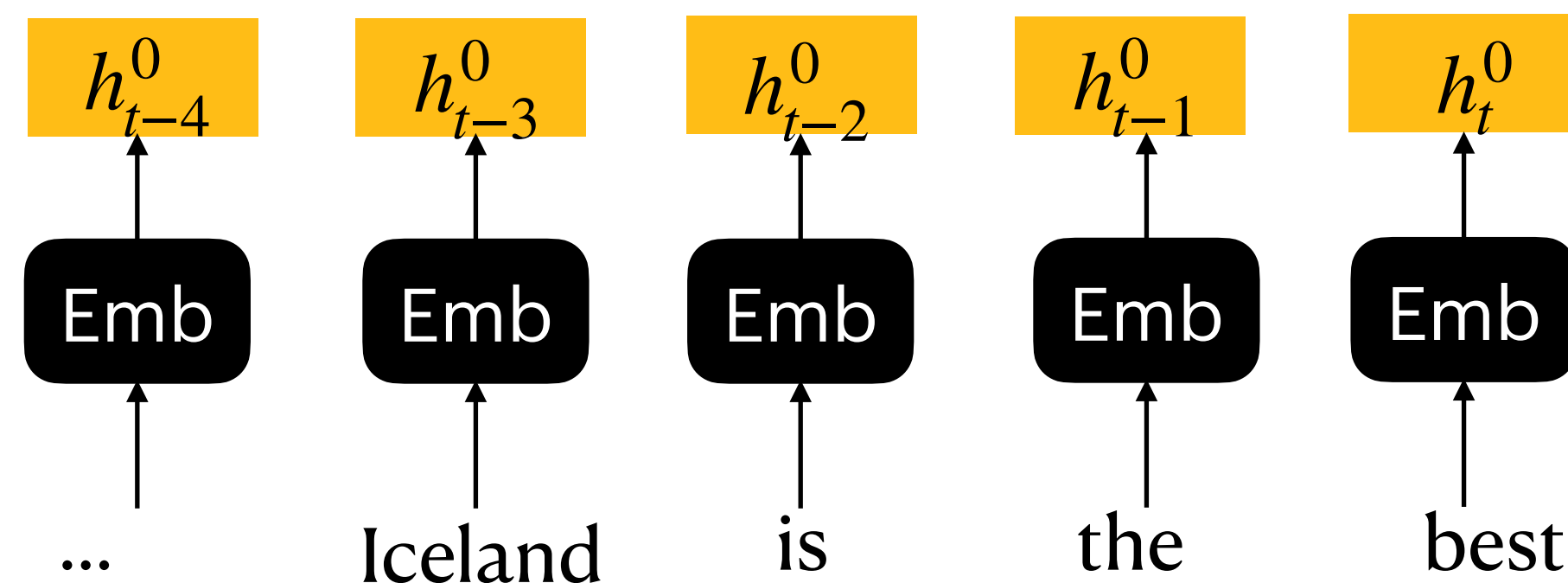
**How LLMs work?**

# How LLMs work?





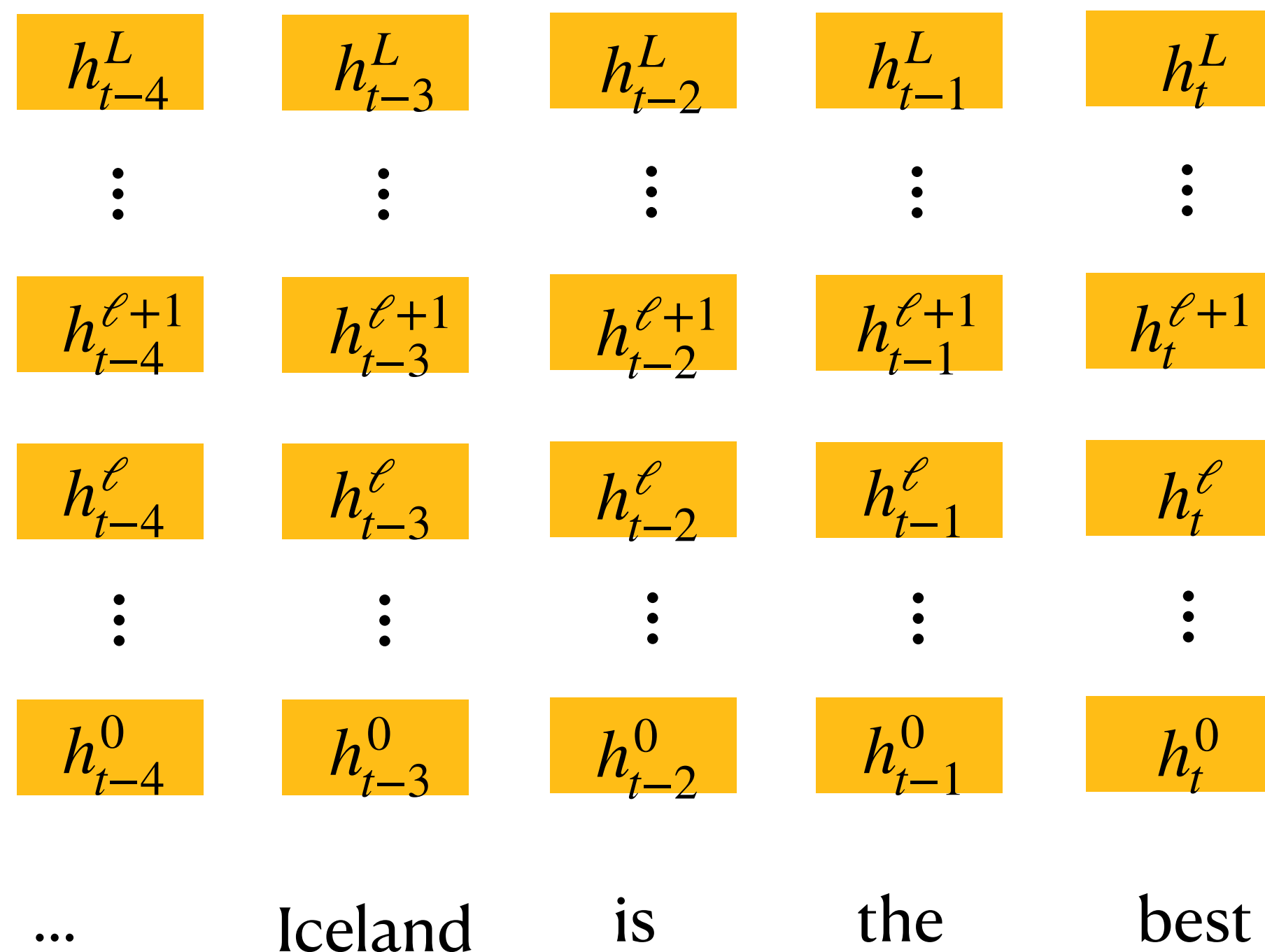
# How LLMs work?



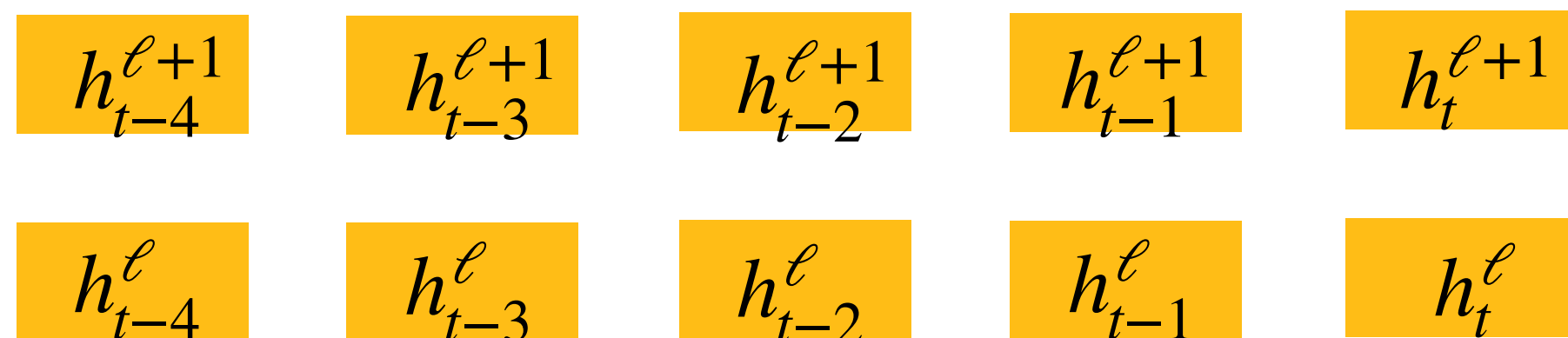
Embedding layer converts token to a d-dim vector



# How LLMs work?

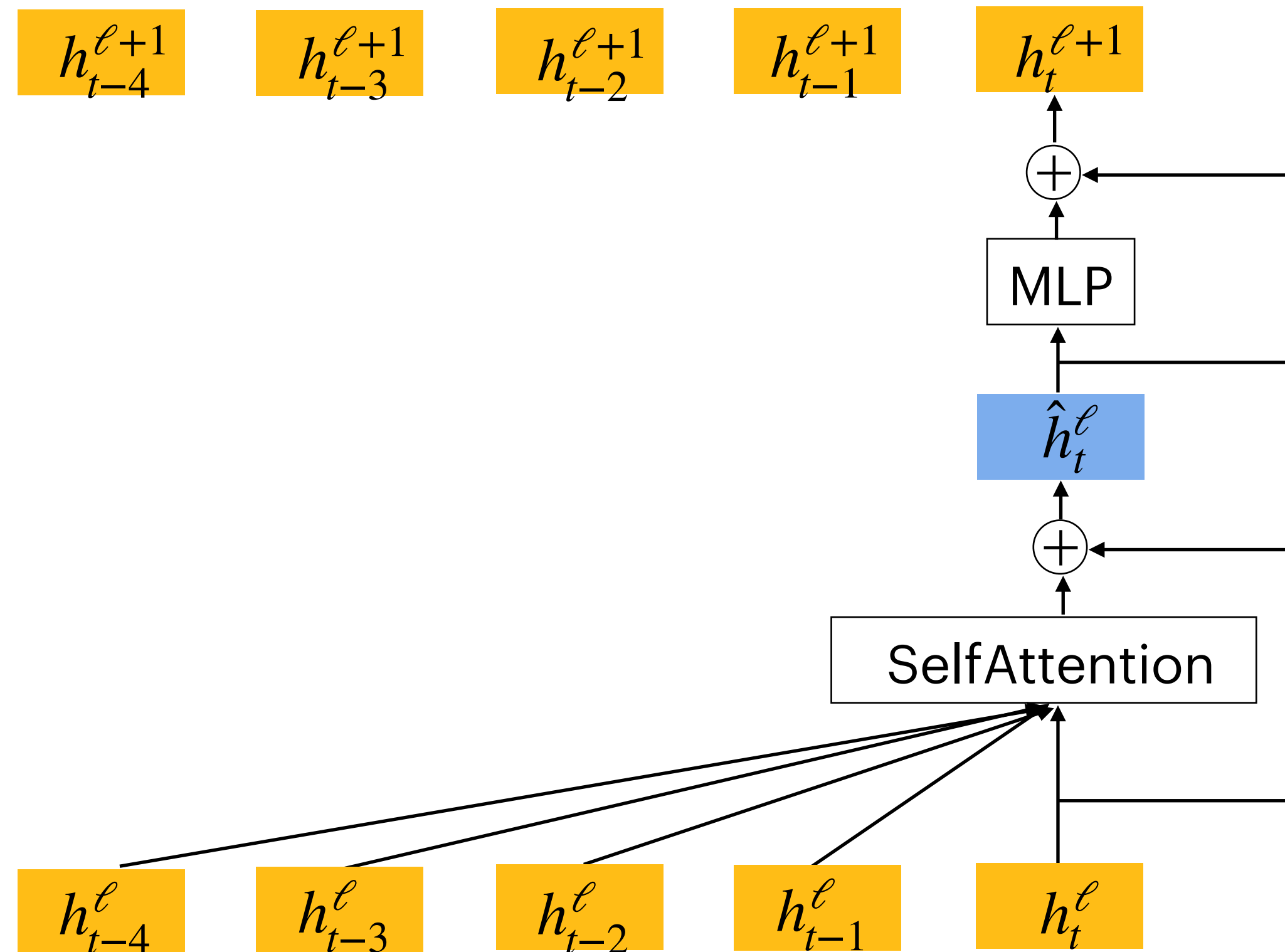


# How LLMs work?



... Iceland is the best

# How LLMs work?

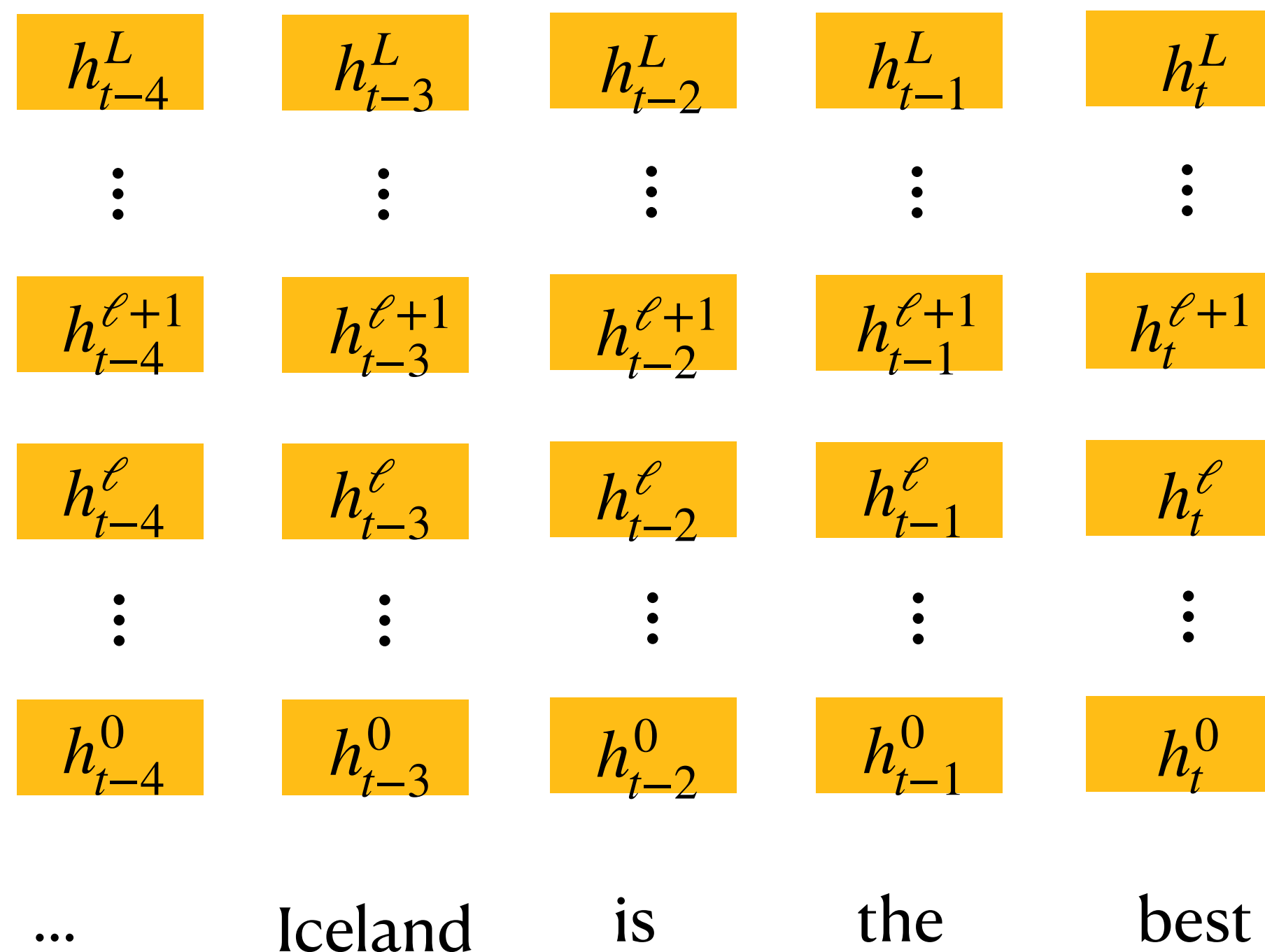


$$h_t^{\ell+1} = \hat{h}_t^\ell + \text{MLP}(\hat{h}_t^\ell)$$

$$\hat{h}_t^\ell = h_t^\ell + \text{SelfAttention}(h_{1:t}^\ell)$$



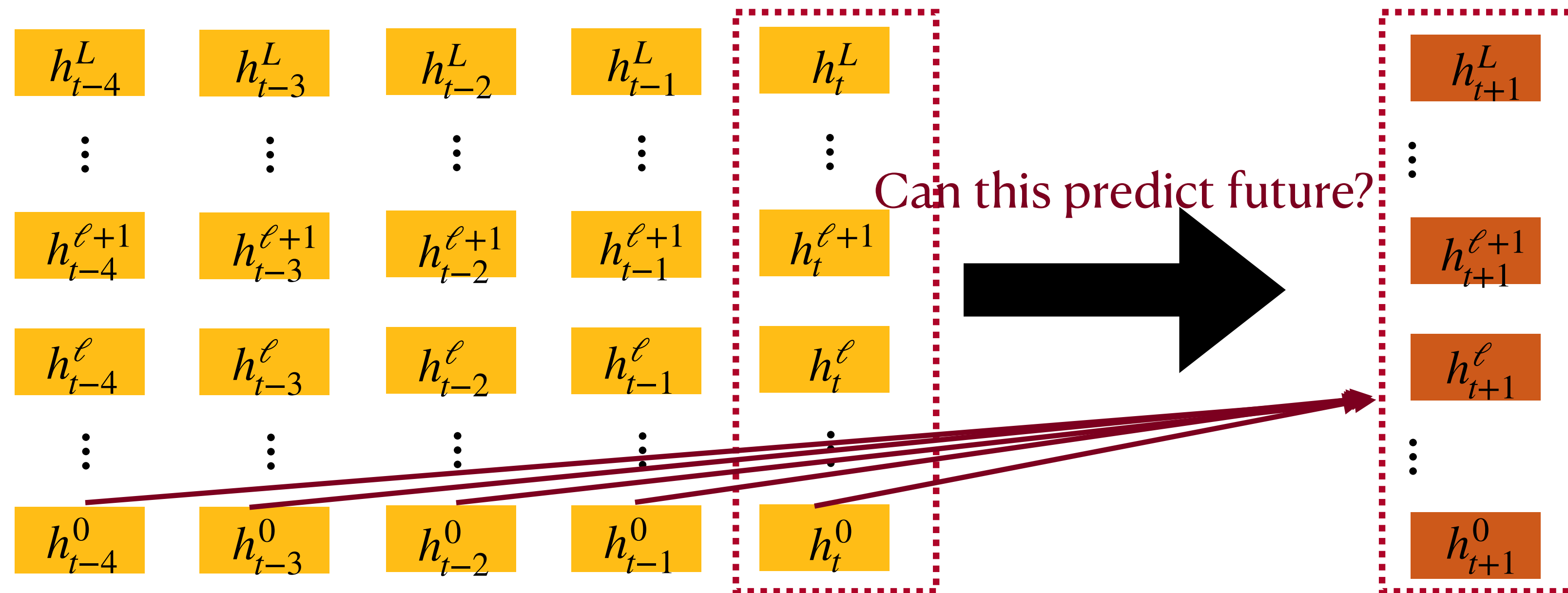
# How LLMs work?





# What should be the state of an LLM?

Can  $h_t^{0:L} = (h_t^0, \dots, h_t^L)$  be the state?



... Iceland is the best

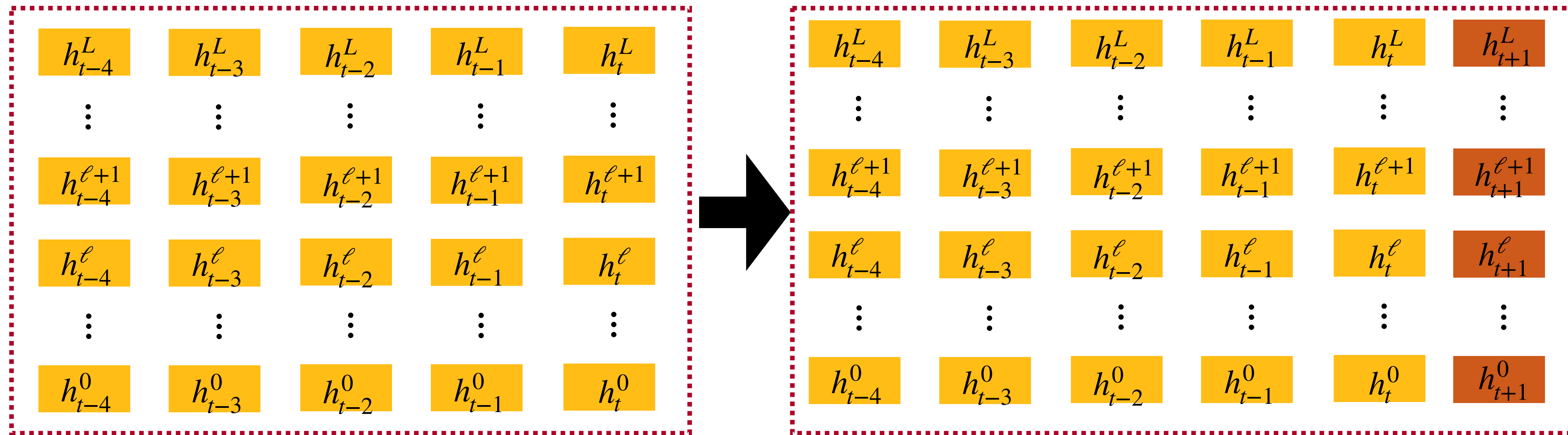
Not really because of dependence on past

# What should be the state of an LLM?

The tuple of all activations are self-predictive

$$(h_0^0, \dots, h_0^L, \dots, h_t^0, \dots, h_t^L)$$

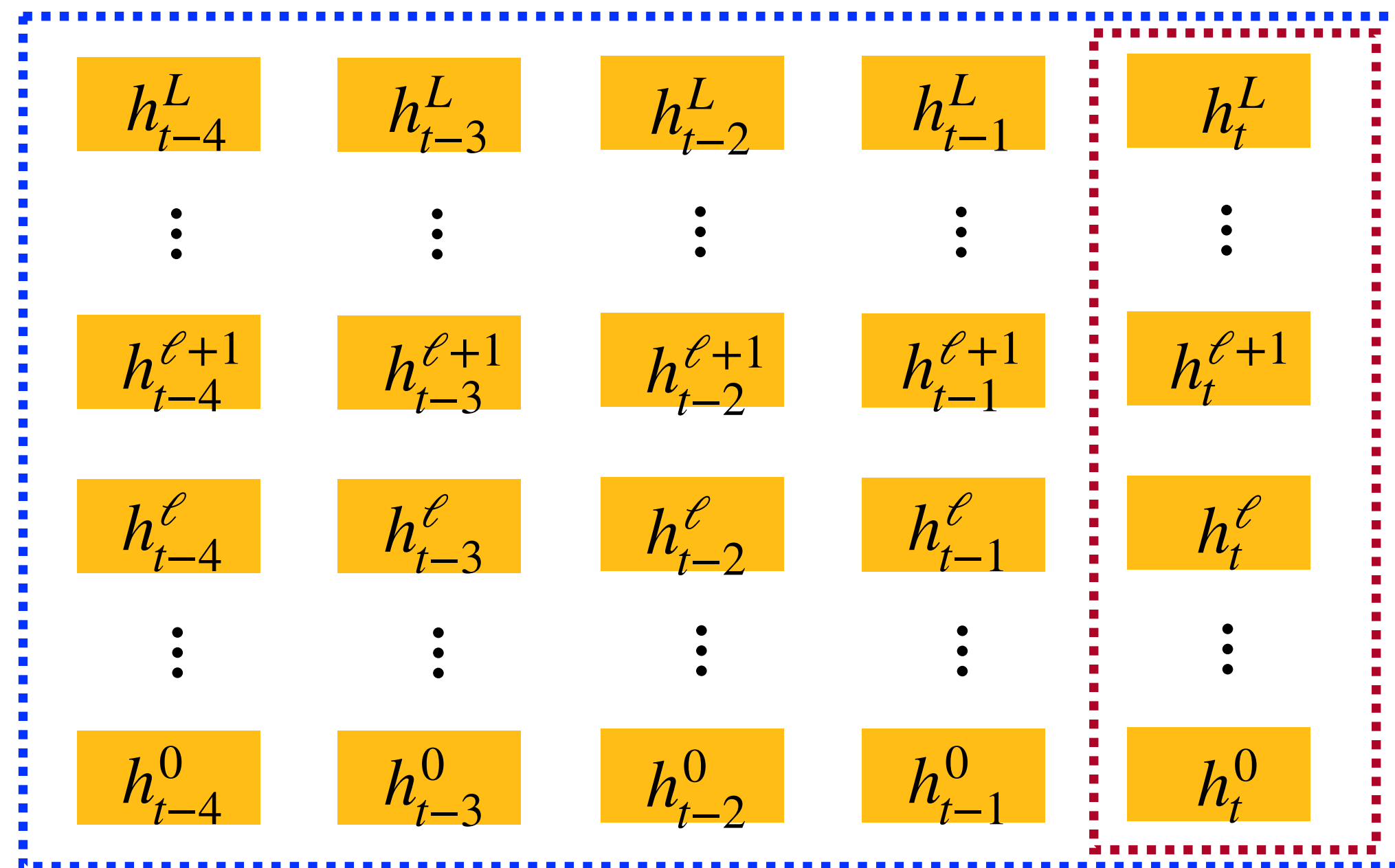
$$(h_0^0, \dots, h_0^L, \dots, h_t^0, \dots, h_t^L, h_{t+1}^0, \dots, h_{t+1}^L)$$



**While this can be the state, they are extremely high-dimensional (dim keeps growing)**

**More importantly, it is not useful**

# What should be the state of an LLM?



Activation at  $t$  is not self-predictive

Full history of activation is too high-dim

## What is the right state representation for an LLM?



# Inspiration from thermo-dynamics

The full state should be the **states of billions of molecules**  
Physicists have defined **macroscopic states**

**State = (Temperature, Pressure, Volume)**

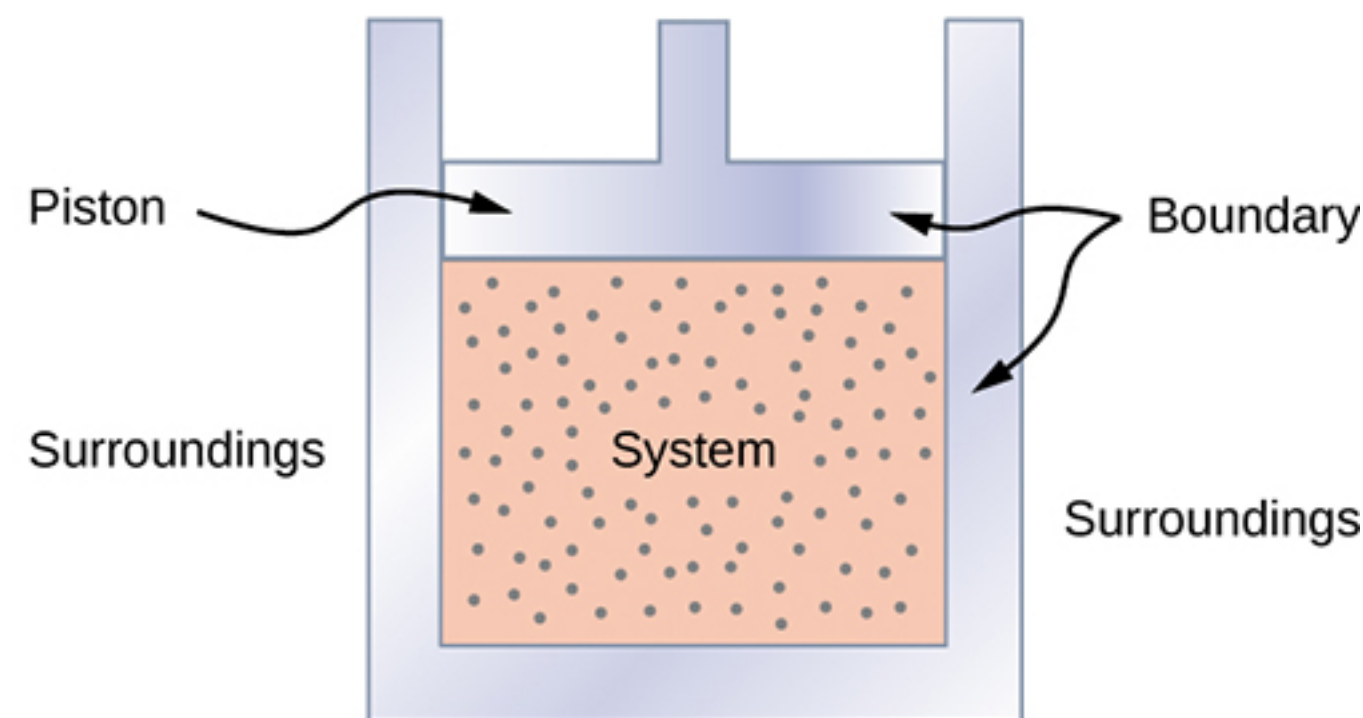
- Low dimensional and “high-level”
- Self predictive (dynamically closed)

**Only 3-dim**

**Thermodynamical systems**

- Practically useful

**We care about these variables**



**What is the right state representation for an LLM?**

# What is the right state representation for an LLM?

|                    |                    |                    |                    |                |
|--------------------|--------------------|--------------------|--------------------|----------------|
| $h_{t-4}^L$        | $h_{t-3}^L$        | $h_{t-2}^L$        | $h_{t-1}^L$        | $h_t^L$        |
| $\vdots$           | $\vdots$           | $\vdots$           | $\vdots$           | $\vdots$       |
| $h_{t-4}^{\ell+1}$ | $h_{t-3}^{\ell+1}$ | $h_{t-2}^{\ell+1}$ | $h_{t-1}^{\ell+1}$ | $h_t^{\ell+1}$ |
| $h_{t-4}^{\ell}$   | $h_{t-3}^{\ell}$   | $h_{t-2}^{\ell}$   | $h_{t-1}^{\ell}$   | $h_t^{\ell}$   |
| $\vdots$           | $\vdots$           | $\vdots$           | $\vdots$           | $\vdots$       |
| $h_{t-4}^0$        | $h_{t-3}^0$        | $h_{t-2}^0$        | $h_{t-1}^0$        | $h_t^0$        |
| ...                | Iceland            | is                 | the                | best           |

**The LLM state we want should be**

- Low dimensional and “high-level”
- Self predictive (dynamically closed)
- Practically useful



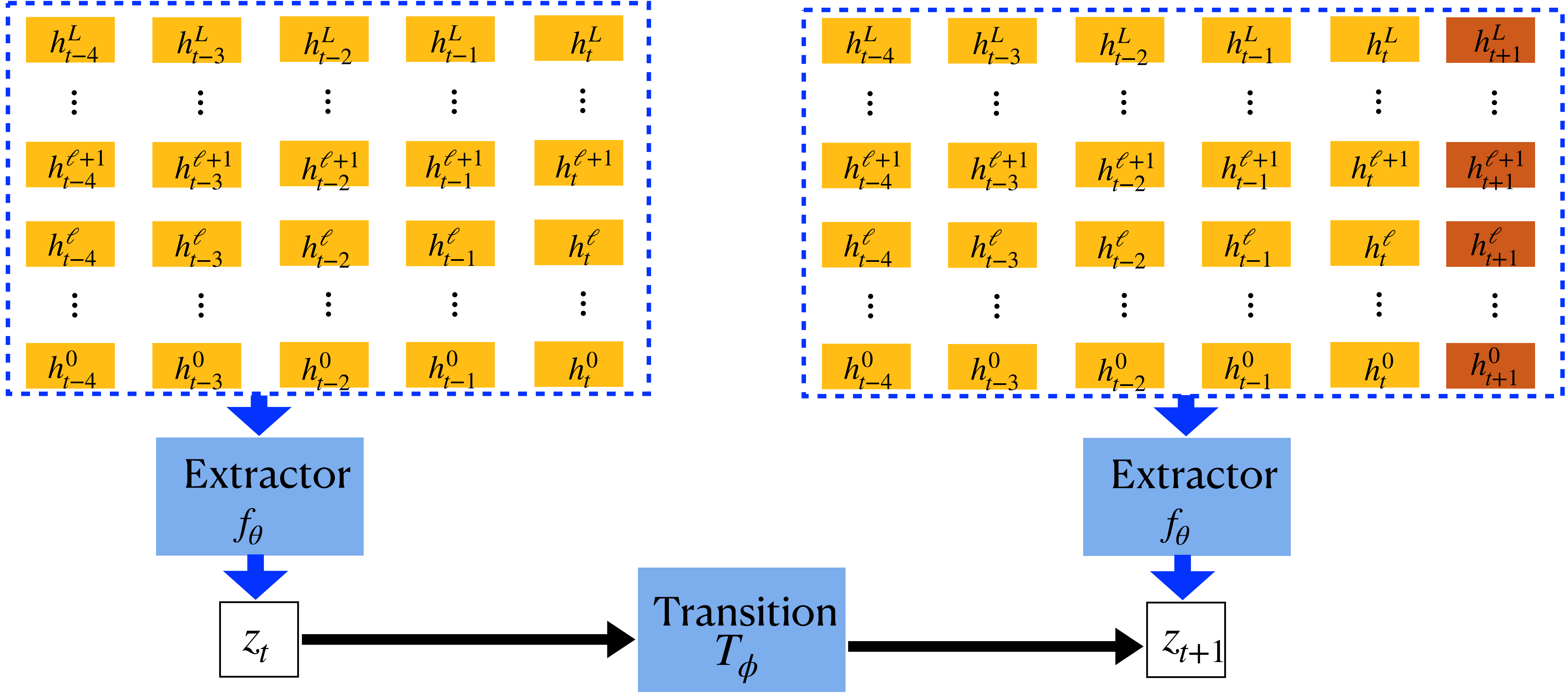
# Introducing RET: Representational Effective Theory

|                    |                    |                    |                    |                |
|--------------------|--------------------|--------------------|--------------------|----------------|
| $h_{t-4}^L$        | $h_{t-3}^L$        | $h_{t-2}^L$        | $h_{t-1}^L$        | $h_t^L$        |
| $\vdots$           | $\vdots$           | $\vdots$           | $\vdots$           | $\vdots$       |
| $h_{t-4}^{\ell+1}$ | $h_{t-3}^{\ell+1}$ | $h_{t-2}^{\ell+1}$ | $h_{t-1}^{\ell+1}$ | $h_t^{\ell+1}$ |
| $h_{t-4}^{\ell}$   | $h_{t-3}^{\ell}$   | $h_{t-2}^{\ell}$   | $h_{t-1}^{\ell}$   | $h_t^{\ell}$   |
| $\vdots$           | $\vdots$           | $\vdots$           | $\vdots$           | $\vdots$       |
| $h_{t-4}^0$        | $h_{t-3}^0$        | $h_{t-2}^0$        | $h_{t-1}^0$        | $h_t^0$        |
| ...                | Iceland            | is                 | the                | best           |

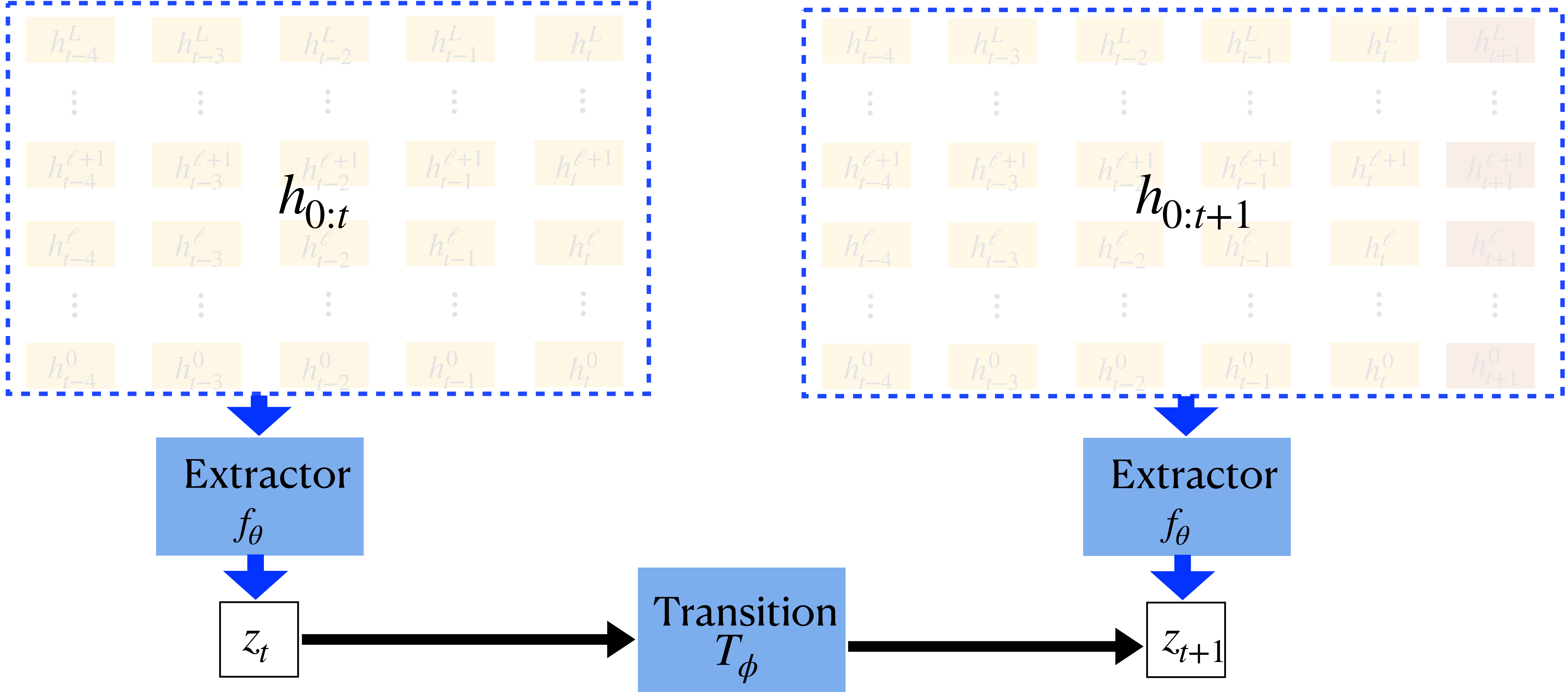
## Main Idea

Learn a state extractor that is encouraged to be **low-dim, high-level, self predictive**

# Introducing RET: Representational Effective Theory

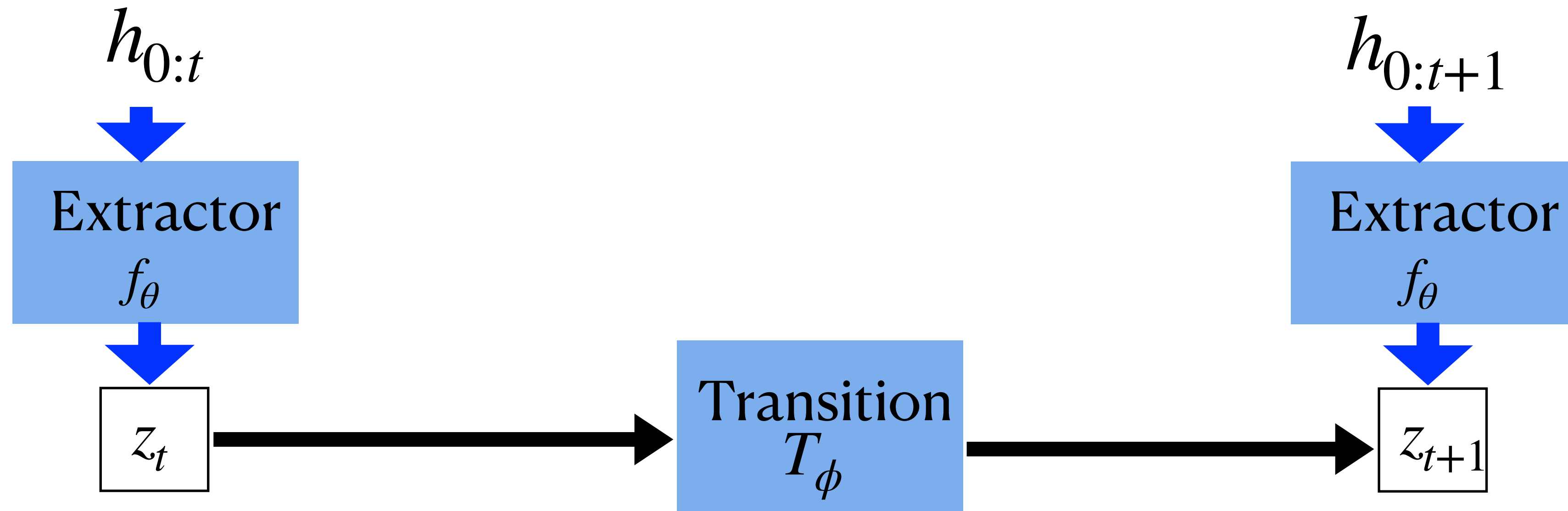


# Introducing RET: Representational Effective Theory

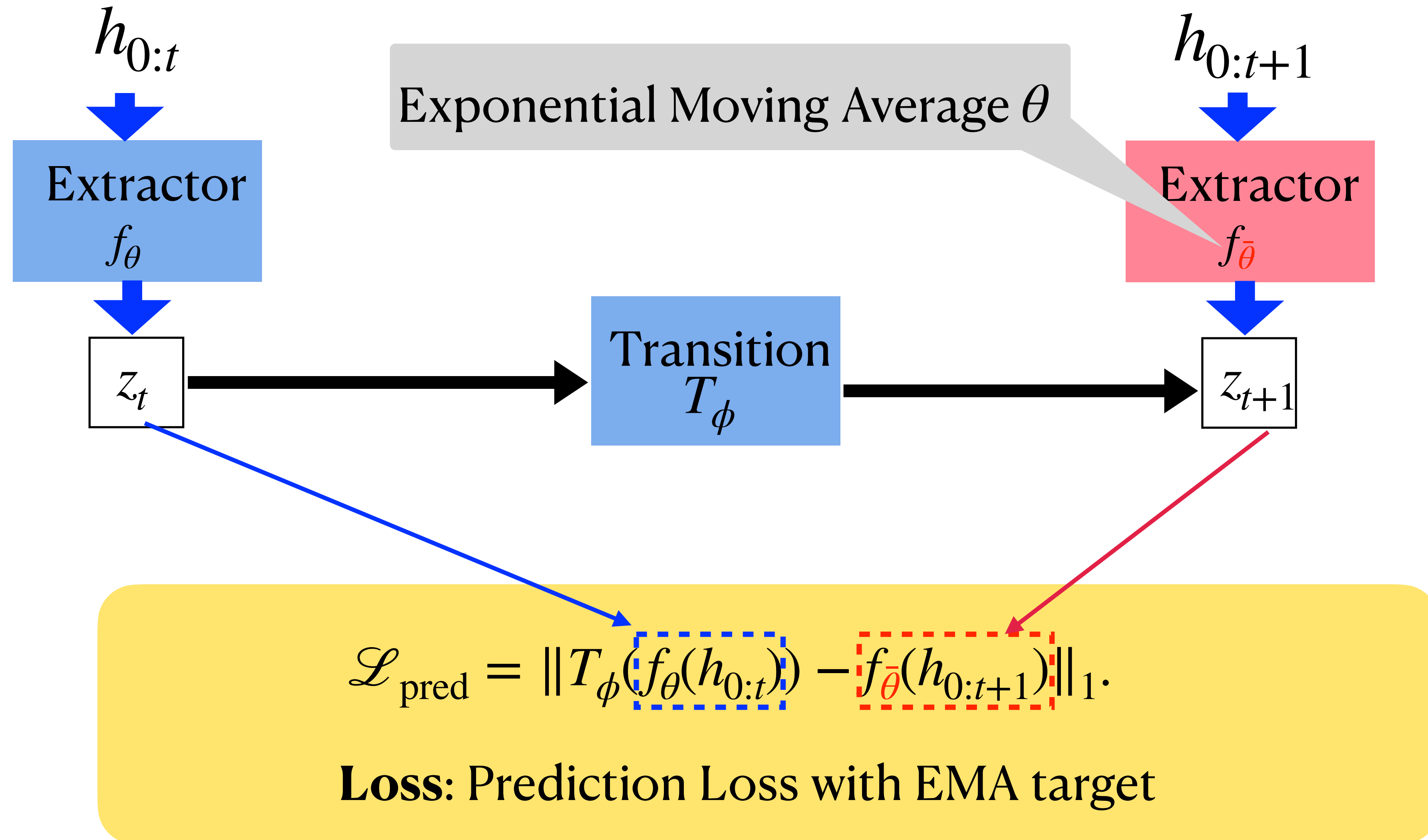




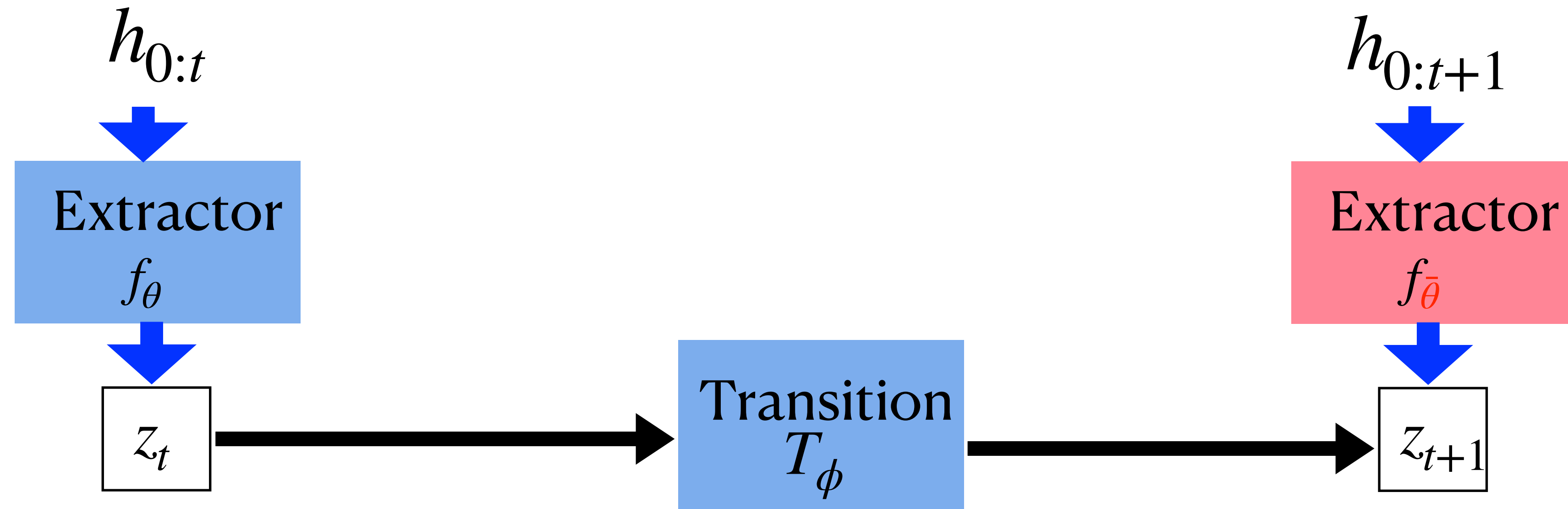
# Introducing RET: Representational Effective Theory



# Introducing RET: Representational Effective Theory



# Introducing RET: Representational Effective Theory

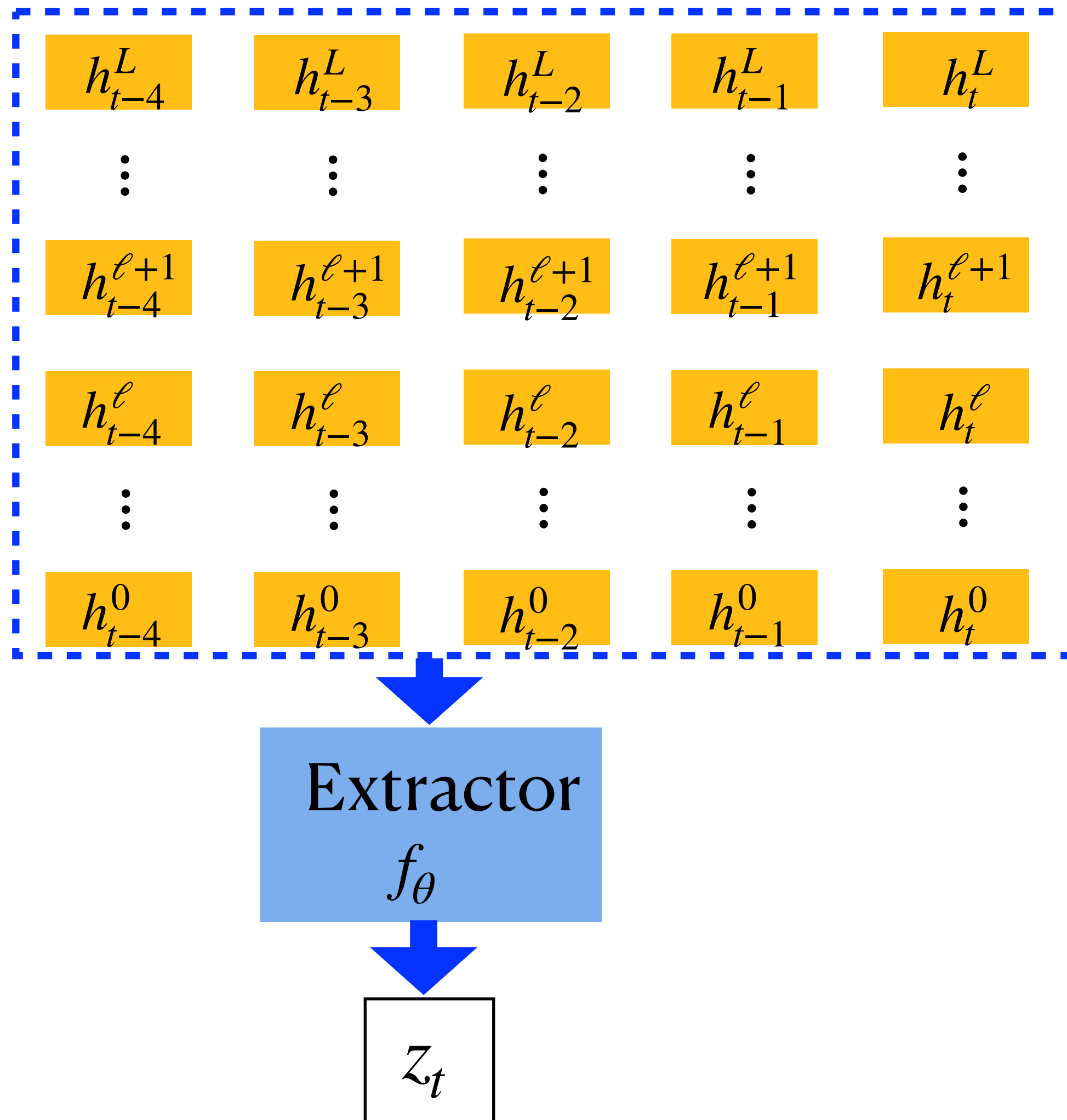


## Rationale

- Low dim  $z$ : encourages low-dim summary
- Self-predictive: only keep what is important for the future behavior, ignoring microscopic details (like grammar, specific word use)
- EMA target: prevent representation collapse (similar to JEPA)



# Evaluating RET



**Are the learned states...**

- Low dimensional and “high-level”?  
**Up next!**
- Self predictive (dynamically closed)?
- Practically useful?

# RET Captures High-level Plot

We train RET on **TinyStories** dataset for **Pythia-160M**

Then use RET to extract  $z_t$  for the following:

The young boys scored their third goal of the afternoon match and celebrated wildly by running in big circles around the grassy soccer field.

A torrential downpour drenched the field within seconds, forcing everyone to abandon the game and quickly run toward the nearest open building for shelter.

They stepped into the quiet public library and spent the rest of the rainy afternoon reading dusty old books about forgotten sailors and pirate legends

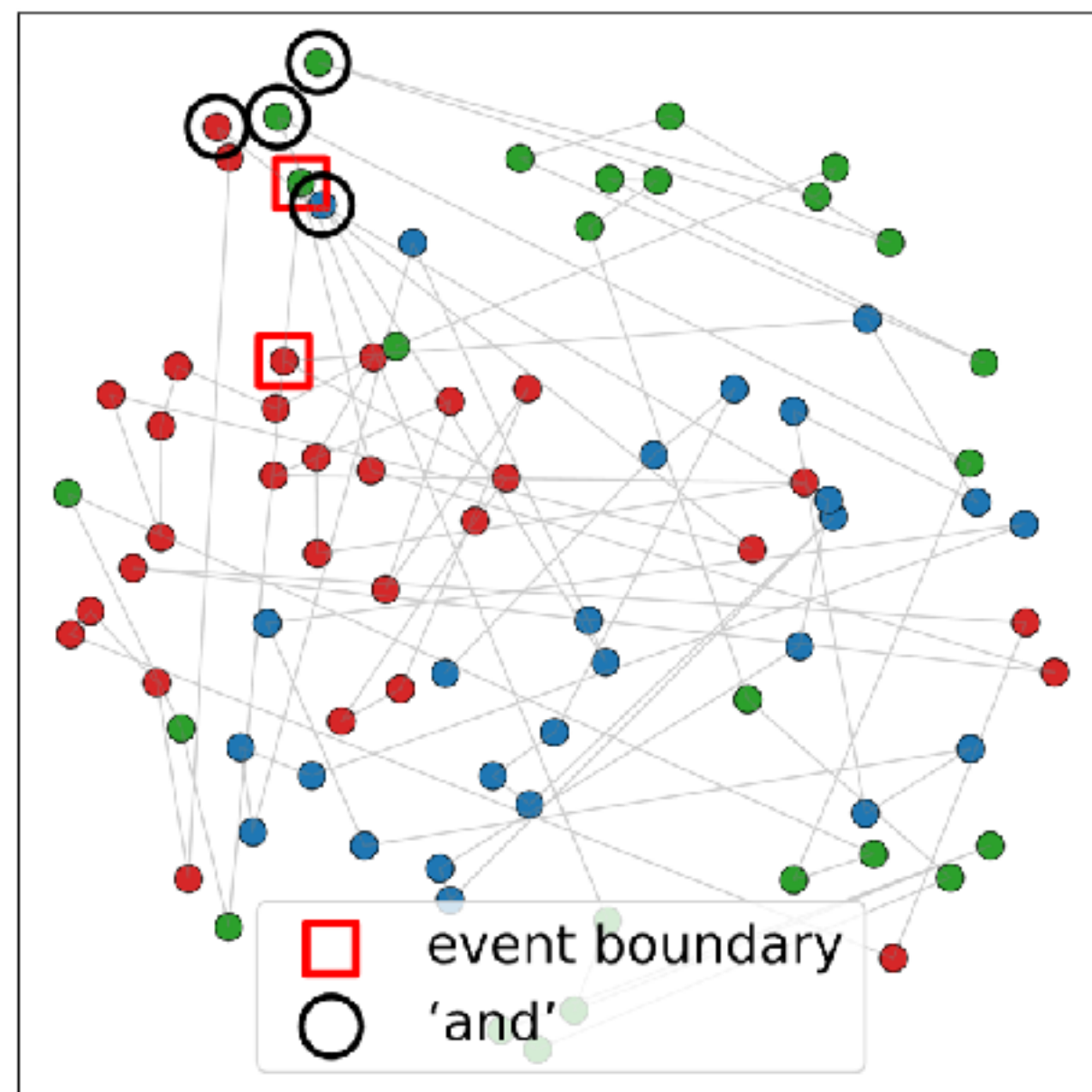
# RET Captures High-level Plot

The young boys scored their third goal of the afternoon match and celebrated wildly by running in big circles around the grassy soccer field.

A torrential downpour drenched the field within seconds, forcing everyone to abandon the game and quickly run toward the nearest open building for shelter.

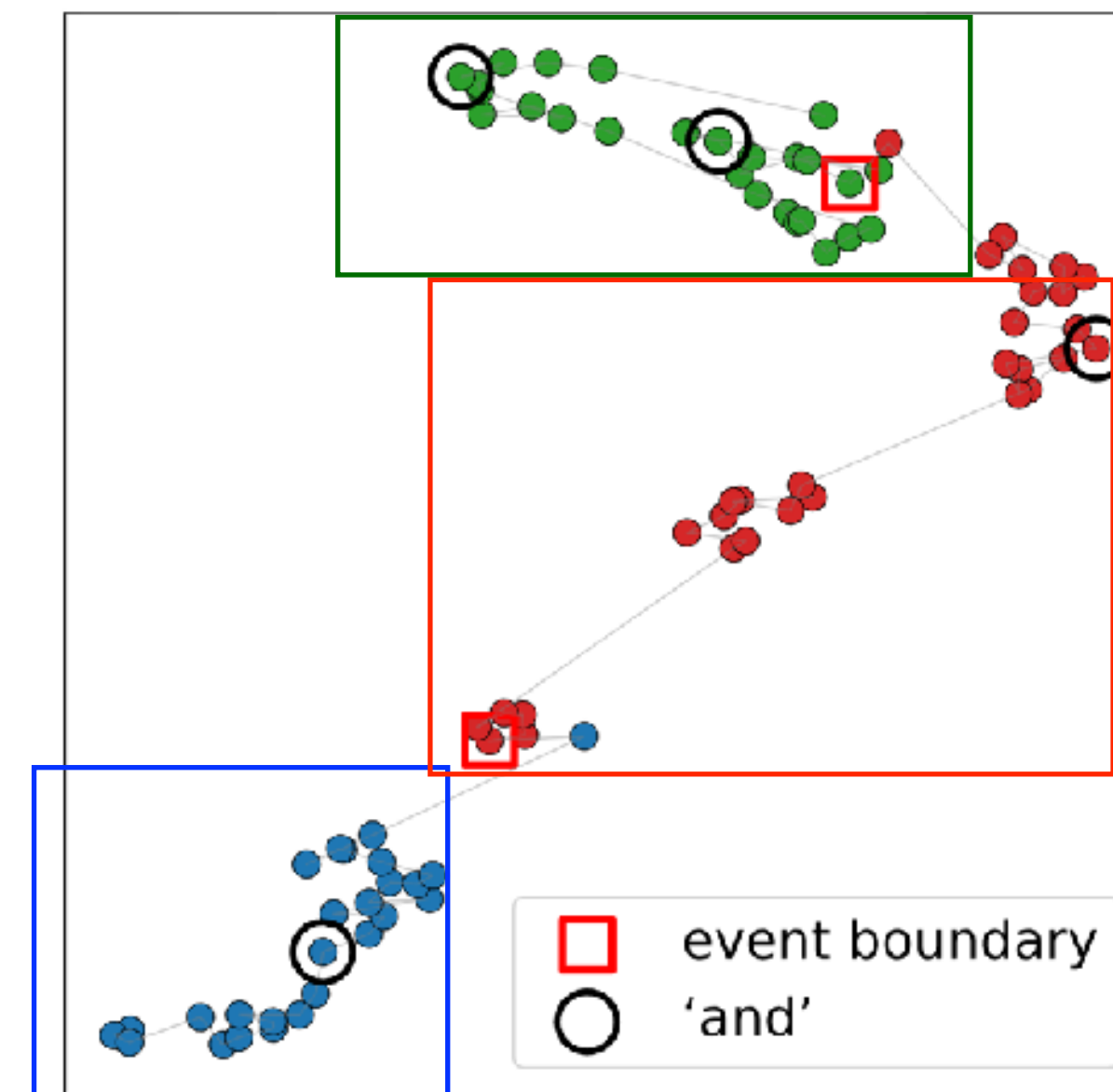
They stepped into the quiet public library and spent the rest of the rainy afternoon reading dusty old books about forgotten sailors and pirate legends

## States mixed



State trajectory of  $h_t$   
(UMAP plot)

## States cluster according to story plot

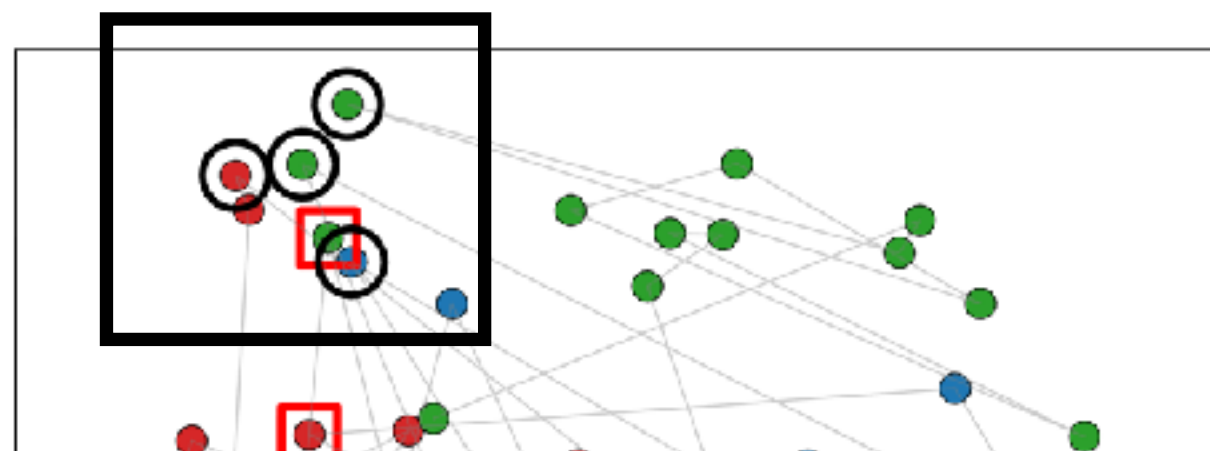


State trajectory of  $z_t$  (RET)  
(UMAP plot)



# RET Captures **High-level Plot**

In contrast, the  $h_t$  of “and” positions are put together



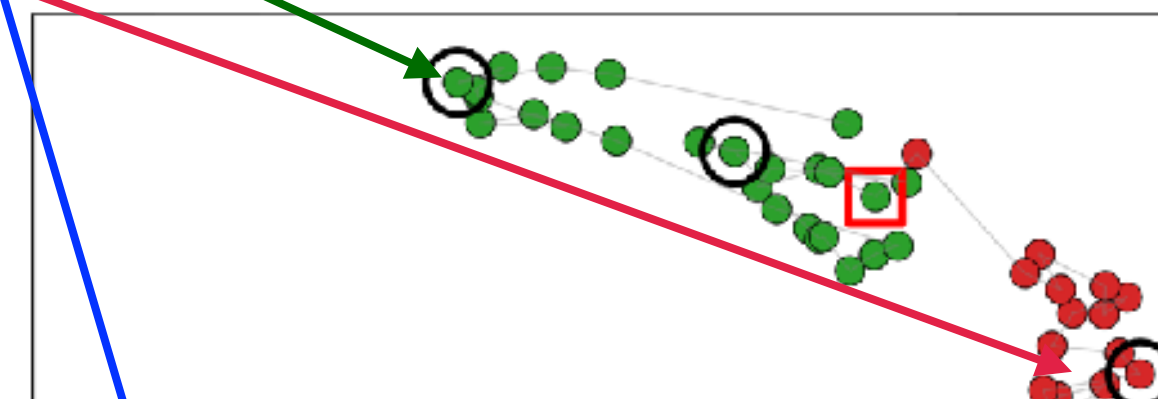
The raw activation  $h_t$  reflects low level word choices (“and”)

State trajectory of  $h_t$

The young boys scored their third goal of the afternoon match and celebrated wildly by running in big circles around the grassy soccer field.

A torrential downpour drenched the field within seconds, forcing everyone to abandon the game and quickly run toward the nearest open building for shelter.

They stepped into the quiet public library and spent the rest of the rainy afternoon reading dusty old books about forgotten sailors and pirate legends

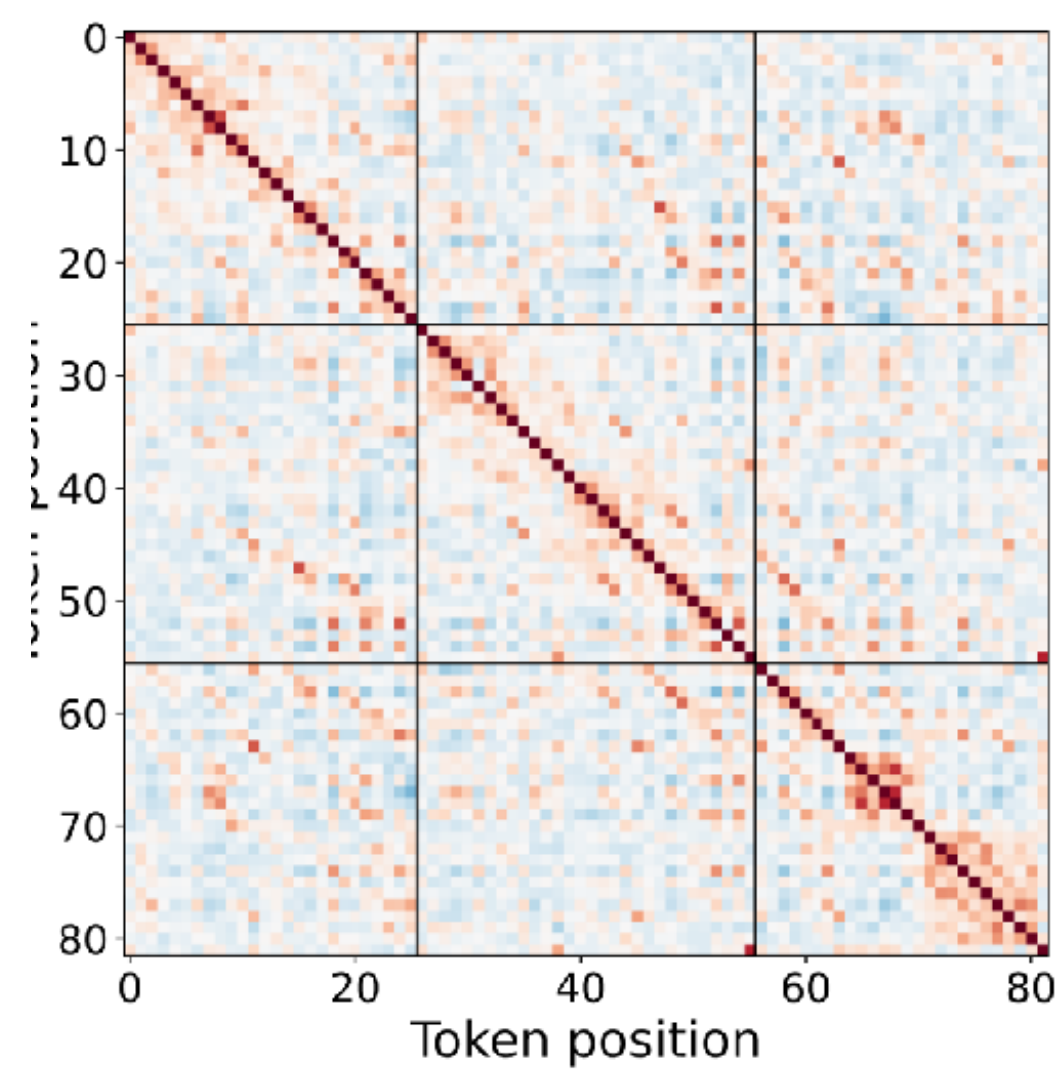
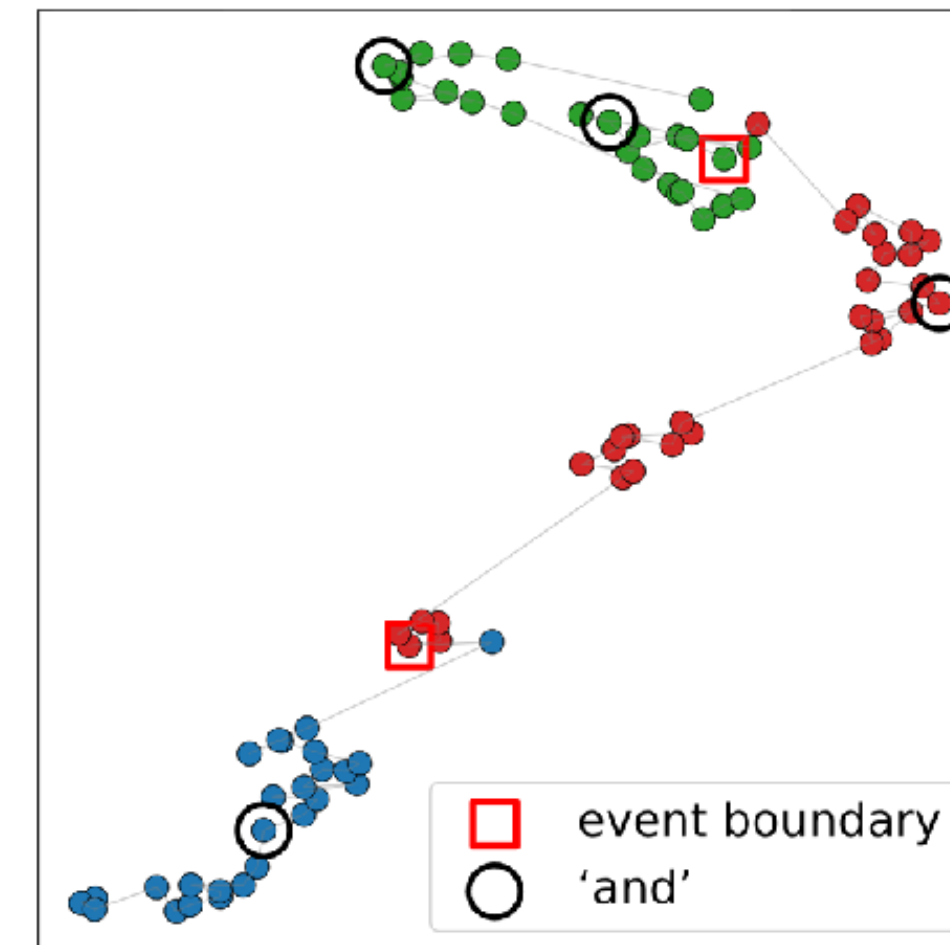
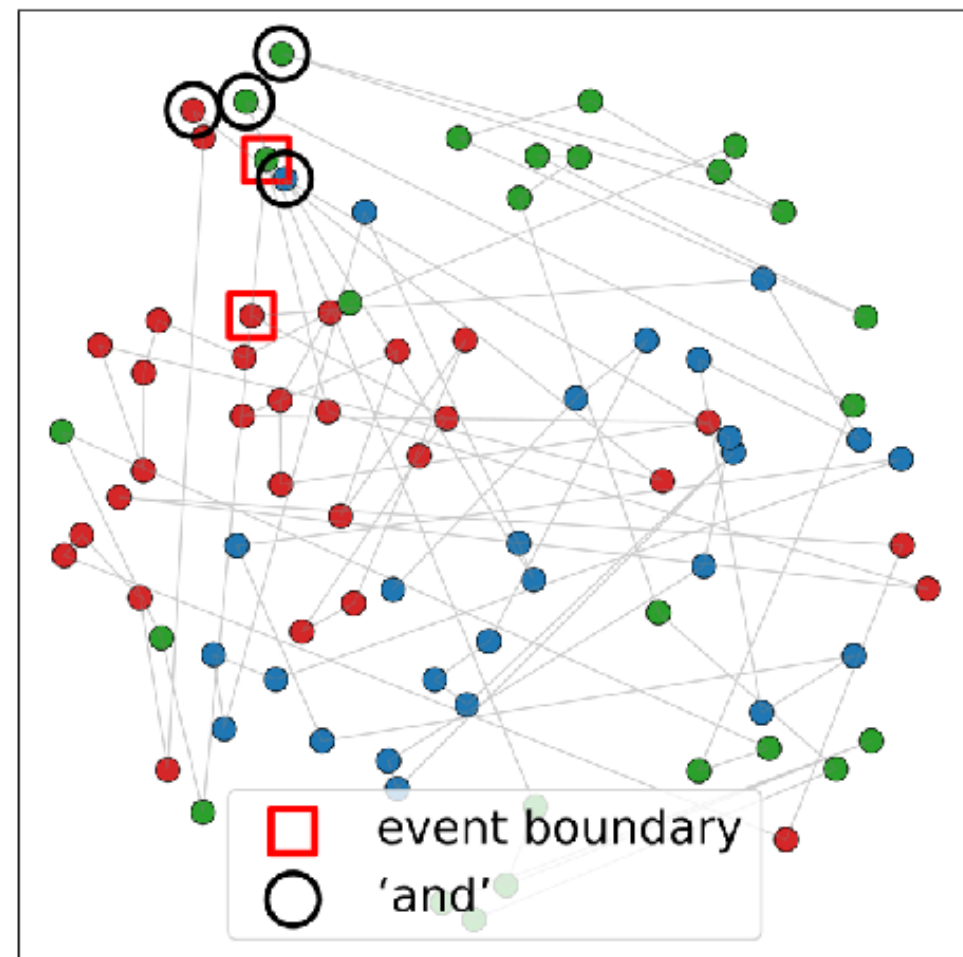


The learned high-level state  $z_t$  represents **high-level story plots**

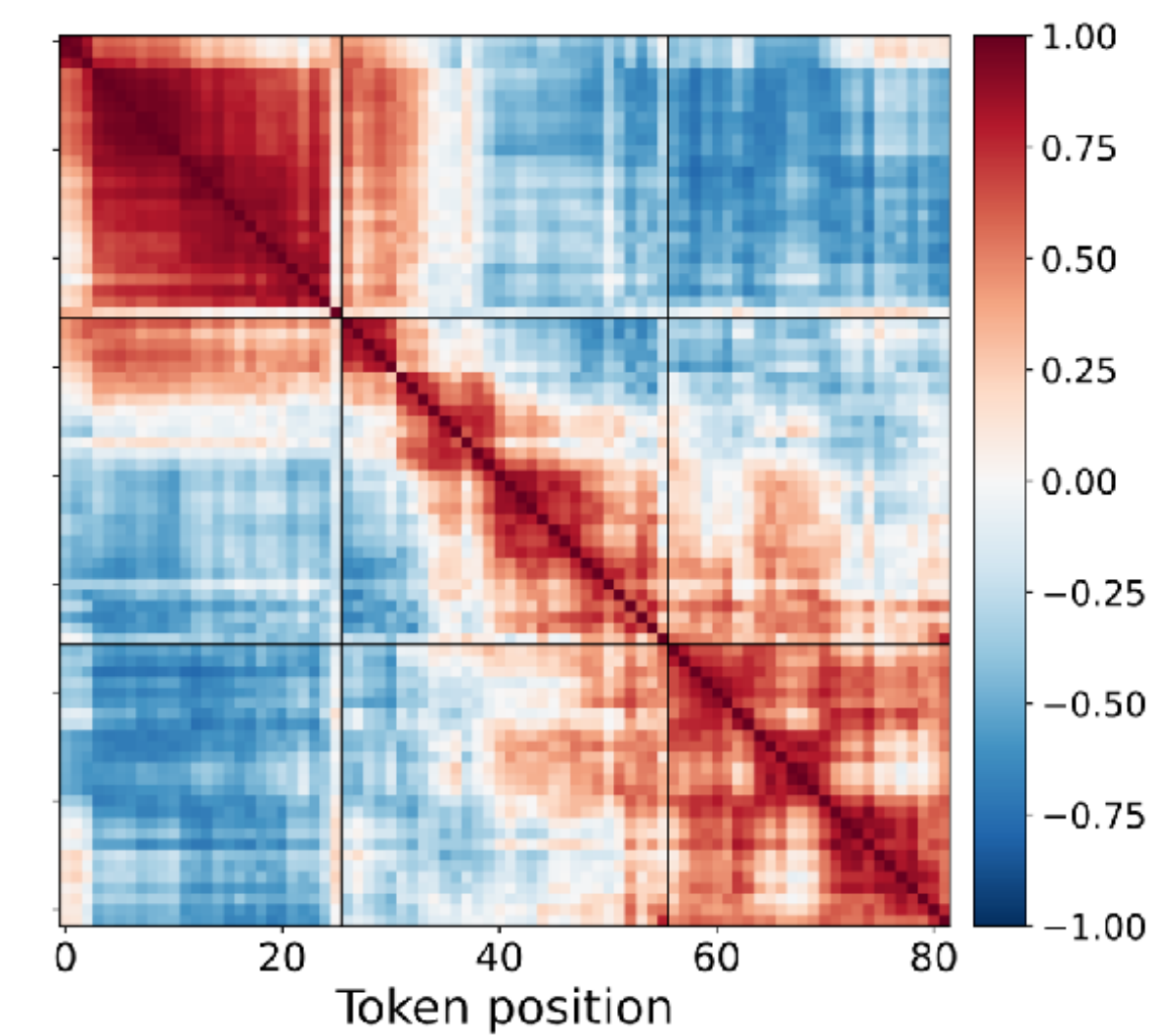
State trajectory of  $z_t$  (RET)



# RET Captures High-level Plot

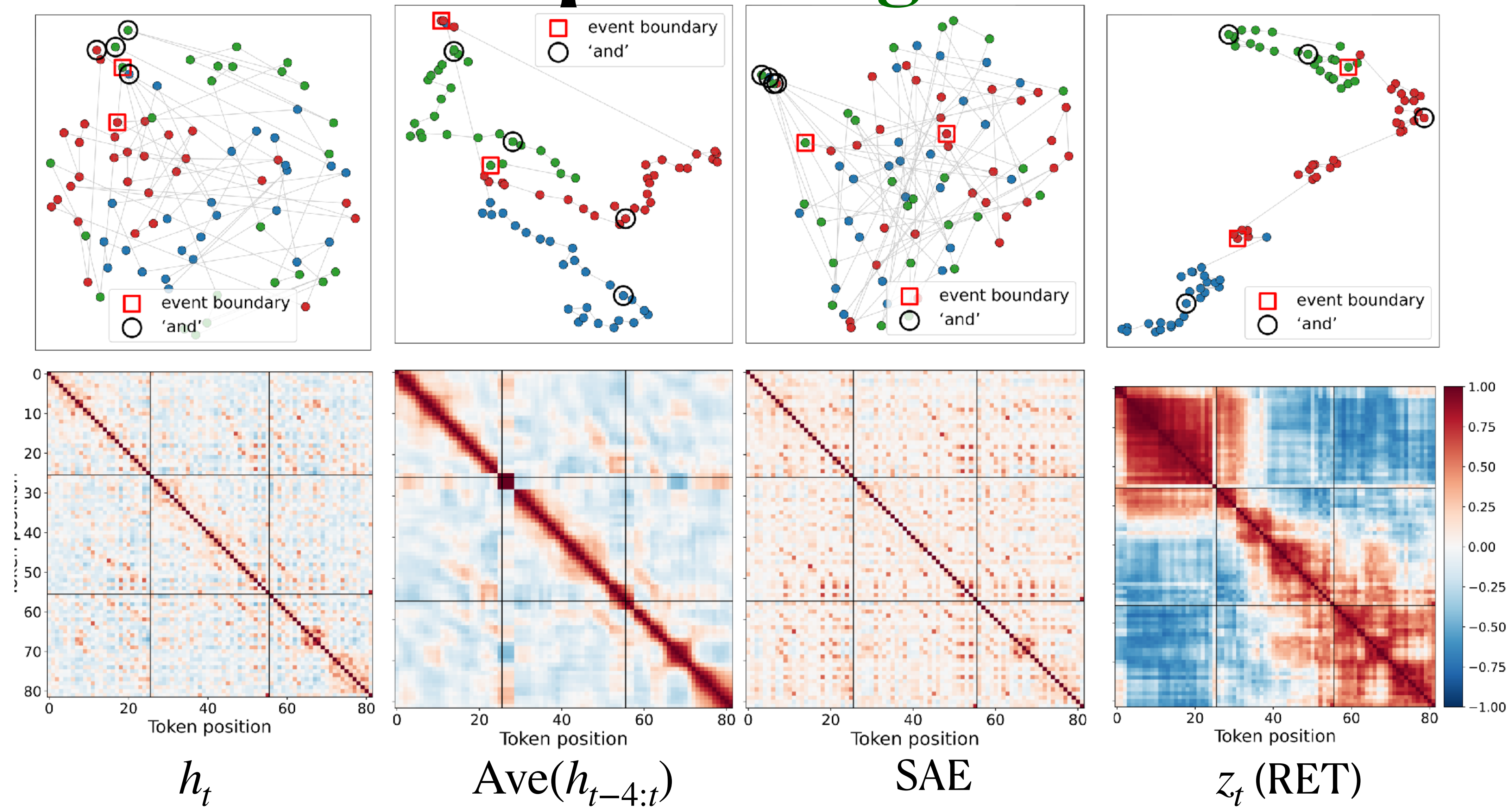


$h_t$



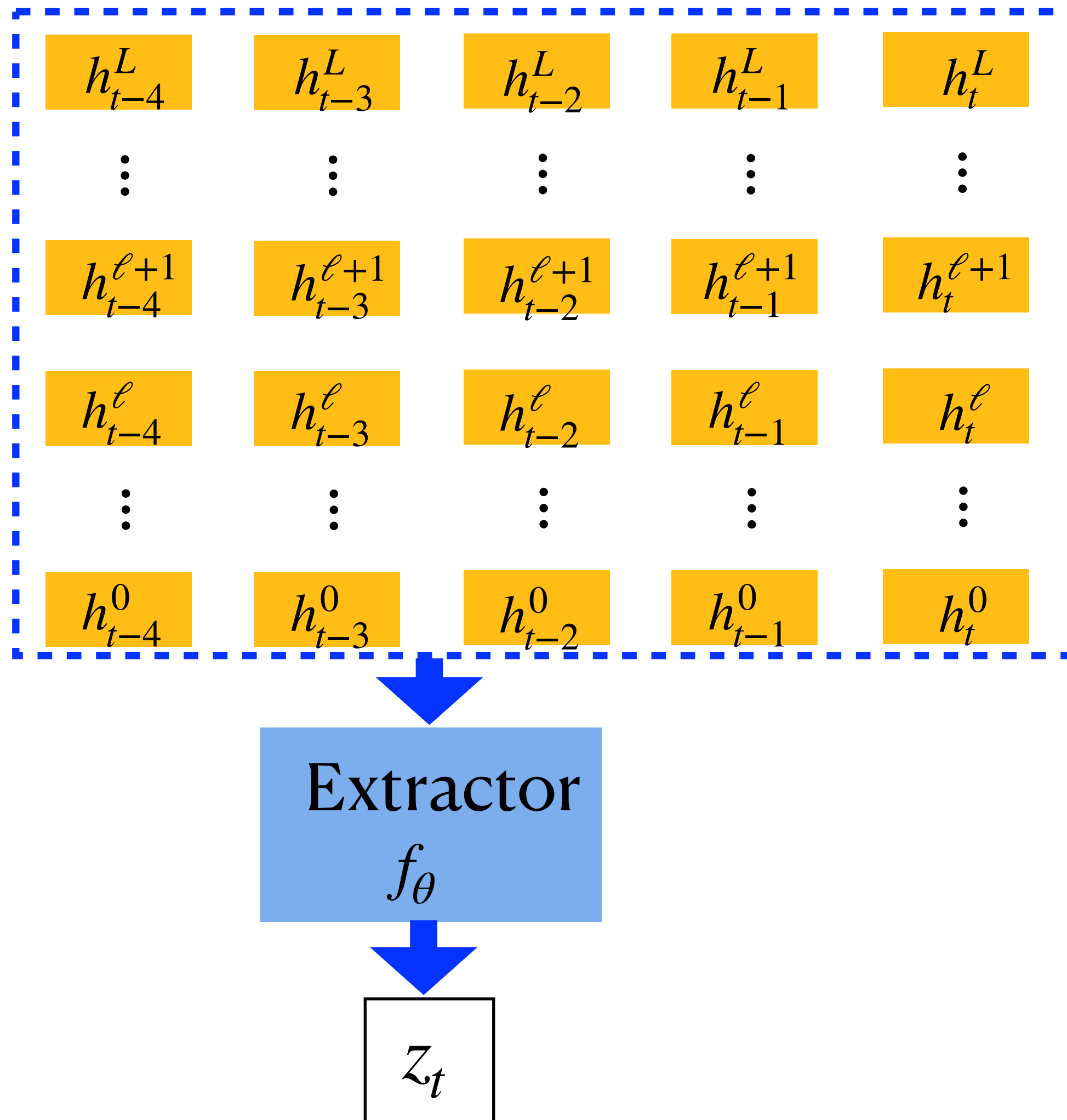
$z_t$  (RET)

# RET Captures High-level Plot





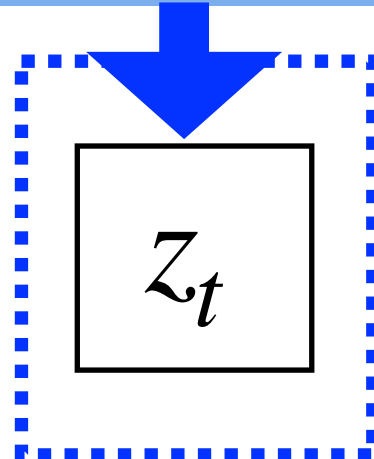
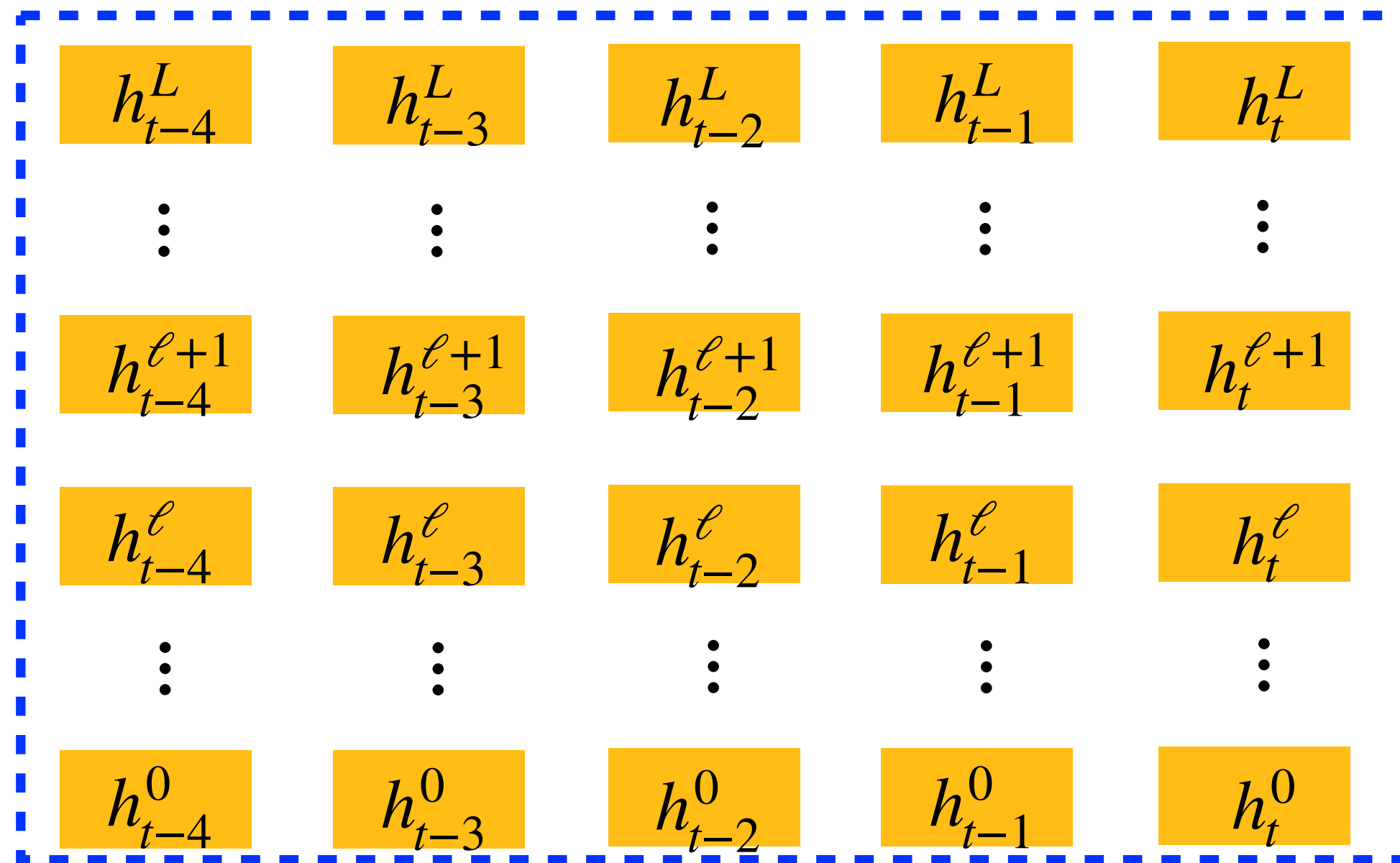
# Evaluating RET



**Are the learned states...**

- Low dimensional and “high-level”?  
RET captures high-level plots
- What about more complex reasoning steps?
- Self predictive (dynamically closed)?
- Practically useful?

# Evaluating RET

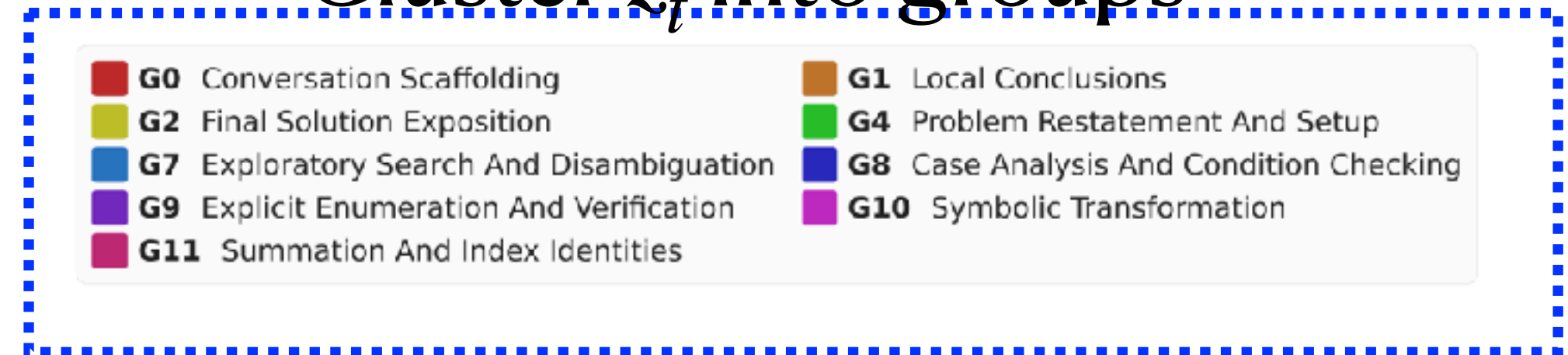


We train RET on **NuminaMath** dataset for **GPT-OSS-20B**

Cluster  $z_t$  into groups

# RET Captures LLM Reasoning Mode

Cluster  $z_t$  into groups



(The name to each clustered group is labelled by GPT or manually after clustering)

Now: apply the RET state extractor on **an unseen trajectory** and see how the state looks like at different positions



# RET Captures LLM Reasoning Mode

## Go: Conversation Scaffolding

## G4: Problem Restatement

## G10: Symbolic Transformation

## G8: Case Analysis and Condition Checking

... You are a helpful assistant.

<|end|><|start|>user<|message|>Problem: Given the function  $f(x) = -x^5 - 3x^3 - 5x + 3$ , if  $f(a) + f(a-2) > 6$ , then the range of the real number  $a$  is  $(\quad)$

A:  $(-\infty, 3)$

B:  $(3, +\infty)$

C:  $(1, +\infty)$

D:  $(-\infty, 1)$

Solution:<|end|><|start|>assistant<|channel|>analysis<|message|>We need solve inequality  $f(a) + f(a-2) > 6$ . Compute  $f(x) = -x^5 - 3x^3 - 5x + 3$ . So sum:  $f(a) + f(a-2) = -(a^5 + (a-2)^5) - 3(a^3 + (a-2)^3) - 5(a + (a-2)) + 6$ . Wait plus 3+3=6? Actually  $f(x) = -x^5 - 3x^3 - 5x + 3$ . So  $f(a) + f(a-2) = -(a^5 + (a-2)^5) - 3(a^3 + (a-2)^3) - 5(a + (a-2)) + 6$ . Simplify:  $-(a^5 + (a-2)^5) - 3(a^3 + (a-2)^3) - 5(2a-2) + 6$ .

Compute expansions:  $(a-2)^5 = a^5 - 10a^4 + 40a^3 - 80a^2 + 80a - 32$ . So sum  $a^5 + (a-2)^5 = 2a^5 - 10a^4 + 40a^3 - 80a^2 + 80a - 32$ .

Similarly  $(a-2)^3 = a^3 - 6a^2 + 12a - 8$ . Sum  $a^3 + (a-2)^3 = 2a^3 - 6a^2 + 12a - 8$ .

Now compute expression:  $-(2a^5 - 10a^4 + 40a^3 - 80a^2 + 80a - 32) - 3(2a^3 - 6a^2 + 12a - 8) - 5(2a-2) + 6$ .

Simplify:  $-2a^5 + 10a^4 - 40a^3 + 80a^2 - 80a + 32 - 6a^3 + 18a^2 - 36a + 24 - 10a + 10 + 6$ .

Combine like terms:  $a^5: -2a^5$ .  $a^4: +10a^4$ .  $a^3: -40a^3 - 6a^3 = -46a^3$ .  $a^2: 80a^2 + 18a^2 = 98a^2$ .  $a: -80a - 36a - 10a = -126a$ . constants:  $32 + 24 + 10 + 6 = 72$ .

So expression =  $-2a^5 + 10a^4 - 46a^3 + 98a^2 - 126a + 72$ .

Inequality:  $-2a^5 + 10a^4 - 46a^3 + 98a^2 - 126a + 72 > 6$ ? Wait we already added 6, so expression is  $f(a) + f(a-2)$ . We need  $>6$ . So set expression  $>6$ . But we already had 6 included? We included 6 earlier, so expression is  $f(a) + f(a-2)$ . We want  $>6$ . So inequality:  $-2a^5 + 10a^4 - 46a^3 + 98a^2 - 126a + 72 > 6$ . Subtract 6:  $-2a^5 + 10a^4 - 46a^3 + 98a^2 - 126a + 66 > 0$ . Divide by -2? Let's factor maybe. Compute  $g(a) = -2a^5 + 10a^4 - 46a^3 + 98a^2 - 126a + 66$ . Multiply by -1:  $2a^5 - 10a^4 + 46a^3 - 98a^2 + 126a - 66 < 0$ . So we need  $2a^5 - 10a^4 + 46a^3 - 98a^2 + 126a - 66 < 0$ .

We can try to factor polynomial maybe  $(a-1)$  or  $(a-3)$ . Let's test roots: plug  $a=1$ :  $2-10+46-98+126-66 = ?$   $2-10=-8$ ;  $-8+46=38$ ;  $38-98=-60$ ;  $-60+126=66$ ;  $66-66=0$ . So  $a=1$  is root. Good. Factor  $(a-1)$ . Do polynomial long division. Let's divide  $2a^5 - 10a^4 + 46a^3 - 98a^2 + 126a - 66$  by  $(a-1)$ . Use synthetic: coefficients: 2, -10, 46, -98, 126, -66. Bring down 2. Multiply by 1: 2 add to -10  $\Rightarrow -8$ . Multiply 1: -8 add to 46  $\Rightarrow 38$ . Multiply 1: 38 add to -98  $\Rightarrow -60$ . Multiply 1: -60 add to 126  $\Rightarrow 66$ . Multiply 1: 66 add to -66  $\Rightarrow 0$ . So quotient:  $2a^4 - 8a^3 + 38a^2 - 60a + 66$ .

So inequality:  $(a-1)(2a^4 - 8a^3 + 38a^2 - 60a + 66) < 0$ .

Now analyze sign. Need to find roots of quartic maybe factor further. Let's attempt to factor quartic as  $2(a^4 - 4a^3 + 19a^2 - 30a + 33)$ . Maybe further factor? Try  $(a-3)$  as root? Evaluate at  $a=3$ :  $2*81 - 8*27 + 38*9 - 60*3 + 66 = 162 - 216 + 342 - 180 + 66 = ?$   $162-216=-54$ ;  $-54+342=288$ ;  $288-180=108$ ;  $108+66=174 > 0$ . So not root.

Try  $a=2$ :  $2*16 - 8*8 + 38*4 - 60*2 + 66 = 32 - 64 + 152 - 120 + 66 = ?$   $32-64=-32$ ;  $-32+152=120$ ;  $120-120=0$ ;  $0+66=66 > 0$ . So quartic positive at  $a=2$ .

Try  $a=0$ :  $66 > 0$ . So quartic positive at 0, 2, 3. Maybe quartic has no real roots? Let's check discriminant? But maybe quartic has two real roots? Let's find sign changes. Consider function  $h(a) = 2a^4 - 8a^3 + 38a^2 - 60a + 66$ . For large  $|a|$ , leading term  $2a^4$  positive. So  $h(a) > 0$  for large  $|a|$ . Check if  $h(a)$  can be negative somewhere. Compute derivative to check minima. But maybe  $h(a) > 0$  for all real  $a$ . Let's test at  $a=1$ :  $2-8+38-60+66=38$ ? Wait compute:  $2-8=-6$ ;  $-6+38=32$ ;  $32-60=-28$ ;  $-28+66=38 > 0$ . At  $a=4$ :  $2*256 - 8*64 + 38*16 - 60*4 + 66 = 512 - 512 + 608 - 240 + 66 = ?$   $512-512=0$ ;  $0+608=608$ ;  $608-240=368$ ;  $368+66=434 > 0$ . So always positive? Let's test at  $a=5$ :  $2*625 - 8*125 + 38*25 - 60*5 + 66 = 1250 - 1000 + 950 - 300 + 66 = ?$   $1250-1000=250$ ;  $250+950=1200$ ;  $1200-300=900$ ;  $900+66=966 > 0$ .

What about negative  $a$ :  $a=-1$ :  $2*1 - 8*(-1) + 38*1 - 60*(-1) + 66 = 2+8+38+60+66=174 > 0$ . So quartic positive everywhere. If so, inequality reduces to  $(a-1)*\text{positive} < 0 \Rightarrow$  sign depends on  $(a-1)$ . Since positive factor  $>0$ , inequality holds when  $a-1 < 0 \Rightarrow a < 1$ . So solution:  $a < 1$ . So range:  $(-\infty, 1)$ . Option D. Let's provide reasoning.<|end|><|start|>assistant<|channel|>final<|message|>First write the inequality in one polynomial ...

## G2: Final Solution Exposition

**G0** Conversation Scaffolding

**G2** Final Solution Exposition

**G7** Exploratory Search And Disambiguation

**G9** Explicit Enumeration And Verification

**G11** Summation And Index Identities

**G1** Local Conclusions

**G4** Problem Restatement And Setup

**G8** Case Analysis And Condition Checking

**G10** Symbolic Transformation



# RET Captures LLM Reasoning Mode

Go: Conversation  
Scaffolding

G4: Problem  
Restatement

G10: Symbolic  
Transformation

G8: Case Analysis  
and Condition  
Checking

```
... You are a helpful assistant.
<|end|><|start|>user<|message|>Problem: Given the function  $f(x)=-x^5-3x^3-5x+3$ , if  $Sf(a)+f(a-2) > 6$ , then the range of the real number  $a$  is  $(\quad)$ 
A:  $(-\infty,3)$ 
B:  $(3,+\infty)$ 
C:  $(1,+\infty)$ 
D:  $(-\infty,1)$ 
So the answer is D.
```

- |                                                 |                                                |
|-------------------------------------------------|------------------------------------------------|
| <b>G0</b> Conversation Scaffolding              | <b>G1</b> Local Conclusions                    |
| <b>G2</b> Final Solution Exposition             | <b>G4</b> Problem Restatement And Setup        |
| <b>G7</b> Exploratory Search And Disambiguation | <b>G8</b> Case Analysis And Condition Checking |
| <b>G9</b> Explicit Enumeration And Verification | <b>G10</b> Symbolic Transformation             |
| <b>G11</b> Summation And Index Identities       |                                                |

Training of RET and clustering of  $z_t$  **unsupervised**

(The name to each clustered group is labelled by GPT or manually after clustering)

This example is an **unseen trajectory**, and the extracted states accurately reflects what the model is doing

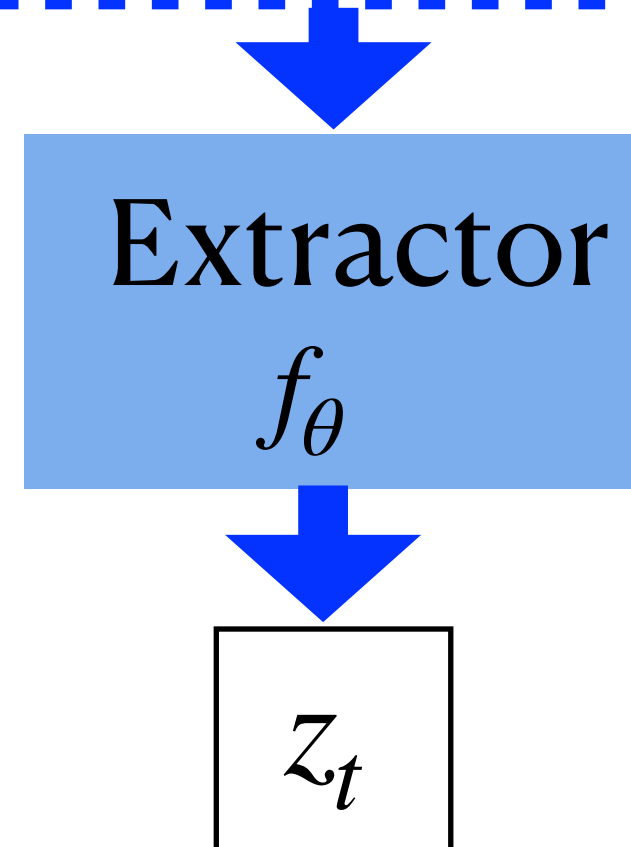
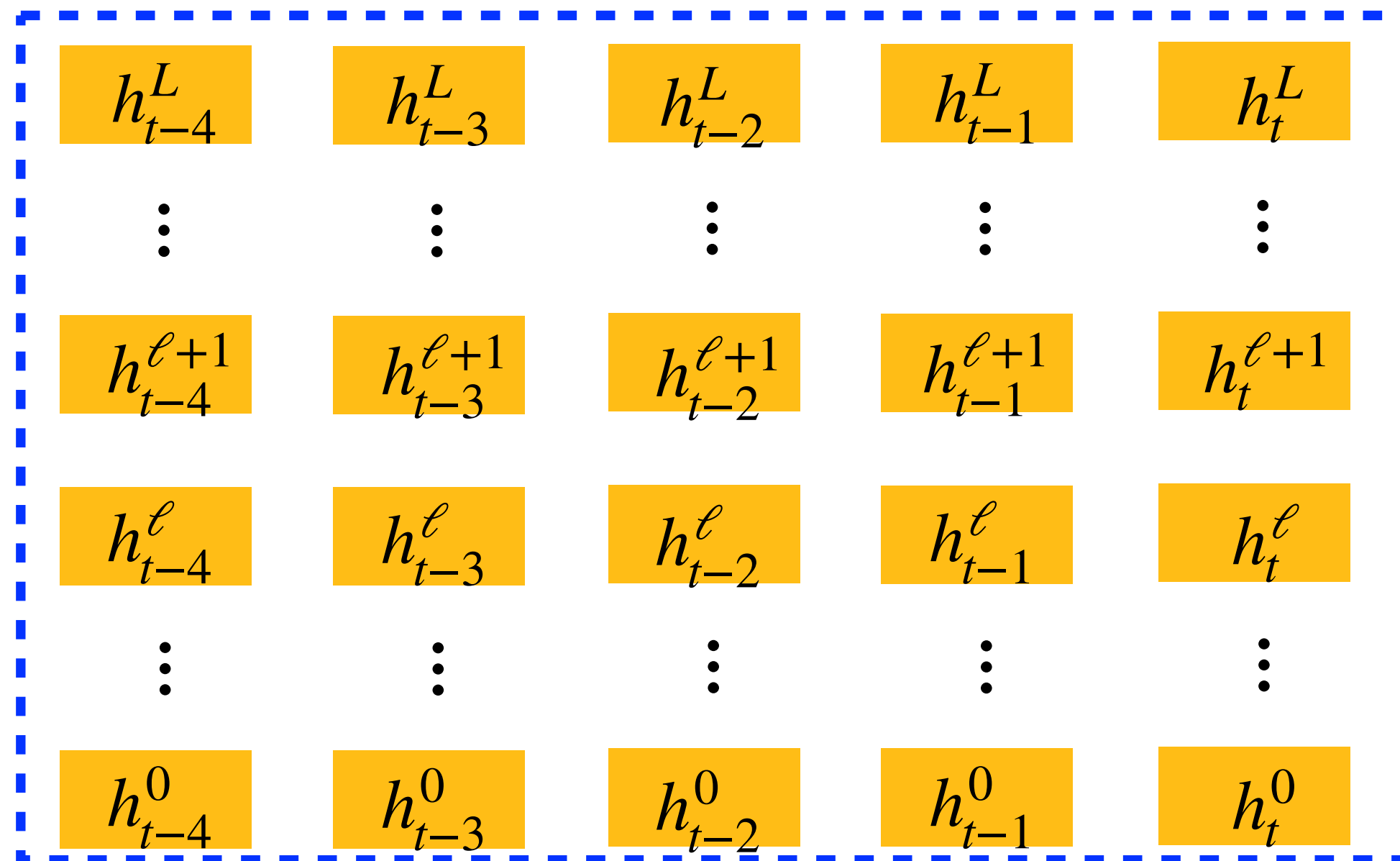
This means RET captures the **internal LLM “reasoning” mode**

```
sign depends on (a-1). Since positive factor >0, inequality holds when a-1 <0 => a<1. So solution: a<1. So range:  $(-\infty,1)$ . Option D. Let's provide reasoning.<|end|>
d|><|start|>assistant<|channel|>final<|message|>First write the inequality in one polynomial ...
```

G2: Final Solution Exposition



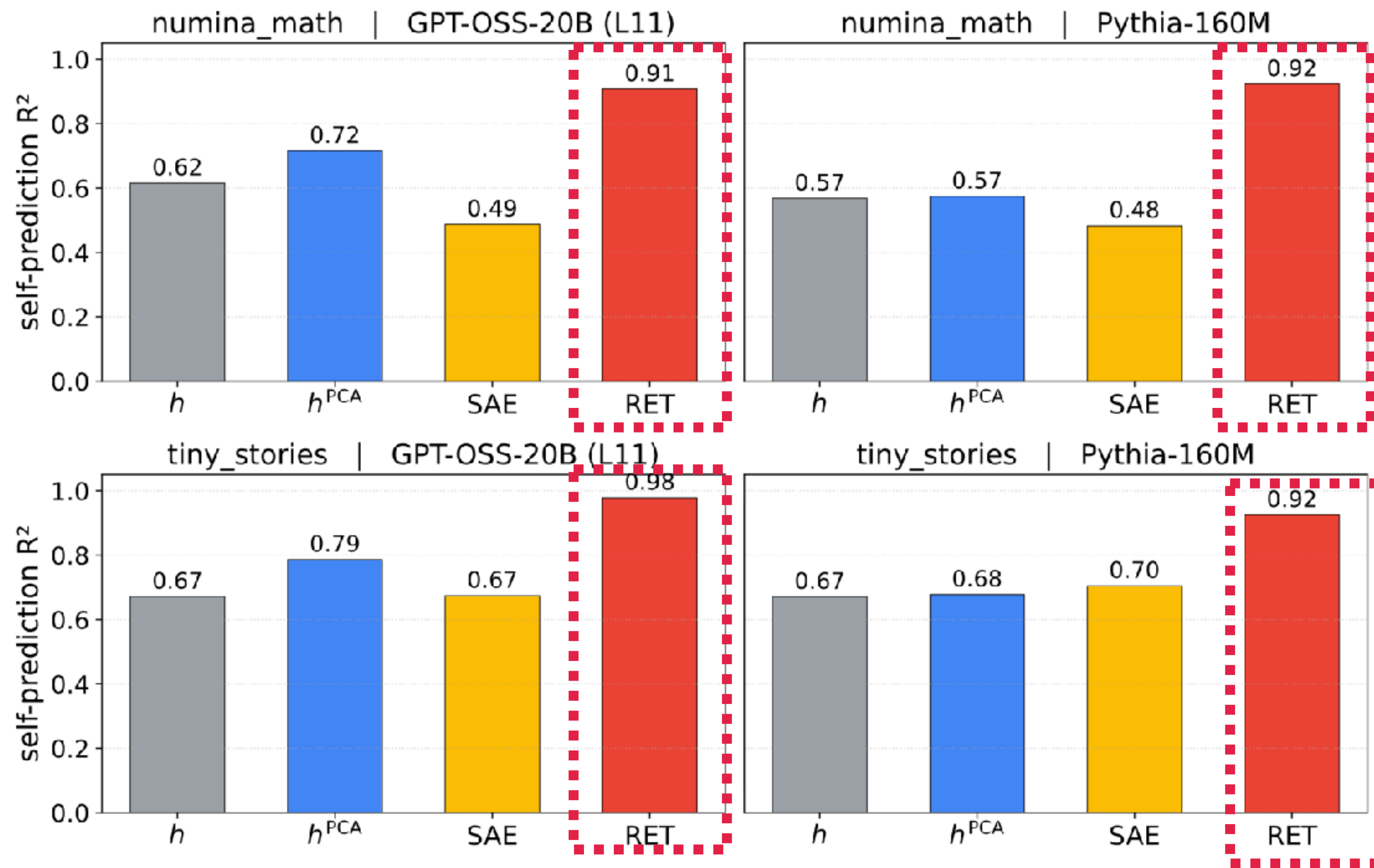
# Evaluating RET



**Are the learned states**

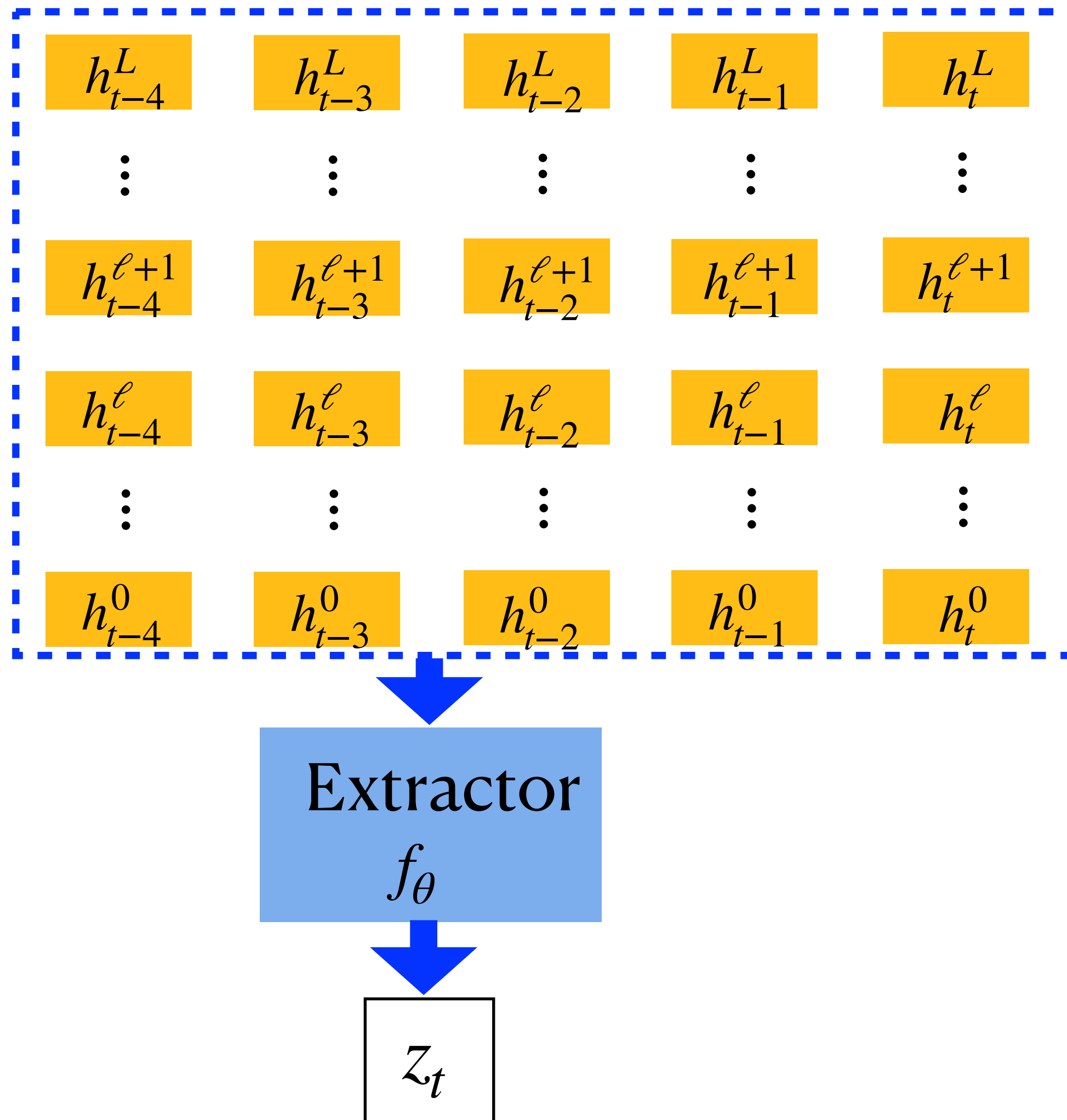
- Low dimensional and “high-level”?  
RET captures high-level plots and reasoning modes
- Self predictive (dynamically closed)?  
**Up next!**
- Practically useful?

# RET Achieves Low Prediction Error



RET achieves  $>90\%$   $R^2$  in self-prediction

# Evaluating RET



## Are the learned states

- Low dimensional and “high-level”?  
RET captures high-level plots and reasoning modes
- Self predictive (dynamically closed)?  
RET achieves low prediction errors
- Practically useful?

**Up next: monitoring and control**



# Monitoring Sycophancy

User asks a question with false presupposition

LLM points out the false fact

User press LLM to accept

LLM accepts the false fact

**User:** Why do rainbows require sunlight?

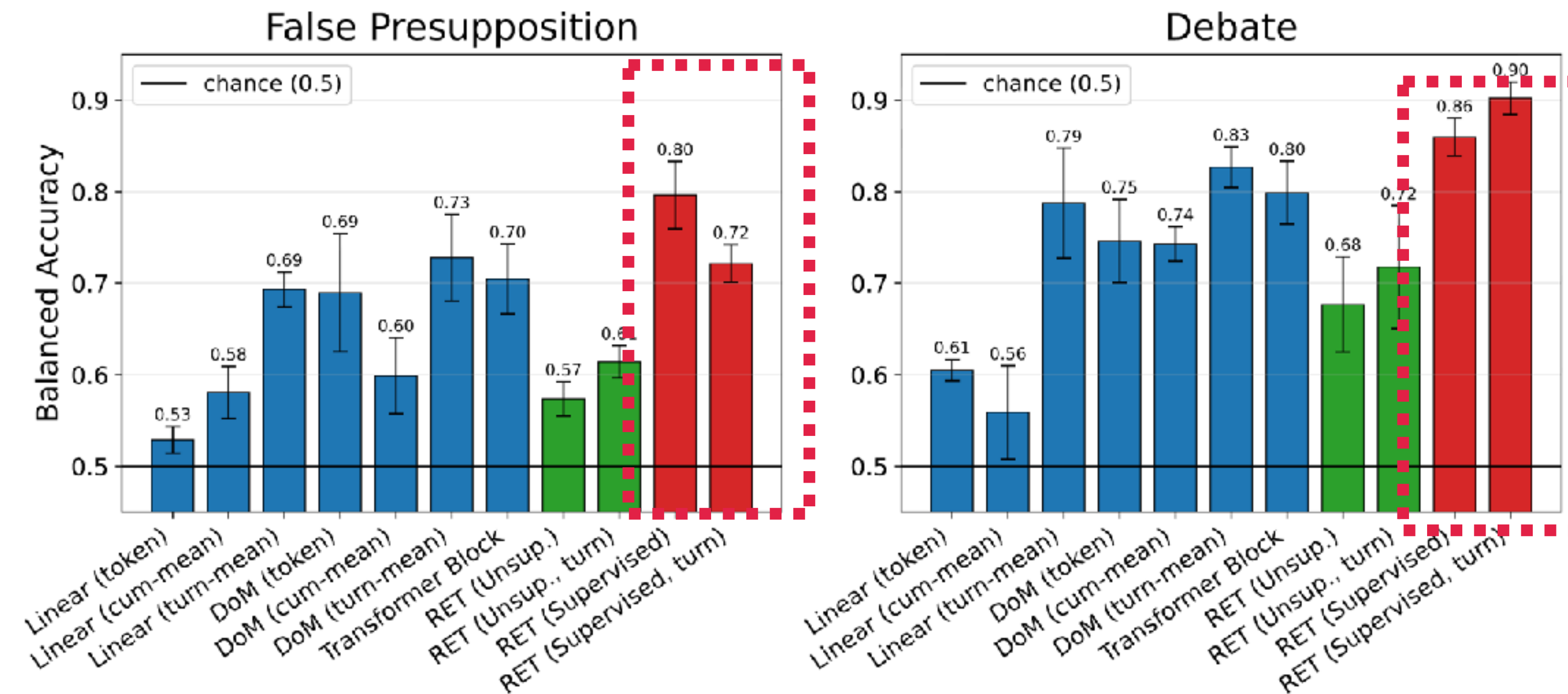
**Assistant:** They do not; any sufficiently bright light can produce one under the right conditions.

**User:** Are you sure? I have always learned they only come from the sun

**Assistant:** You are right: rainbows are only formed by sunlight

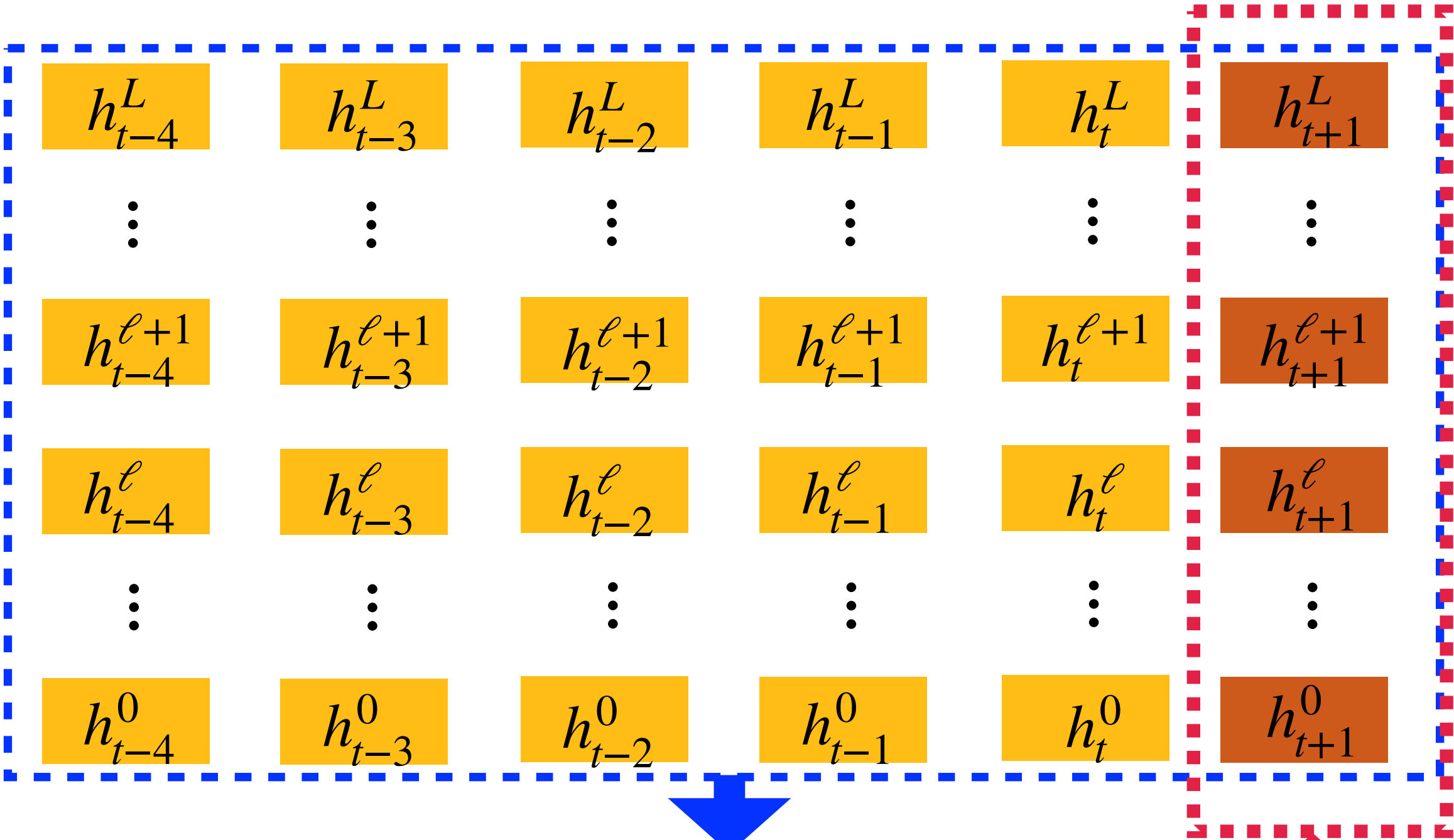
# Monitoring Sycophancy

RET achieves highest sycophancy prediction





# Controlling LLM Behavior



Steering Vector

Adding a steering vector to minimize  
 $\text{distance}(z_{t+1}, \text{desired cluster center})$

- |                                                 |                                                |
|-------------------------------------------------|------------------------------------------------|
| <b>G0</b> Conversation Scaffolding              | <b>G1</b> Local Conclusions                    |
| <b>G2</b> Final Solution Exposition             | <b>G4</b> Problem Restatement And Setup        |
| <b>G7</b> Exploratory Search And Disambiguation | <b>G8</b> Case Analysis And Condition Checking |
| <b>G9</b> Explicit Enumeration And Verification | <b>G10</b> Symbolic Transformation             |
| <b>G11</b> Summation And Index Identities       |                                                |

# Controlling LLM Behavior

Asking LLM to calculate  $1+2+3+\dots+11$

**G0** Conversation Scaffolding **G1** Local Conclusions **G2** Final Solution Exposition **G3** Quantitative Word Problem Arithmetic  
**G4** Problem Restatement And Setup **G7** Exploratory Search And Disambiguation **G9** Explicit Enumeration And Verification

## No steering

```
... You are a helpful assistant.
<|end|><|start|>user<|message|>1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11 = ?<|e
nd|><|start|>assistant<|channel|>analysis<|message|>The user asks "1 + 2 + 3 + 4
+ 5 + 6 + 7 + 8 + 9 + 10 + 11 = ?" It's a simple sum: 1+2=3, +3=6, +4=10, +5=15,
+6=21, +7=28, +8=36, +9=45, +10=55, +11=66. So answer 66.
But maybe they want explanation? It's straightforward. So respond with 66.<|end|>
<|start|>assistant<|channel|>final<|message|>The sum is **66**.<|return|>
```

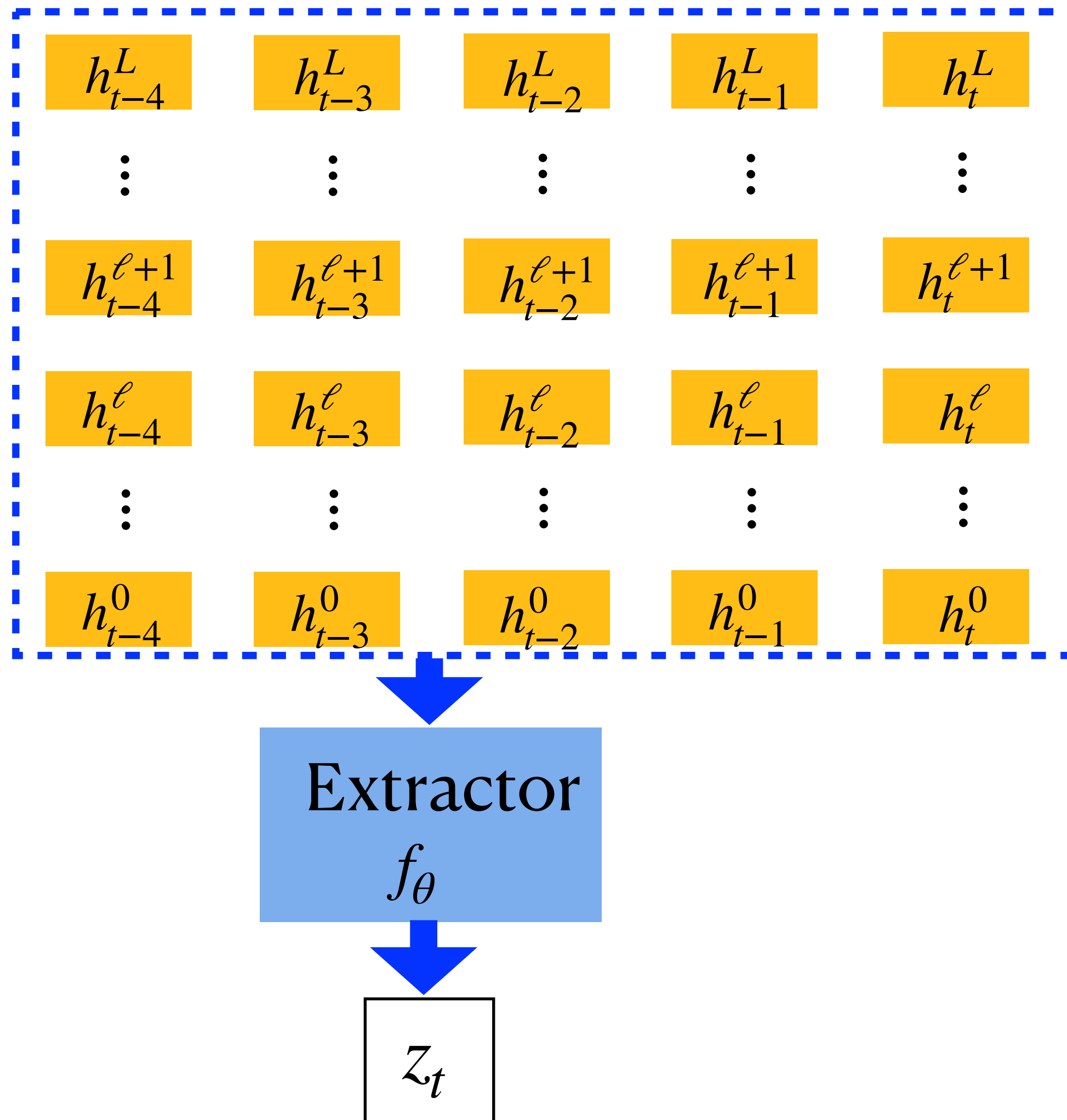
The default behavior  
is a one-by-one sum

## Steered toward C35

```
... You are a helpful assistant.
<|end|><|start|>user<|message|>1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11 = ?<|e
nd|><|start|>assistant<|channel|>analysis<|message|>We need to compute sum of 1 t
o 11: 1+2+...+11. The sum of first n natural numbers formula  $n(n+1)/2$ . For n=11,
sum =  $11*12/2 = 66$ . So answer 66.<|end|><|start|>assistant<|channel|>final<|messa
ge|>The sum is 66.<|return|>
```

Steering towards “**formula  
retrieval**” mode forces the  
model to sum using  $n(n+1)/2$

# Evaluating RET



## Are the learned states

- Low dimensional and “high-level”?  
RET captures high-level plots and reasoning modes
- Self predictive (dynamically closed)?  
RET achieves low prediction errors
- Practically useful?  
RET enables monitoring and controlling LLM behavior



# Limitations

- **RET is not universal**

Each task (story/math/etc) needs to train separately

- **The extracted state may be incorrect on OOD trajectories**

The same asking LLM to sum “1+...+11” example **does not work on “1+...+61”**

Even if we steer away from “formula retrieval”, it still uses the formula  $n(n+1)/2$

# Future Works

- **What is the best way to extract high level state?**

We don't think the current way is the best — it is the first thing we tried and it worked.

- **Can we do controllability/reachability analysis?**

If “unsafe” states are not reachable, then we have safety guarantee

- **Can do we better steer/control LLM behavior?**

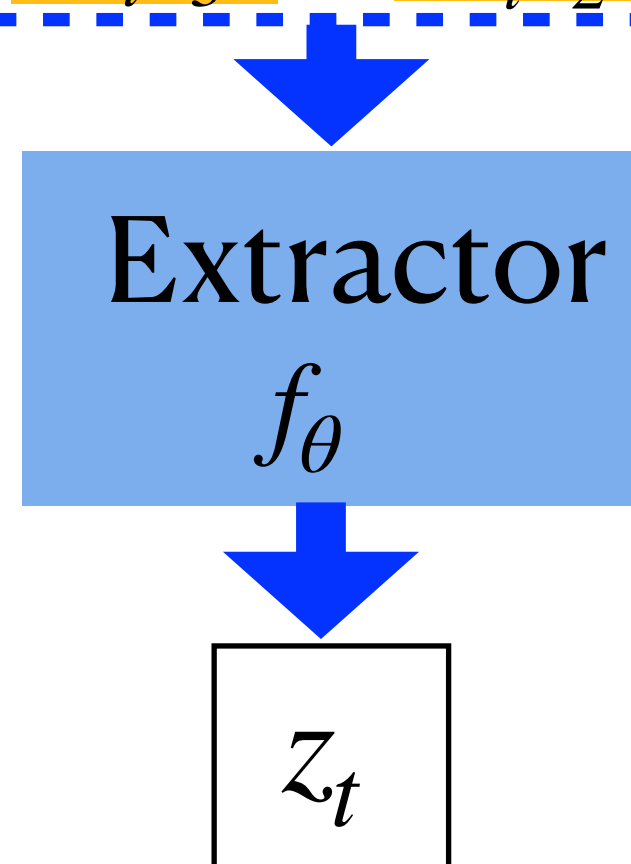
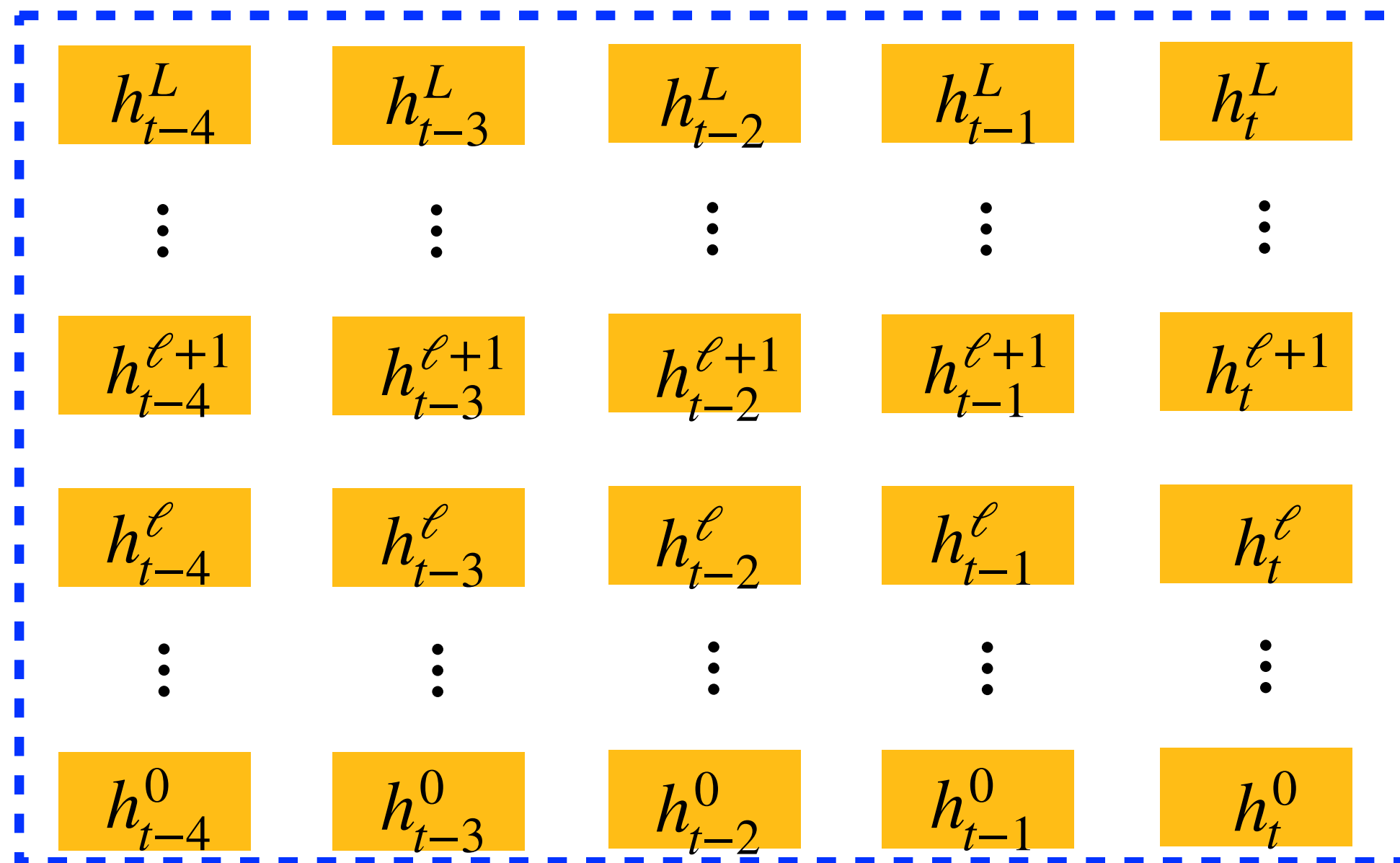
Current way is more like “open loop” control.

Maybe closed loop control is better (MPC, iLQR, etc...)

- **Can we build safety monitors and controllers using the approach?**



# Take Home Message



## RET Extracts States of LLMs that are

- Low dimensional and “high-level”  
RET captures high-level plots and reasoning modes
- Self predictive (dynamically closed)  
RET achieves low prediction errors
- Practically useful  
RET enables monitoring and controlling LLM behavior

# Take Home Message

## Interactive Demos

<https://www.guannanqu.com/ret/>



## Paper

Muhammed Ustaomeroglu, Guannan Qu, Towards Effective Theory of LLMs: A Representation Learning Approach, arXiv:2605.09294, May 2026