# Zero-shot Model-based Reinforcement Learning using Large Language Models

**Abdelhakim Benechehab**[1,2], **Youssef Attia El Hili**[1], Ambroise Odonnat[1,3], Oussama Zekri[4], Albert Thomas[1], Giuseppe Paolo[1], Maurizio Filippone[5], Ievgen Redko[1], Balázs Kégl[1]

[1] Huawei Noah's Ark Lab, Paris, France [2] Department of Data Science, EURECOM [3] Inria, Univ. Rennes 2, CNRS, IRISA [4] ENS Paris-Saclay [5] Statistics Program, KAUST

## TL;DR

- We propose **DICL**, a method to adapt **LLMs** for **dynamics** prediction in **MBRL**

## Problem Setup

- Given a trajectory $\mathcal{T} = (s_0, a_0, r_1, s_1, \ldots, r_{T-1}, s_{T-1}, a_{T-1})$, with $s_t \in \mathbb{R}^{d_s}$, $a_t = \pi(s_t) \in \mathbb{R}^{d_a}$, and $r_t \in \mathbb{R}$, $\rightarrow$ **predict future states** $s_T$

**LLMs for numerical data:**

- Time series tokenization

$$\underbrace{[0.2513, 5.2387, 9.7889]}_{\text{time series}} \rightarrow \underbrace{[1.5, 5.16, 8.5]}_{\text{rescaled}} \rightarrow \underbrace{\text{“150, 516, 850”}}_{\text{prompt}}$$

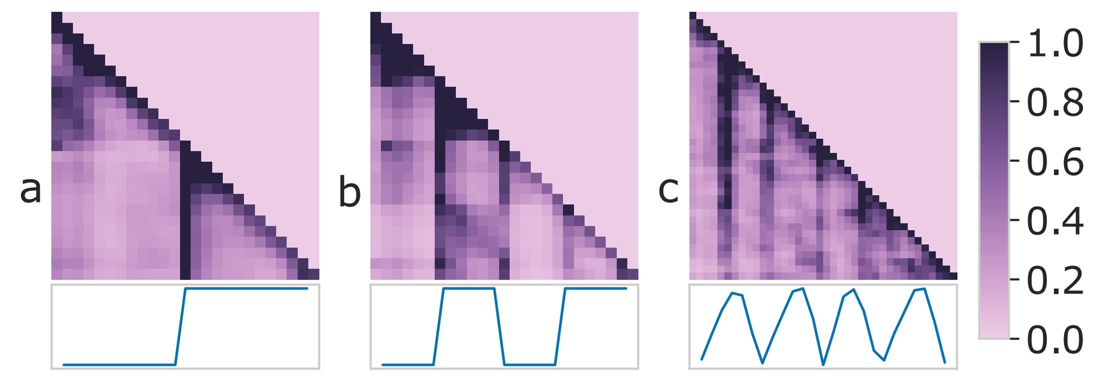
- In-Context Learning (ICL)

$$\underbrace{x_1, x_2, \ldots, x_{T-1},}_{\text{input}} \ \underbrace{x_T}_{\text{target}}$$

- Sampling the numbers sub-vocabulary $\mathcal{V}_{\text{num}}$

$$\{P(X_{i+1}|x_i, \ldots, x_0)\}_{i \le T} \leftarrow \text{softmax}(\text{logits}(\mathcal{V}_{\text{num}}))$$

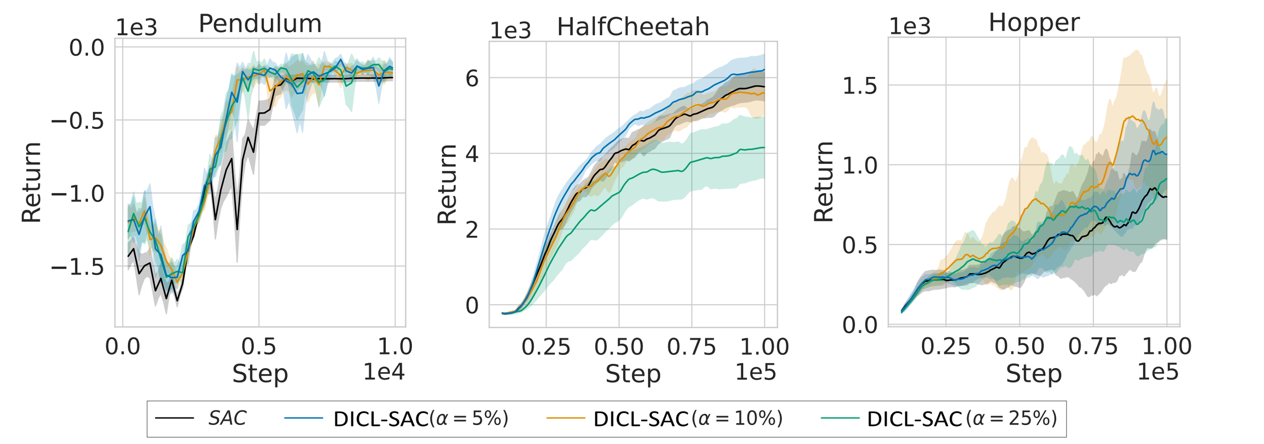
## Motivation: LLMs can perceive time patterns




▸ Visualizing **Llama 3-8B** attention matrices.

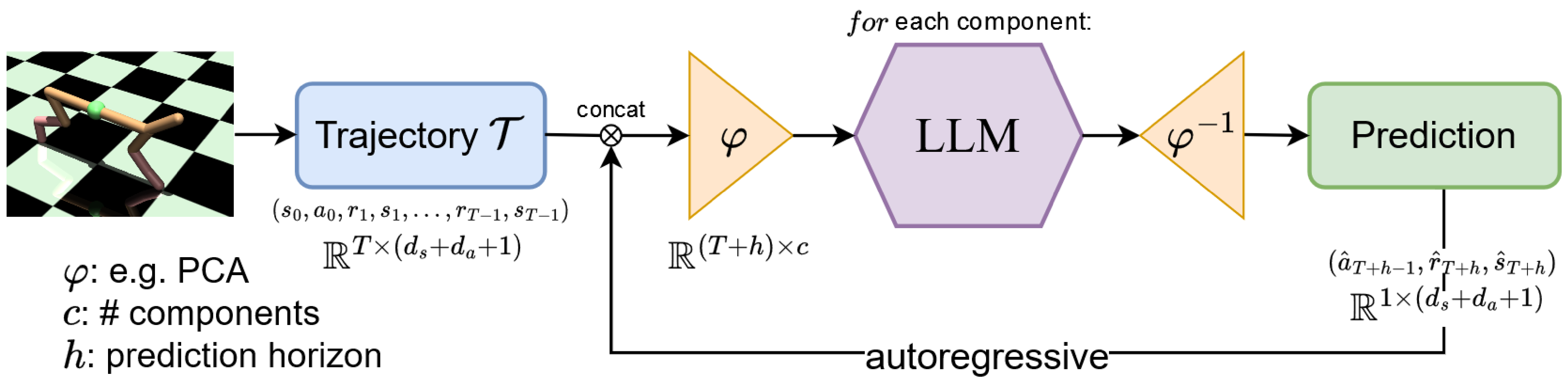
## Application: Data-augmented off-policy RL

- Augment **off-policy** RL (e.g. SAC) with **DICL**-generated transitions.




▸ **DICL-SAC** improves the **convergence speed** in the beginning of training

## Method: Disentangled In-context learning (DICL)



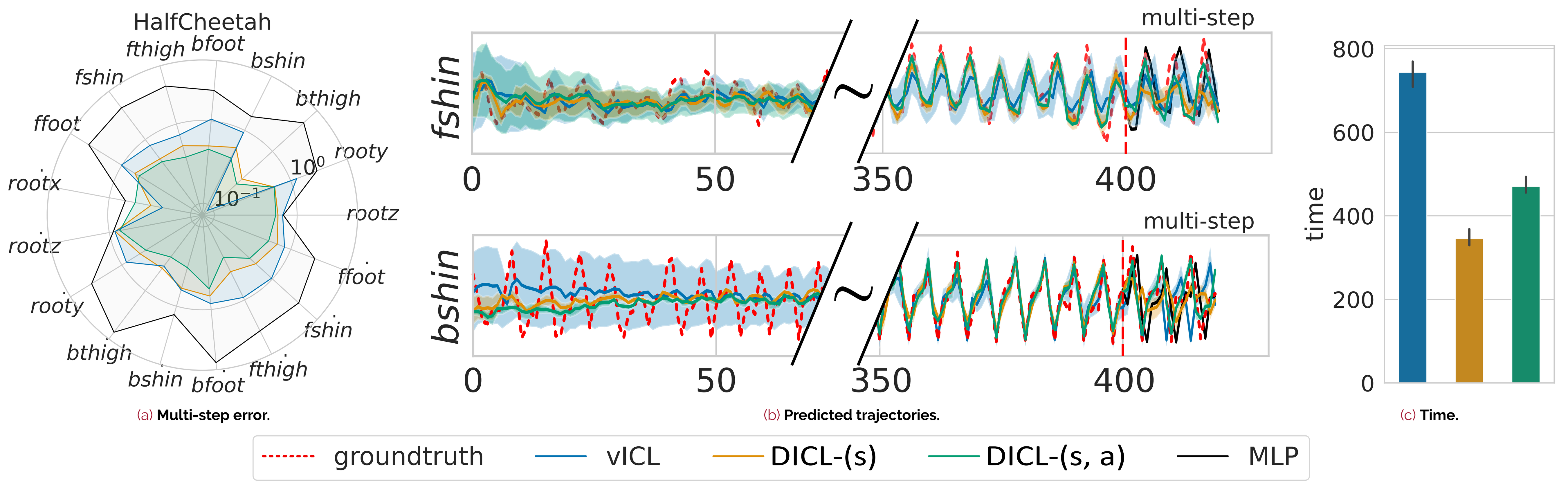

We solve two main challenges:

- Multivariate state $d_s > 1$
- Including the control signal $(a_i)_{i \le T}$

} **Feature-mixing transformation** $\varphi$

In practice, We use **PCA** as the $\varphi$ transformation

## Dynamics learning using DICL




(a) Multi-step error. (b) Predicted trajectories. (c) Time.

▸ **DICL improves upon the MLP baseline** in multi-step prediction error (MSE)

▸ The prediction **error gets smaller with** $T$, the in-context trajectory length

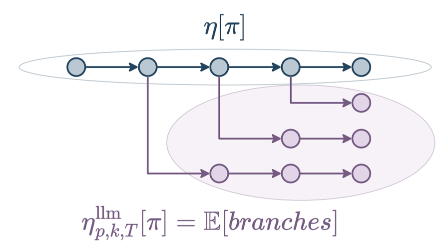
▸ **DICL** runs in **less time** thanks to PCA-enabled **dimensionality reduction**

## Return bound under DICL dynamics

**Theorem.**

- $T$ context length, $p$ branching proba, $k$ branch length
- $\varepsilon_{\text{llm}}$ LLM in-context prediction error

$$|\eta(\pi) - \eta^{\text{llm}}_{p,k,T}(\pi)| \le 2\frac{\gamma^T}{1-\gamma} r_{\max} k^2 p\, \varepsilon_{\text{llm}}(T)$$

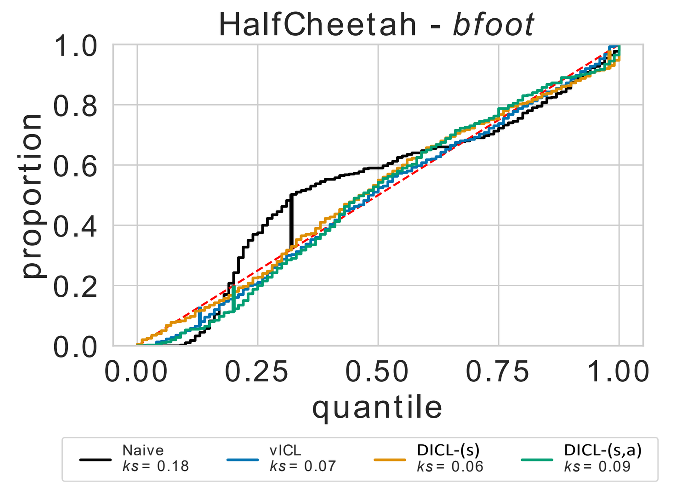



## More findings from our paper

→ Ablation study on **LLaMA** LLMs.

→ **DICL** for **Off-Policy Evaluation** (OPE) !!

→ LLMs are **well-calibrated** in-context forecasters.




## Take Home Message

**LLMs** are powerful foundation models trained on vast amounts of data
→ **DICL** is an effective way to adapt them to MBRL

## Main References

- **LLMTime, Gruver et al.** - Neurips 2023
  *Large Language Models Are Zero-Shot Time Series Forecasters*
- **Liu et al.** - EMNLP 2024
  *LLMs learn governing principles of dynamical systems, revealing an in-context neural scaling law*
- **Benechehab et al.** - SCOPE workshop @ ICLR 2025 (Follow-up work)
  *AdaPTS: Adapting Univariate Foundation Models to Probabilistic Multivariate Time Series Forecasting*

## Want to Know More?

paper & code

Abdelhakim Benechehab



Correspondence to abdelhakim.benechehab@gmail.com