# Emergence of meta-stable clustering in mean-field transformer models

ICLR 2025 Oral

*Giuseppe Bruno, Federico Pasqualotto, Andrea Agazzi*

UNIVERSITÄT BERN

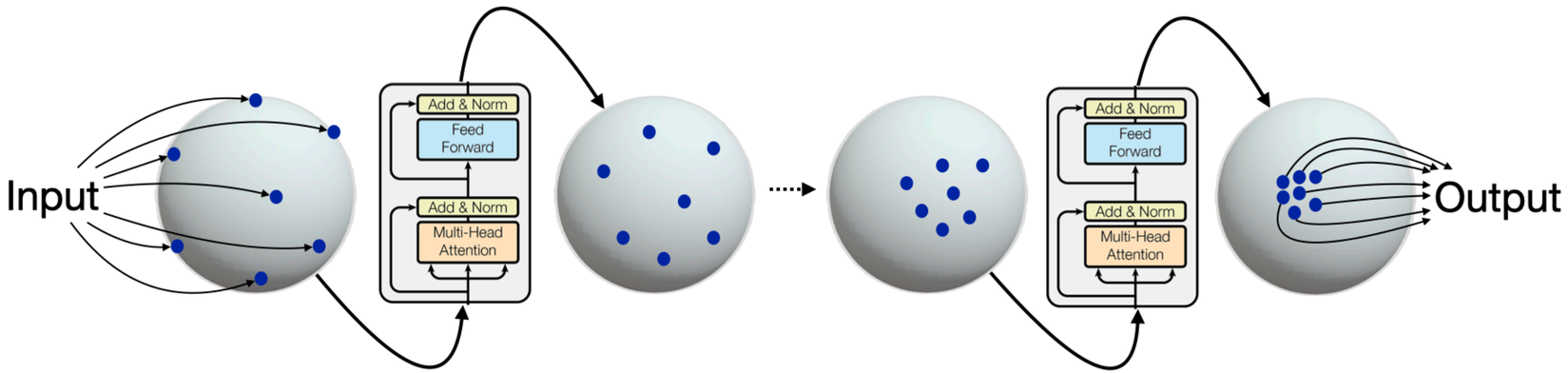
UC San Diego

ICLR

# Introduction

## Transformers as interacting particles systems

We build upon the mathematical **mean-field** model for **Transformers** dynamics introduced in [1, 2], modeling the evolution of $N$ tokens $x_i(k)_{i=1}^N$ on $\mathbb{S}^{d-1}$ via a system of coupled ODEs:

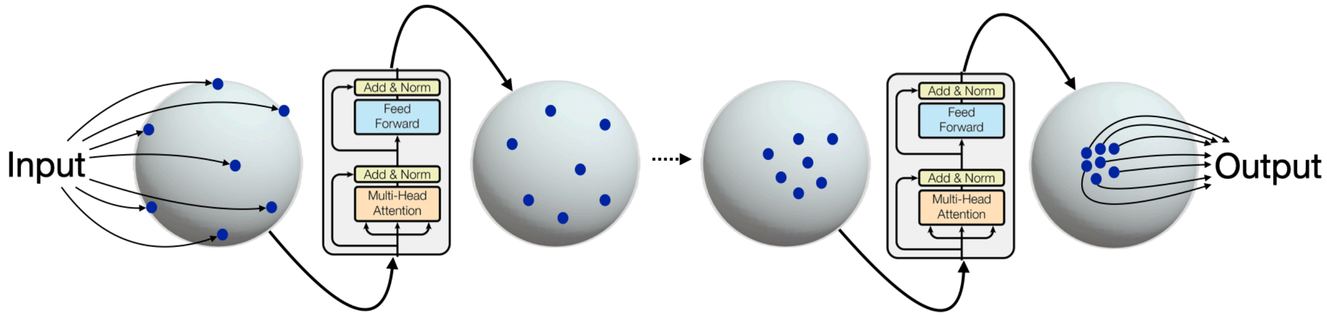



[1] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. arXiv preprint arXiv:2312.10794, 2023.
[2] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sink-formers: Transformers with doubly stochastic attention. In International Conference on Artificial Intelligence and Statistics, pages 3515–3530. PMLR,

# Introduction

## The infinite-depth limit

We build upon the mathematical **mean-field** model for **Transformers** dynamics introduced in [1, 2], modeling the evolution of $N$ tokens $x_i(k)_{i=1}^N$ on $\mathbb{S}^{d-1}$ via a system of coupled ODEs:



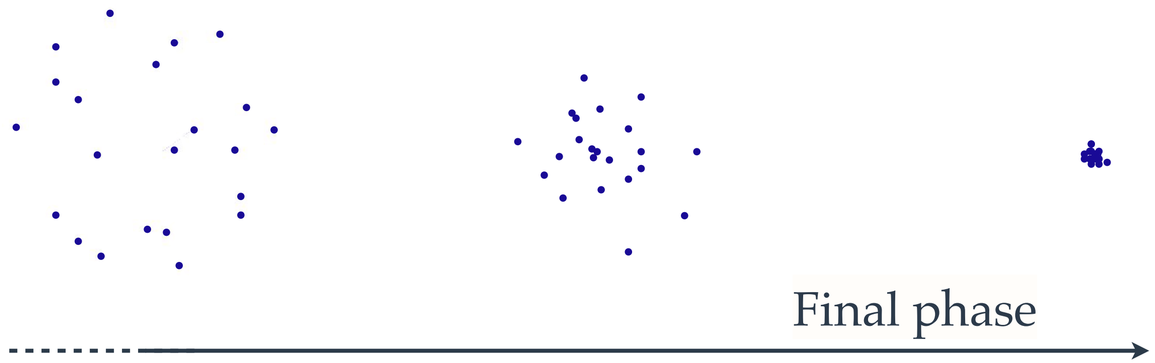

$$\begin{cases} \dot{x}_i(t) = P_{x_i(t)}\left(\frac{1}{Z_{\beta,i}(t)}\sum_{j=1}^{N} e^{\beta\langle Qx_i(t), Kx_j(t),\rangle} Vx_j(t)\right), \\ x_i(0) = x_i\,. \end{cases}$$

Interacting particles system


[1] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. arXiv preprint arXiv:2312.10794, 2023.
[2] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sink-formers: Transformers with doubly stochastic attention. In International Conference on Artificial Intelligence and Statistics, pages 3515–3530. PMLR,

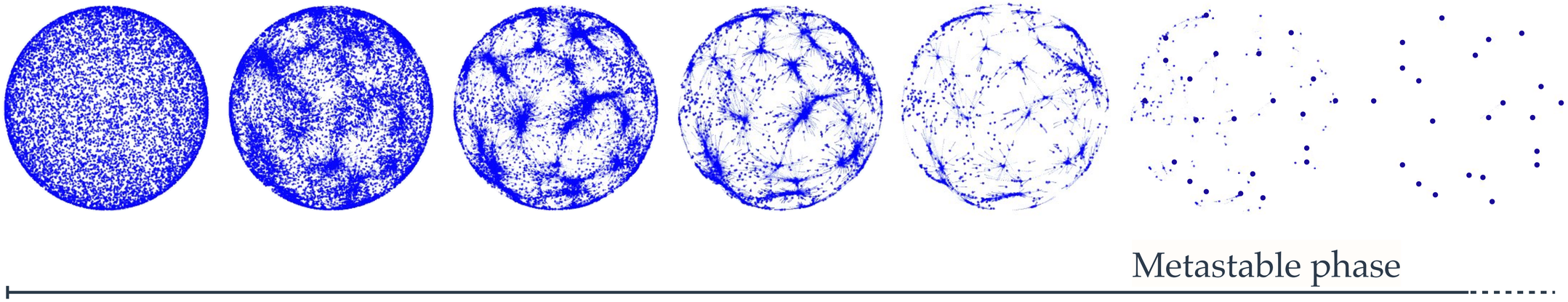
# Dynamic meta-stability

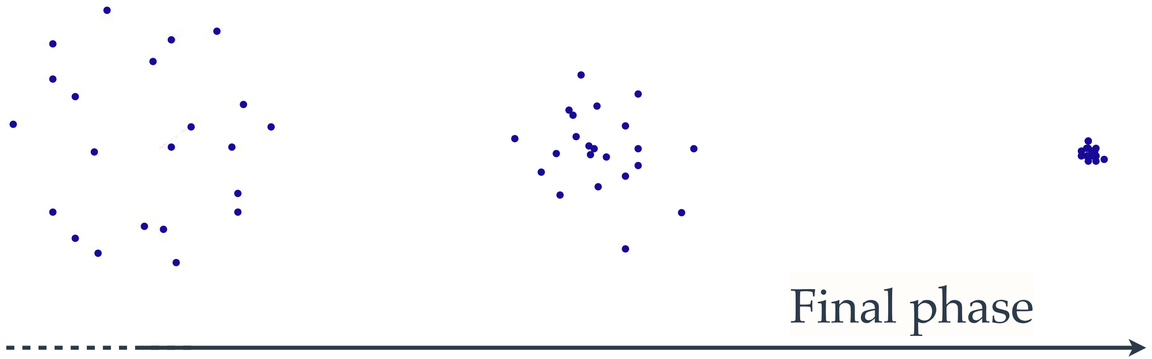
## Clustering



# Dynamic meta-stability

## Clustering



# Dynamic meta-stability

## Our analysis

In the following regime:

$$N \gg 1,$$
$$x_i(0) \sim U\left(\mathbb{S}^{d-1}\right),$$
$$Q, K, V = Id,$$

we study the nonlinear PDE:

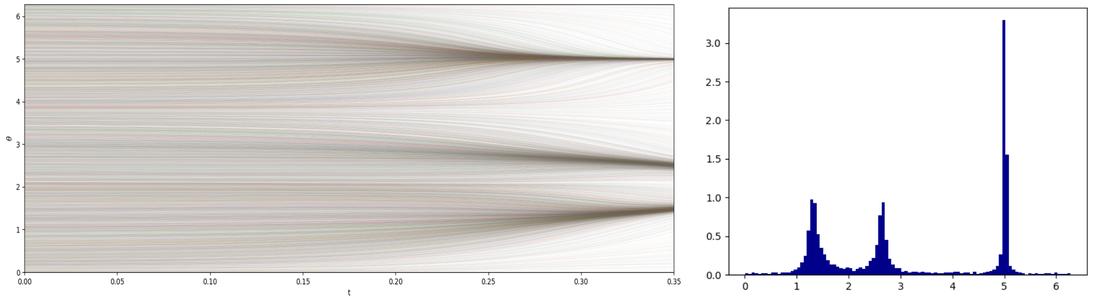
$$\begin{cases} \partial_t \mu + \operatorname{div}(\chi[\mu]\mu) = 0 \\ \mu_{|t=0} = \mu(0) \end{cases}$$

with $\chi[\mu] = P_x\left(\int_{\mathbb{S}^{d-1}} e^{\beta\langle x,y\rangle} y \; d\mu(y)\right).$

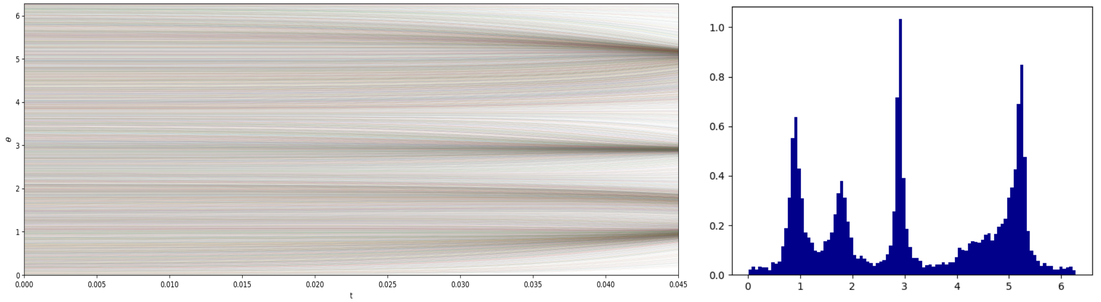
In particular, we prove:

$$T \approx \ln(N),$$
$$N_{\text{cluster}} \approx \beta^{-1/2}.$$

$\beta = 5$



$\beta = 7$



$\beta = 50$