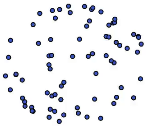
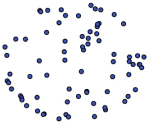
Mode collapse and metastability in Transformers





Berlin-Leipzig hybrid Seminar: Mathematics of Machine Learning

Viktor Stein, 15.09.2025

Outline

I. What are transformers?

II. ODE and PDE description of transformers

III. Long-time behavior - emergence of clusters

LLMs, GPTs, etc

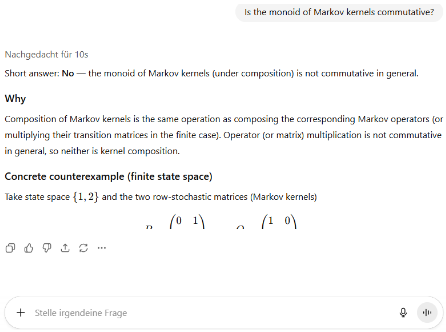


Fig. 1: ChatGPT5' UI.

LLMs, GPTs, etc

- GPT = generative pretrained transformer, a type of LLM = large language model.

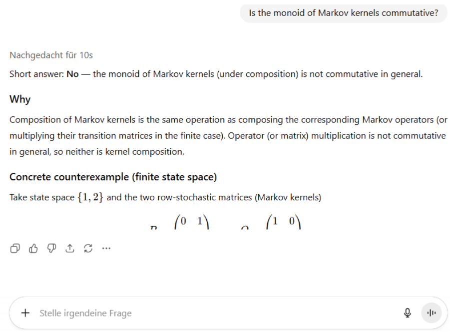


Fig. 1: ChatGPT5' UI.

LLMs, GPTs, etc

- GPT = generative pretrained transformer, a type of LLM = large language model.
- ChatGPT receives “question” (text input sequence) and *generates* “answer” (text output sequence) left-to-right.

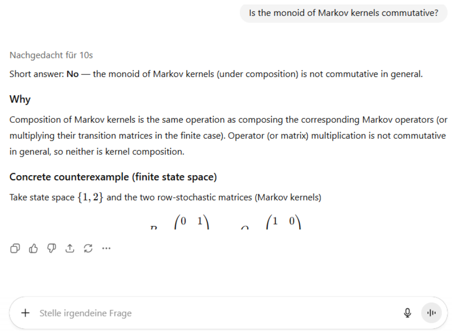


Fig. 1: ChatGPT5’ UI.

LLMs, GPTs, etc

- GPT = generative pretrained transformer, a type of LLM = large language model.
- ChatGPT receives "question" (text input sequence) and *generates* "answer" (text output sequence) left-to-right.
- Before transformers: sequence-to-sequence (Seq2Seq) models use two particular RNNs (called Long-Short-Term-Memory, LSTM) in an *encoder-decoder architecture*. CNNs alike struggle to capture long-range dependencies.

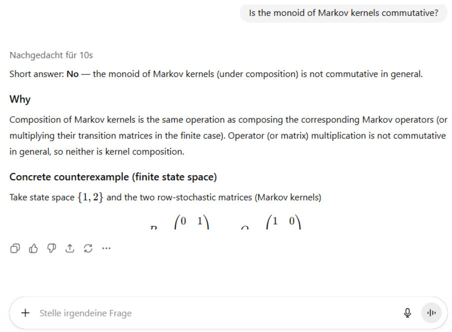


Fig. 1: ChatGPT5' UI.

LLMs, GPTs, etc

- GPT = generative pretrained transformer, a type of LLM = large language model.
- ChatGPT receives "question" (text input sequence) and *generates* "answer" (text output sequence) left-to-right.
- Before transformers: sequence-to-sequence (Seq2Seq) models use two particular RNNs (called Long-Short-Term-Memory, LSTM) in an *encoder-decoder architecture*. CNNs alike struggle to capture long-range dependencies.
- text is not only sequential (order matters), but also structured: there is *context*!

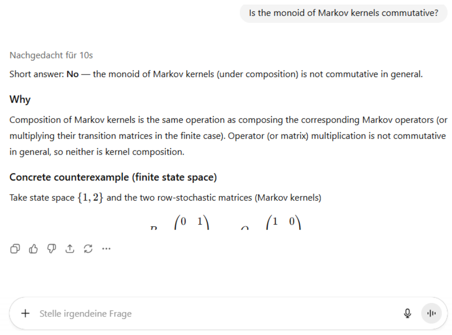


Fig. 1: ChatGPT5' UI.

The preprocessing step of *tokenization* uses a *vocabulary*, an *embedding* and positional encoding.

The preprocessing step of *tokenization* uses a *vocabulary*, an *embedding* and positional encoding.

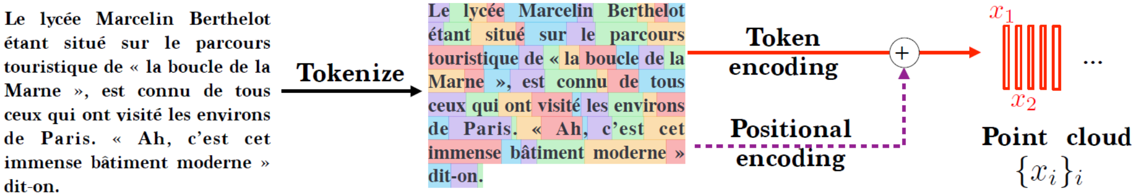


FIG. 2: Text is encoded into a point cloud. © G. Peyré

Tokenization

The preprocessing step of *tokenization* uses a *vocabulary*, an *embedding* and positional encoding.

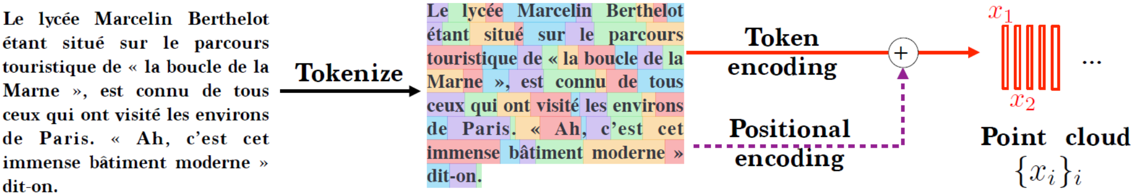


Fig. 2: Text is encoded into a point cloud. © G. Peyré

The points x_i are called (context) *tokens*.

Transformer architecture

The transformer architecture consists of stacked decoder-like *sublayers*,

Transformer architecture

The transformer architecture consists of stacked decoder-like *sublayers*, made up of (masked multi-head) *self-attention* + (token-wise) feed-forward neural networks aka MLP with residual aka skip connections + layer normalization

The transformer architecture consists of stacked decoder-like *sublayers*, made up of (masked multi-head) *self-attention* + (token-wise) feed-forward neural networks aka MLP with residual aka skip connections + layer normalization

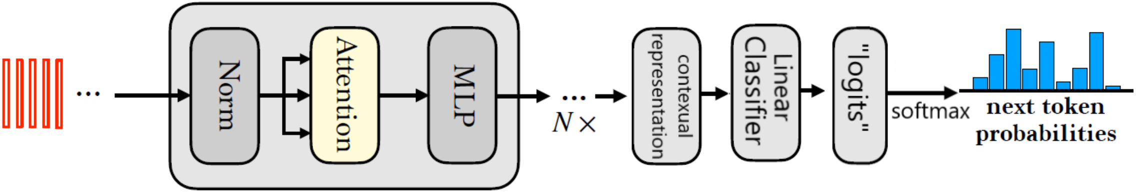


Fig. 3: Autoregressive decoder-only transformer architecture (GPT). Figure modified from Peyré 2024.

The transformer architecture consists of stacked decoder-like *sublayers*, made up of (masked multi-head) *self-attention* + (token-wise) feed-forward neural networks aka MLP with residual aka skip connections + layer normalization

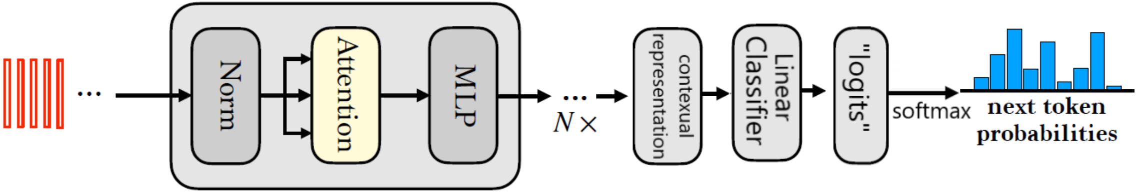


Fig. 3: Autoregressive decoder-only transformer architecture (GPT). Figure modified from Peyré 2024.

(Variant of encoder-decoder transformer "T5" [Vaswani et al. 2017])

Transformer architecture

The transformer architecture consists of stacked decoder-like *sublayers*, made up of (masked multi-head) *self-attention* + (token-wise) feed-forward neural networks aka MLP with residual aka skip connections + layer normalization

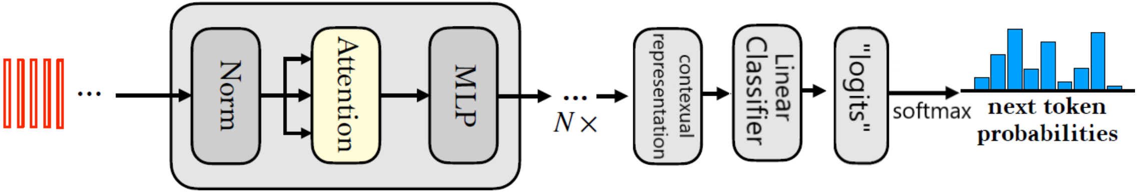


Fig. 3: Autoregressive decoder-only transformer architecture (GPT). Figure modified from Peyré 2024.

(Variant of encoder-decoder transformer "T5" [Vaswani et al. 2017])

Training via backpropagation: loss = predict next token given the previous ones

Transformer architecture

The transformer architecture consists of stacked decoder-like *sublayers*, made up of (masked multi-head) *self-attention* + (token-wise) feed-forward neural networks aka MLP with residual aka skip connections + layer normalization

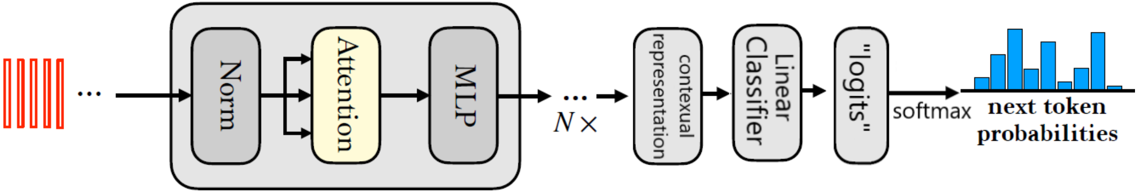


Fig. 3: Autoregressive decoder-only transformer architecture (GPT). Figure modified from Peyré 2024.

(Variant of encoder-decoder transformer "T5" [Vaswani et al. 2017])

Training via backpropagation: loss = predict next token given the previous ones

Generation: predict next token, add to rest ("context"), repeat ("autoregression")

k-th layer with step size $\tau > 0$:

k-th layer with step size $\tau > 0$:

$$x_i^{(k+1)} = x_i^k + \tau \sum_{j=1}^{n} \frac{\exp\left(\langle Qx_i^{(k)}, Kx_j^{(k)}\rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Qx_i^{(k)}, Kx_\ell^{(k)}\rangle\right)} Vx_j^{(k)}, \qquad k \in \{1, \ldots, L\}, i \in \{1, \ldots, n\}.$$

k-th layer with step size $\tau > 0$:

$$x_i^{(k+1)} = x_i^k + \tau \sum_{j=1}^{n} \frac{\exp\left(\langle Q x_i^{(k)}, K x_j^{(k)} \rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Q x_i^{(k)}, K x_\ell^{(k)} \rangle\right)} V x_j^{(k)}, \qquad k \in \{1, \ldots, L\}, i \in \{1, \ldots, n\}.$$

query, key, value matrices Q, K, V *learned during training*

k-th layer with step size $\tau > 0$:

$$x_i^{(k+1)} = x_i^k + \tau \sum_{j=1}^{n} \frac{\exp\left(\langle Q x_i^{(k)}, K x_j^{(k)} \rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Q x_i^{(k)}, K x_\ell^{(k)} \rangle\right)} V x_j^{(k)}, \qquad k \in \{1, \ldots, L\}, i \in \{1, \ldots, n\}.$$

query, key, value matrices Q, K, V *learned during training*

$\langle Qx, Ky \rangle$ is (non-symmetric!) *alignment score* between x and y.

k-th layer with step size $\tau > 0$:

$$x_i^{(k+1)} = x_i^k + \tau \sum_{j=1}^{n} \frac{\exp\left(\langle Q x_i^{(k)}, K x_j^{(k)} \rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Q x_i^{(k)}, K x_\ell^{(k)} \rangle\right)} V x_j^{(k)}, \qquad k \in \{1, \dots, L\}, i \in \{1, \dots, n\}.$$

query, key, value matrices Q, K, V *learned during training*

$\langle Qx, Ky \rangle$ is (non-symmetric!) *alignment score* between x and y. If alignment is high, then x is relevant for y.

k-th layer with step size $\tau > 0$:

$$x_i^{(k+1)} = x_i^k + \tau \sum_{j=1}^{n} \frac{\exp\left(\langle Q x_i^{(k)}, K x_j^{(k)} \rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Q x_i^{(k)}, K x_\ell^{(k)} \rangle\right)} V x_j^{(k)}, \qquad k \in \{1, \ldots, L\}, i \in \{1, \ldots, n\}.$$

query, key, value matrices Q, K, V *learned during training*

$\langle Qx, Ky \rangle$ is (non-symmetric!) *alignment score* between x and y. If alignment is high, then x is relevant for y. Other alignment: $v^\top \tanh(Wx + Uy)$.

The single-head attention block [Vaswani et al. 2017; Bahdanau, Cho, and Bengio 2015]

k-th layer with step size $\tau > 0$:

$$x_i^{(k+1)} = x_i^k + \tau \sum_{j=1}^{n} \frac{\exp\left(\langle Q x_i^{(k)}, K x_j^{(k)} \rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Q x_i^{(k)}, K x_\ell^{(k)} \rangle\right)} V x_j^{(k)}, \qquad k \in \{1, \ldots, L\}, i \in \{1, \ldots, n\}.$$

query, key, value matrices Q, K, V *learned during training*

$\langle Qx, Ky \rangle$ is (non-symmetric!) *alignment score* between x and y. If alignment is high, then x is relevant for y. Other alignment: $v^\top \tanh(Wx + Uy)$.

Compactly: $\operatorname{Attention}(Q, K, V) := \operatorname{softmax}\left(QK^\top\right) V$,

k-th layer with step size $\tau > 0$:

$$x_i^{(k+1)} = x_i^k + \tau \sum_{j=1}^{n} \frac{\exp\left(\langle Q x_i^{(k)}, K x_j^{(k)} \rangle\right)}{\sum_{\ell=1}^n \exp\left(\langle Q x_i^{(k)}, K x_\ell^{(k)} \rangle\right)} V x_j^{(k)}, \qquad k \in \{1, \ldots, L\}, i \in \{1, \ldots, n\}.$$

query, key, value matrices Q, K, V *learned during training*

$\langle Qx, Ky \rangle$ is (non-symmetric!) *alignment score* between x and y. If alignment is high, then x is relevant for y. Other alignment: $v^\top \tanh(Wx + Uy)$.

Compactly: $\text{Attention}(Q, K, V) := \text{softmax}\left(QK^\top\right) V$, where the "soft argmax" is

$$\text{softmax} \colon \mathbb{R}^d \to \text{int}(\Delta_{d-1}), \qquad x \mapsto \left(\frac{\exp(x_j)}{\sum_{\ell=1}^d \exp(x_\ell)} \right).$$

Table of Contents

From now on: ignore normalization & MLP.

Each transformer sublayer = one discrete time step.

From now on: ignore normalization & MLP.

Each transformer sublayer = one discrete time step. Letting the step size $\Delta \to 0$ (like in neural ODEs) we obtain (unmasked single-head) *self-attention*

$$\boxed{\dot{x}_i(t) = \sum_{j=1}^{n} \underbrace{\frac{\exp\left(\langle Q x_i(t), K x_j(t)\rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Q x_i(t), K x_\ell(t)\rangle\right)}}_{=: P_{i,j}(t)} V x_j(t), \qquad i \in [n],\ t > 0.} \tag{1}$$

$x_i(t)$ - tokens or *representations* at time t,

From now on: ignore normalization & MLP.

Each transformer sublayer = one discrete time step. Letting the step size $\Delta \to 0$ (like in neural ODEs) we obtain (unmasked single-head) *self-attention*

$$\boxed{\dot{x}_i(t) = \sum_{j=1}^{n} \underbrace{\frac{\exp\left(\langle Q x_i(t), K x_j(t)\rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Q x_i(t), K x_\ell(t)\rangle\right)}}_{=: P_{i,j}(t)} V x_j(t), \qquad i \in [n],\ t > 0.} \tag{1}$$

$x_i(t)$ - tokens or *representations* at time t, $P_{i,j}$ is called the (stochastic) *attention matrix*

From now on: ignore normalization & MLP.

Each transformer sublayer = one discrete time step. Letting the step size $\Delta \to 0$ (like in neural ODEs) we obtain (unmasked single-head) *self-attention*

$$\boxed{\dot{x}_i(t) = \sum_{j=1}^{n} \underbrace{\frac{\exp\left(\langle Q x_i(t), K x_j(t)\rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Q x_i(t), K x_\ell(t)\rangle\right)}}_{=: P_{i,j}(t)} V x_j(t), \qquad i \in [n],\ t > 0. \qquad (1)}$$

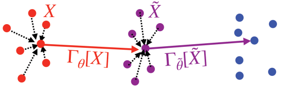
$x_i(t)$ - tokens or *representations* at time t, $P_{i,j}$ is called the (stochastic) *attention matrix*

> (1) is a simplified version of forward pass through the infinitely deep *trained* transformer with the same Q, K, V in all layers ("weight sharing").

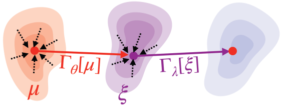
Mean field limit - the transformer PDE

[Burger et al. 2025]

Mean field limit of infinitely many tokens:



$$\{x_i\}_{i=1}^n \longleftrightarrow \frac{1}{n}\sum_{i=1}^n \delta_{x_i} \xrightarrow{n\to\infty} \mu \in \mathcal{P}(\mathbb{R}^d)$$



On probability measures $\mathcal{P}(\mathbb{R}^d)$, the transformer ODE becomes the *transformer PDE*

$$\dot{\mu}_t = -\nabla \cdot \big(\mu_t \Gamma(\mu_t)\big), \quad t > 0, \qquad [\Gamma(\mu)](x) := \int_{\mathbb{R}^d} V y \frac{\exp\left(\langle Qx, Ky\rangle\right)}{\int_{\mathbb{R}^d} \exp\left(\langle Qx, Kz\rangle\right) \mathrm{d}\mu(z)} \,\mathrm{d}\mu(y)$$

Γ is called *softmax attention mapping*.

Other forms of attention: Sinkhorn, L2, linaer, unnormalized, masked)

Table of Contents

Theorem (Geshkovski et al. 2023, Thm. 2.1)

Let $d = 1$, $V > 0$, $QK > 0$.

Asymptotic low-rankness of attention matrix

Theorem (Geshkovski et al. 2023, Thm. 2.1)

Let $d = 1$, $V > 0$, $QK > 0$. *For any sequence of pairwise distinct initial tokens* $(x_i(0))_{i=1}^n \in \mathbb{R}^{d \times n}$,

Theorem (Geshkovski et al. 2023, Thm. 2.1)

Let $d = 1$, $V > 0$, $QK > 0$. For any sequence of pairwise distinct initial tokens $(x_i(0))_{i=1}^n \in \mathbb{R}^{d\times n}$, there exists a permutation matrix $\Pi \in \mathbb{R}^{n\times n}$

Theorem (Geshkovski et al. 2023, Thm. 2.1)

Let $d = 1$, $V > 0$, $QK > 0$. For any sequence of pairwise distinct initial tokens $(x_i(0))_{i=1}^n \in \mathbb{R}^{d \times n}$, there exists a permutation matrix $\Pi \in \mathbb{R}^{n \times n}$ such that

$$\lim_{t \to \infty} P(t) = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \dots & \dots & \vdots \\ 1 & 0 & \dots & \dots & 0 \\ a_1 & a_2 & \dots & \dots & a_d \\ 0 & \dots & \dots & 0 & 1 \\ \vdots & \dots & \dots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} \Pi \in \mathbb{R}^{n \times n}, \qquad a_i \geq 0, \ \sum_{i=1}^n a_i = 1.$$

THEOREM (GESHKOVSKI ET AL. 2023, THM. 2.1)

Let $d = 1$, $V > 0$, $QK > 0$. *For any sequence of pairwise distinct initial tokens* $(x_i(0))_{i=1}^n \in \mathbb{R}^{d \times n}$, *there exists a permutation matrix* $\Pi \in \mathbb{R}^{n \times n}$ *such that*

$$\lim_{t \to \infty} P(t) = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \dots & \dots & \vdots \\ 1 & 0 & \dots & \dots & 0 \\ a_1 & a_2 & \dots & \dots & a_d \\ 0 & \dots & \dots & 0 & 1 \\ \vdots & \dots & \dots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} \Pi \in \mathbb{R}^{n \times n}, \qquad a_i \geq 0, \ \sum_{i=1}^n a_i = 1.$$

For almost all initial tokens $(a_i)_{i=1}^n \in \{e_1, e_n\}$.

Asymptotic low-rankness of attention matrix

Theorem (Geshkovski et al. 2023, Thm. 2.1)

Let $d = 1$, $V > 0$, $QK > 0$. For any sequence of pairwise distinct initial tokens $(x_i(0))_{i=1}^n \in \mathbb{R}^{d \times n}$, there exists a permutation matrix $\Pi \in \mathbb{R}^{n \times n}$ such that

$$\lim_{t \to \infty} P(t) = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \dots & \dots & \vdots \\ 1 & 0 & \dots & \dots & 0 \\ a_1 & a_2 & \dots & \dots & a_d \\ 0 & \dots & \dots & 0 & 1 \\ \vdots & \dots & \dots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} \Pi \in \mathbb{R}^{n \times n}, \qquad a_i \geq 0, \ \sum_{i=1}^n a_i = 1.$$

For almost all initial tokens $(a_i)_{i=1}^n \in \{e_1, e_n\}$. Conjecture: this also holds for $d \geq 2$.

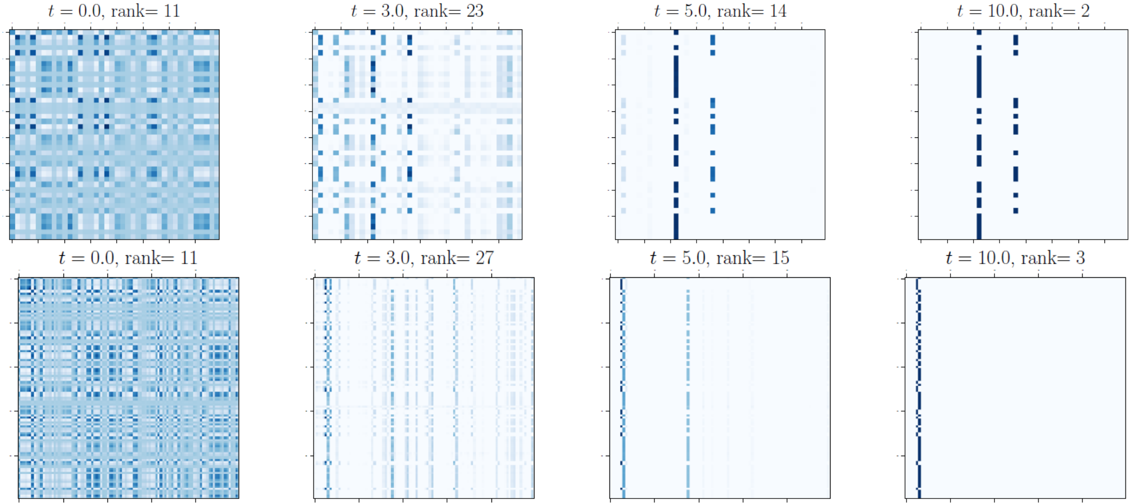


Fig. 4: $d = 1$ and $Q = K = V = 1$. Top: $n = 40$, bottom $n = 100$. The attention matrix converges to a rank two matrix at a doubly exponential rate.

Preprocessing step: spatial rescaling

Motivation. Degenerate case: $Q^\top K = 0$.

Preprocessing step: spatial rescaling

Motivation. Degenerate case: $Q^\top K = 0$. Transformer ODE becomes $\dot{x}_i(t) = V x_i(t)$

Preprocessing step: spatial rescaling

Motivation. Degenerate case: $Q^\top K = 0$. Transformer ODE becomes $\dot{x}_i(t) = V x_i(t) \rightsquigarrow$ no interaction, closed form $x_i(t) = e^{tV} x_i(0)$

Preprocessing step: spatial rescaling

Motivation. Degenerate case: $Q^\top K = 0$. Transformer ODE becomes $\dot{x}_i(t) = V x_i(t) \rightsquigarrow$ no interaction, closed form $x_i(t) = e^{tV} x_i(0) \rightsquigarrow$ divergence: $\|x_i(t)\| \in O(e^t)$ for $t \to \infty$.

Preprocessing step: spatial rescaling

Motivation. Degenerate case: $Q^\mathsf{T} K = 0$. Transformer ODE becomes $\dot{x}_i(t) = V x_i(t) \rightsquigarrow$ no interaction, closed form $x_i(t) = e^{tV} x_i(0) \rightsquigarrow$ divergence: $\|x_i(t)\| \in O(e^t)$ for $t \to \infty$.

Solution: spatial rescaling: $z_i(t) := e^{-tV} x_i(t) \rightsquigarrow$ controls $\|z_i(t)\|$ for $t \to \infty$.

Preprocessing step: spatial rescaling

Motivation. Degenerate case: $Q^\mathsf{T} K = 0$. Transformer ODE becomes $\dot{x}_i(t) = V x_i(t) \rightsquigarrow$ no interaction, closed form $x_i(t) = e^{tV} x_i(0) \rightsquigarrow$ divergence: $\|x_i(t)\| \in O(e^t)$ for $t \to \infty$.

Solution: spatial rescaling: $z_i(t) := e^{-tV} x_i(t) \rightsquigarrow$ controls $\|z_i(t)\|$ for $t \to \infty$.

Transformer ODE becomes

$$\dot{z}_i(t) = \sum_{j=1}^n \left(\frac{\exp\left(\langle Q e^{tV} z_i(t) K e^{tV} z_j(t) \rangle\right)}{\sum_{\ell=1}^n \exp\left(\langle Q e^{tV} z_i(t) K e^{tV} z_\ell(t) \rangle\right)} \right) V \left(z_j(t) - z_i(t)\right).$$

Preprocessing step: spatial rescaling

Motivation. Degenerate case: $Q^{\mathsf{T}}K = 0$. Transformer ODE becomes $\dot{x}_i(t) = Vx_i(t) \rightsquigarrow$ no interaction, closed form $x_i(t) = e^{tV}x_i(0) \rightsquigarrow$ divergence: $\|x_i(t)\| \in O(e^t)$ for $t \to \infty$.

Solution: spatial rescaling: $z_i(t) := e^{-tV}x_i(t) \rightsquigarrow$ controls $\|z_i(t)\|$ for $t \to \infty$.

Transformer ODE becomes

$$\dot{z}_i(t) = \sum_{j=1}^{n} \left(\frac{\exp\left(\langle Qe^{tV}z_i(t)Ke^{tV}z_j(t)\rangle\right)}{\sum_{\ell=1}^{n}\exp\left(\langle Qe^{tV}z_i(t)Ke^{tV}z_\ell(t)\rangle\right)} \right) V\left(z_j(t) - z_i(t)\right).$$

Looks like *Krause model* for flocking phenomena / opinion dynamics:
$\dot{x}_i(t) = \sum_{i=1}^{n} P_{i,j}\left(x_j(t) - x_i(t)\right)$ (note that $P_{i,j}$ *not* time-dependent).

Motivation. Degenerate case: $Q^\mathsf{T} K = 0$. Transformer ODE becomes $\dot{x}_i(t) = V x_i(t) \rightsquigarrow$ no interaction, closed form $x_i(t) = e^{tV} x_i(0) \rightsquigarrow$ divergence: $\|x_i(t)\| \in O(e^t)$ for $t \to \infty$.

Solution: spatial rescaling: $z_i(t) := e^{-tV} x_i(t) \rightsquigarrow$ controls $\|z_i(t)\|$ for $t \to \infty$.
Transformer ODE becomes

$$\dot{z}_i(t) = \sum_{j=1}^{n} \left(\frac{\exp\left(\langle Q e^{tV} z_i(t) K e^{tV} z_j(t)\rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Q e^{tV} z_i(t) K e^{tV} z_\ell(t)\rangle\right)} \right) V \left(z_j(t) - z_i(t)\right).$$

Looks like *Krause model* for flocking phenomena / opinion dynamics:
$\dot{x}_i(t) = \sum_{i=1}^{n} P_{i,j} \left(x_j(t) - x_i(t)\right)$ (note that $P_{i,j}$ *not* time-dependent).

Spatial rescaling is a mathematical surrogate for normalization

Key results from [Geshkovski et al. 2023]

Value	Key and query	Limit geometry
$V = \mathrm{I}_d$	$Q^\mathsf{T} K \succ 0$	vertices of convex polytope

Key results from [Geshkovski et al. 2023]

Value	Key and query	Limit geometry
$V = \mathrm{I}_d$	$Q^\mathsf{T} K \succ 0$	vertices of convex polytope
$\lambda_{\max}(V) > 0$ simple	$\langle Q\varphi_1, K\varphi_1 \rangle > 0$	union of 3 parallel hyperplanes

Key results from [Geshkovski et al. 2023]

Value	Key and query	Limit geometry
$V = \mathrm{I}_d$	$Q^\mathsf{T} K \succ 0$	vertices of convex polytope
$\lambda_{\max}(V) > 0$ simple	$\langle Q\varphi_1, K\varphi_1\rangle > 0$	union of 3 parallel hyperplanes
V paranormal	$Q^\mathsf{T} K \succ 0$	polytope $\times$ subspaces

Key results from [Geshkovski et al. 2023]

Value	Key and query	Limit geometry
$V = \mathrm{I}_d$	$Q^\top K \succ 0$	vertices of convex polytope
$\lambda_{\max}(V) > 0$ simple	$\langle Q\varphi_1, K\varphi_1\rangle > 0$	union of 3 parallel hyperplanes
V paranormal	$Q^\top K \succ 0$	polytope $\times$ subspaces
$V = -\mathrm{I}_d$	$Q^\top K = \mathrm{I}$	single cluster at origin

Value	Key and query	Limit geometry
$V = \mathrm{I}_d$	$Q^\top K \succ 0$	vertices of convex polytope
$\lambda_{\max}(V) > 0$ simple	$\langle Q\varphi_1, K\varphi_1\rangle > 0$	union of 3 parallel hyperplanes
V paranormal	$Q^\top K \succ 0$	polytope $\times$ subspaces
$V = -\mathrm{I}_d$	$Q^\top K = \mathrm{I}$	single cluster at origin

TABLE 1: Clustering taxonomy for rescaled dynamics (except last row).

Interesting: last row $\leftrightarrow$ heat equation, for Sinkhorn attention [Agarwal et al. 2024].

V *paranormal* $\iff \exists F, G \subset \mathbb{R}^d$ with $F \oplus G = \mathbb{R}^d$, $VF = F$, $VG = G$, $V|_F = \lambda \mathrm{I}$, $\rho(V|_G) < \lambda$ ($\rho =$ spectral radius).

Key results from [Geshkovski et al. 2023]

Value	Key and query	Limit geometry
$V = \mathrm{I}_d$	$Q^\top K \succ 0$	vertices of convex polytope
$\lambda_{\max}(V) > 0$ simple	$\langle Q\varphi_1, K\varphi_1\rangle > 0$	union of 3 parallel hyperplanes
V paranormal	$Q^\top K \succ 0$	polytope $\times$ subspaces
$V = -\mathrm{I}_d$	$Q^\top K = \mathrm{I}$	single cluster at origin

Table 1: Clustering taxonomy for rescaled dynamics (except last row).

Interesting: last row $\leftrightarrow$ heat equation, for Sinkhorn attention [Agarwal et al. 2024].

V *paranormal* $\iff \exists F, G \subset \mathbb{R}^d$ with $F \oplus G = \mathbb{R}^d$, $VF = F$, $VG = G$, $V|_F = \lambda \mathrm{I}$, $\rho(V|_G) < \lambda$ ($\rho =$ spectral radius). Also, $\varphi_1 \in \ker(V - \lambda_{\max}(V)\,\mathrm{I})$.

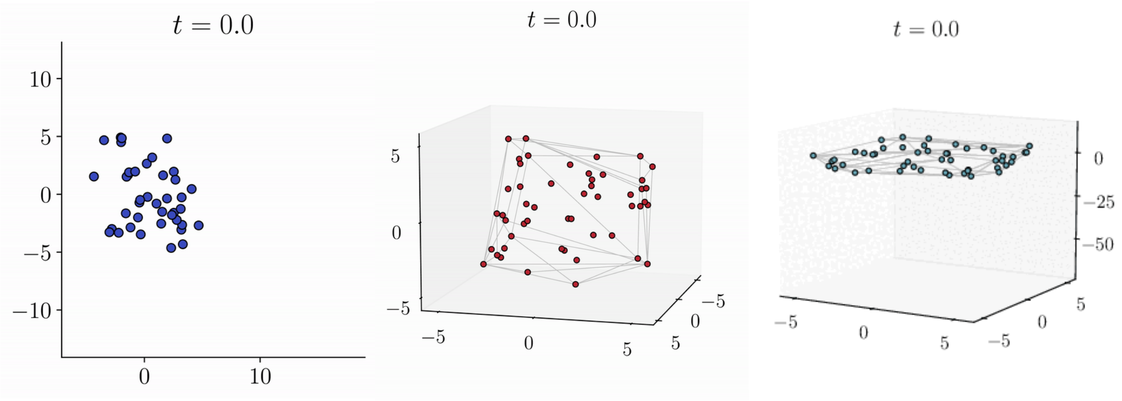


FIG. 5: Clustering means that leaders (="leading" tokens) emerge, that capture attention of all tokens (except one) & carry the largest amount of information ("context awareness").

Empirical results beyond the stated assumptions

- empirically also for non-PSD $Q^\top K$, clustering occurs as outlined above (depending on structure of V).

Empirical results beyond the stated assumptions

- empirically also for non-PSD $Q^\top K$, clustering occurs as outlined above (depending on structure of V).
- empirically, adding a 2-layer MLP ($\sigma \in \{\tanh, \mathrm{ReLU}\}$, $W \in \mathbb{R}^{d\times d}$) $\rightsquigarrow$ same clustering

$$\dot{z}_i(t) = W\sigma\left(\sum_{j=1}^{n} \left(\frac{\exp\left(\langle Qe^{tV} z_i(t) K e^{tV} z_j(t)\rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Qe^{tV} z_i(t) K e^{tV} z_\ell(t)\rangle\right)} \right) V\left(z_j(t) - z_i(t)\right) \right).$$

- empirically also for non-PSD $Q^\mathsf{T} K$, clustering occurs as outlined above (depending on structure of V).
- empirically, adding a 2-layer MLP ($\sigma \in \{\tanh, \mathrm{ReLU}\}$, $W \in \mathbb{R}^{d \times d}$) $\rightsquigarrow$ same clustering

$$\dot{z}_i(t) = W\sigma\left(\sum_{j=1}^{n} \left(\frac{\exp\left(\langle Q e^{tV} z_i(t) K e^{tV} z_j(t) \rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Q e^{tV} z_i(t) K e^{tV} z_\ell(t) \rangle\right)} \right) V\left(z_j(t) - z_i(t)\right) \right).$$

- *Conjecture:* convergence to one of three parallel subspaces of $\mathbb{R}^d$ of codimension k, where k is the number of eigenvalues with positive real part.

Plots: reincorporating the MLP

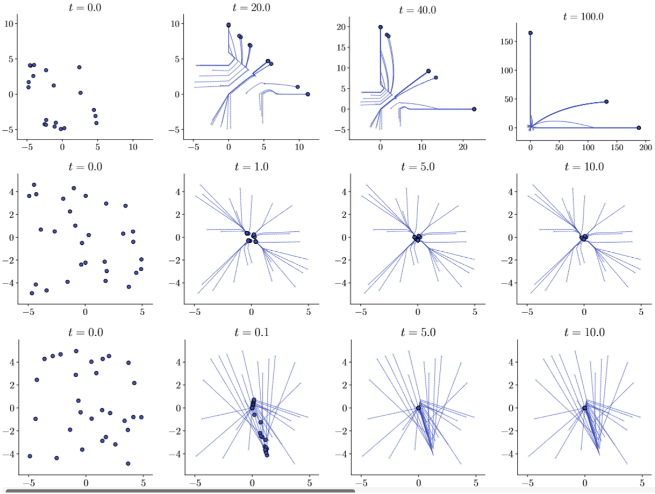


Fig. 6: Top: $\sigma = \mathrm{ReLU}$, $W = \mathrm{I}$, middle: $\sigma = \tanh$, $W = \mathrm{I}$, bottom: $\sigma = \mathrm{ReLU}$, W random.

Let $A \coloneqq K^\mathsf{T} Q$.

Clustering for Gaussian initial data

[Castin et al. 2025]

Let $A \coloneqq K^\mathsf{T} Q$. For $\mu_0 \sim \mathcal{N}(\alpha_0, \Sigma_0)$ we have $\mu_t \sim \mathcal{N}(\alpha_t, \Sigma_t)$

Let $A \coloneqq K^{\mathsf{T}} Q$. For $\mu_0 \sim \mathcal{N}(\alpha_0, \Sigma_0)$ we have $\mu_t \sim \mathcal{N}(\alpha_t, \Sigma_t)$ with

$$\dot{\Sigma}_t = 2\,\mathrm{Sym}(V \Sigma A \Sigma), \qquad \dot{\alpha} = V(\mathrm{I} + \Sigma A)\alpha_t$$

Let $A \coloneqq K^\top Q$. For $\mu_0 \sim \mathcal{N}(\alpha_0, \Sigma_0)$ we have $\mu_t \sim \mathcal{N}(\alpha_t, \Sigma_t)$ with

$$\dot{\Sigma}_t = 2\,\mathrm{Sym}(V\Sigma A\Sigma), \qquad \dot{\alpha} = V(\mathrm{I}+\Sigma A)\alpha_t$$

For Q, K, V constant in time, the covariance equation has the following properties:

- Limiting points have low rank (under commutativity assumptions)

Let $A \coloneqq K^\top Q$. For $\mu_0 \sim \mathcal{N}(\alpha_0, \Sigma_0)$ we have $\mu_t \sim \mathcal{N}(\alpha_t, \Sigma_t)$ with

$$\dot{\Sigma}_t = 2\,\mathrm{Sym}(V\Sigma A\Sigma), \qquad \dot{\alpha} = V(\mathrm{I} + \Sigma A)\alpha_t$$

For Q, K, V constant in time, the covariance equation has the following properties:

- Limiting points have low rank (under commutativity assumptions)
- Rank 1 is preserved
- Stationary points have rank 1 if $V = \mathrm{I}$ and $A = A^\top$.

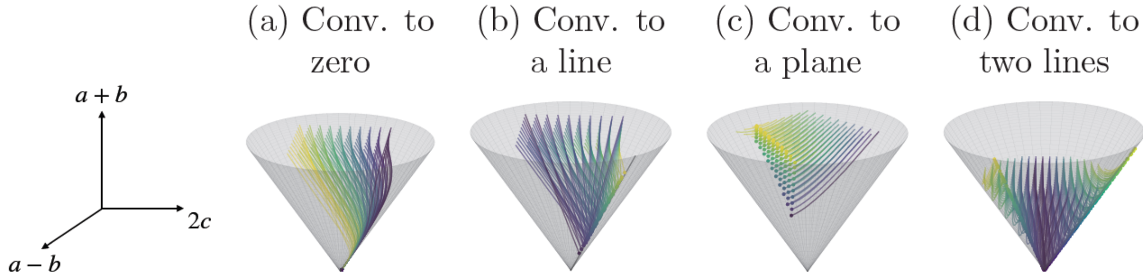


Fig. 7: (a) V random, $A + A^\top \prec 0$, (b) $V = \mathrm{I}$, $A + A^\top \prec 0$ of rank 1, (c) multi-head, $V = \mathrm{I}_2$, $A + A^\top \preceq 0$ of rank 1 (d) A, V chosen specifically to obtain this pattern.

Why do these results look so different?

But papers consider infinitely deep transformers and study $\lim_{t\to\infty} x_i(t)$.

Why do these results look so different?

But papers consider infinitely deep transformers and study $\lim_{t\to\infty} x_i(t)$.

- No spatial rescaling in [Castin et al. 2025], but also treats transformers without weight sharing.

Why do these results look so different?

But papers consider infinitely deep transformers and study $\lim_{t\to\infty} x_i(t)$.

- No spatial rescaling in [Castin et al. 2025], but also treats transformers without weight sharing.
- does finite particle clustering “survive” in the mean field limit?

Thank you for your *attention*!

References I

Agarwal, Medha et al. [2024]. "Iterated Schrödinger bridge approximation to Wasserstein Gradient Flows". arXiv preprint arXiv:2406.10823.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio [May 2015]. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *International Conference on Learning Representations (ICLR)*. Published as a conference paper. San Diego, CA, USA. URL: `https://arxiv.org/abs/1409.0473`.

Burger, Martin et al. [2025]. "Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization". In: *Philosophical Transactions A* 383.2298, p. 20240233.

Castin, Valérie et al. [2025]. "A unified perspective on the dynamics of deep transformers". arXiv preprint arXiv:2501.18322.

References II

Geshkovski, Borjan et al. [2023]. "The Emergence of Clusters in Self-Attention Dynamics". In: *Advances in Neural Information Processing Systems 36*. Ed. by H. Larochelle et al., pp. 57026–57037. DOI: 10.5555/3666122.3668615. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/b2b3e1d9840eba17ad9bbf073e009afe-Abstract-Conference.html.

Lu, Yiping et al. [2019]. "Understanding and improving transformer from a multi-particle dynamic system point of view". In: *Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.*

Peyré, Gabriel [2024]. *Transformers are universal in context leaners*. Slides of a talk given at the conference "Learning and Optimization in Luminy".

Vaswani, Ashish et al. [2017]. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vuckovic, James, Aristide Baratin, and Remi Tachet des Combes [2020]. "A mathematical theory of attention". arXiv preprint arXiv:2007.02876.