# 連続無限層 Transformer の平均場ダイナミクスについて

理化学研究所　磯部 伸
Noboru Isobe
RIKEN
noboru.isobe@riken.jp

## 1　導入

機械学習の中心的課題の一つは，データから予測や推論に有用な表現を自動的に構成することである。例えば，主成分分析に代表される次元削減法や，$k$ 平均法といったクラスタリングアルゴリズムは，高次元データを低次元構造やクラスタ構造として要約する古典的手法である。Bengio et al. (2013) は，このような表現の学習こそが，高性能な機械学習モデルの構築に欠かせない要素であると論じた。

深層学習では，この表現自体をデータから学習する。典型的には，入力データに対してパラメータを含む非線形関数を繰り返し作用させることで，各層において中間表現が生成される。その後，深層ニューラルネットワークにより複雑な表現を獲得させようと，残差接続 (He et al. 2016)，バッチ正規化 (Ioffe and Szegedy 2015)，Attention 機構 (Vaswani et al. 2017) など，さまざまな機構が導入された。特に Attention 機構を内蔵した Transformer は，大規模言語モデルの標準的な基盤アーキテクチャとなり，Generative Pre-trained Transformer (GPT) (Radford et al. 2018) を通じて広く知られるようになった。

一方で，これらの高性能な機械学習モデルが，内部でどのような表現を獲得しているのかについては，いまだ未解明の点が多く残っている。実験的には，高次元空間に埋め込まれた言語表現が，層に沿って特徴的な幾何構造や内在次元を示すことが報告されている (Cheng et al. 2025)。また，中間層の潜在ベクトルを語彙空間に射影して観察する Logit Lens (nostalgebraist 2020) によって，多言語モデルの中間層が英語に近い表現を経由する場合があることも示唆されている (Wendler et al. 2024)。しかし，これらの解析の多くは，訓練済みモデルの各層を経験的に観察するものであり，層を経る過程で獲得されるデータ表現を記述する数理的枠組みは別途必要になる。

ここでは，その枠組みの一つである平均場 Transformer (Rigollet 2026) の理論的枠組みについて紹介する。この数理モデルは，Neural ODE (R. T. Q. Chen et al. 2018) に動機づけられた，深層ニューラルネットワークの連続時間力学系によるモデリングの一種である。このモデリングのアイデアを Transformer に適用することで，モデル内で動的に表れるデータ表現が，Attention 機構が誘導する力学系の性質として理解できることが期待される。

本稿では，このアイデアをシンプルに理解することを目的に，関連する既存研究のうち，

Wasserstein 勾配流として解釈できる設定を取り上げる。この設定は，実際の Transformer を大きく単純化したものではある。しかし，Attention 機構による表現ベクトルの凝集現象を理論的に解析できる設定であり，より一般の Transformer ダイナミクスを解析するための基礎的なモデルとして位置づけられる。

## 2　Transformer の連続時間力学系としての定式化

本節では，(Geshkovski et al. 2025) に基づき，Transformer の層方向発展を $(d-1)$ 次元球面 $\mathbb{S}^{d-1}$ 上の相互作用粒子系として定式化する。

### 2.1　離散時間粒子モデルとしての Transformer

まず，Attention 機構を導入する。Attention 機構 $\mathrm{Attn}\colon (\mathbb{R}^d)^N \to (\mathbb{R}^d)^N$ は，ベクトルの列で表されるデータ $X = (x_1, \dots, x_N) \in (\mathbb{R}^d)^N$ に対して，

$$\mathrm{Attn}(X)_i \coloneqq \sum_{j=1}^{N} \frac{\mathrm{e}^{\beta\langle Qx_i, Kx_j\rangle}}{\sum_{k=1}^{N} \mathrm{e}^{\beta\langle Qx_i, Kx_k\rangle}} V x_j, \quad i = 1, \dots, N,$$

で定義される。ここで，$Q, K, V \in \mathbb{R}^{d\times d}$ は（本来は学習可能な）重み行列，$\beta > 0$ は逆温度パラメータである。Transformer においては，系列 $X$ が Attn を含む人工ニューラルネットワークによって再帰的に処理される。ここでは，残差接続を持つ次のネットワークを考える。

$$x_i^{l+1} = \mathcal{N}\Big(x_i^l + \tau\,\mathrm{Attn}(X^l)_i\Big), \quad i = 1, \dots, N, \quad X^l = \Big\{x_1^l, \dots, x_N^l\Big\}, \quad l = 1, 2, \dots. \tag{2.1}$$

ここで $\mathcal{N}\colon \mathbb{R}^d \ni x \mapsto x/|x| \in \mathbb{S}^{d-1}$ は，RMS 層正規化 (Zhang and Sennrich 2019) を単純化したノルム正規化であり，$\tau > 0$ は便宜上導入した一層あたりの時間刻みである。

### 2.2　連続時間力学系モデルの導出

式 (2.1) において $\tau \to 0$ の形式極限を取る。$\mathcal{N}$ の $x \in \mathbb{S}^{d-1}$ における微分は，接空間 $T_x\mathbb{S}^{d-1}$ への直交射影 $P_x \coloneqq I_d - xx^\top \in \mathbb{R}^{d\times d}$ であるから，連続時間極限として

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}(t) = P_{x_i(t)}\,\mathrm{Attn}(X(t))_i, \quad x_i(t) \in \mathbb{S}^{d-1}, \quad i = 1, \dots, N,$$

を得る。明らかに $\frac{\mathrm{d}}{\mathrm{d}t}\|x_i(t)\|^2 = 0$ であるから，初期条件 $X(0) = [x_1(0), \dots, x_N(0)]$ が $\mathbb{S}^{d-1}$ 上にあれば解は球面上に留まる。

### 2.3　平均場 Transformer モデル

ここで，今の定式化では $\mathrm{Attn}(X) \in (\mathbb{R}^d)^N$ は，添え字 $i$ に対する置換に対して同変であることに注意する。そこで，$\mathcal{P}(\mathbb{S}^{d-1})$ を球面上の Borel 確率測度の集合として，$X(t) = (x_i(t))_{i=1}^N$ がなす経験分布

$$\mu_t \coloneqq \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i(t)} \in \mathcal{P}\Big(\mathbb{S}^{d-1}\Big)$$

を導入する。すると，滑らかなテスト関数 $\varphi \in C^\infty(\mathbb{S}^{d-1})$ に対して

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} \varphi \,\mathrm{d}\mu_t = \int_{\mathbb{S}^{d-1}} \langle \nabla_{\mathbb{S}^{d-1}}\varphi(x), \chi_{\beta,Q,K,V}[\mu_t](x)\rangle \,\mathrm{d}\mu_t(x)$$

が成り立つ (cf. (Ambrosio et al. 2008))。ここで，確率測度 $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ に対して $\chi_{\beta,Q,K,V}[\mu]$ は Attention の測度論的表現

$$\chi_{\beta,Q,K,V}[\mu](x) \coloneqq P_x \frac{\displaystyle\int_{\mathbb{S}^{d-1}} \mathrm{e}^{\beta\langle Qx,Ky\rangle} Vy \,\mathrm{d}\mu(y)}{\displaystyle\int_{\mathbb{S}^{d-1}} \mathrm{e}^{\beta\langle Qx,Ky\rangle} \,\mathrm{d}\mu(y)} \tag{2.2}$$

である。実際，$\mu^X = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$, $X \in (\mathbb{S}^{d-1})^N$ に対し，$P_{x_i}\,\mathrm{Attn}(X)_i = \chi_{\beta,Q,K,V}[\mu^X](x_i)$ である。以上の議論に基づき，我々は Transformer を，確率測度 $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ に関する方程式

$$\partial_t \mu_t + \mathrm{div}_{\mathbb{S}^{d-1}}(\mu_t \chi_{\beta,Q,K,V}[\mu_t]) = 0 \tag{2.3}$$

が誘導する $\mathcal{P}(\mathbb{S}^{d-1})$ 上の力学系とみなす。$\mu$ のことを平均場と呼び，(2.3) を平均場 Transformer と呼ぶ。この方程式の well-posedness や $N \to \infty$ の極限などの解析学的な基礎については (Castin et al. 2025) に体系的な扱いがある。

# 3　勾配流構造を用いた平均場 Transformer の動的解析

ここからは平均場 Transformer を対象にする。簡単のため，本節では (Geshkovski et al. 2025) にならい

$$Q = K = V = I_d, \qquad \beta > 0 \tag{3.1}$$

と仮定し，さらに，式 (2.2) の分母 $Z_\beta(x) \coloneqq \int \mathrm{e}^{\beta\langle x,y\rangle}\,\mathrm{d}\mu(y)$ を取り除いた

$$\chi_\beta[\mu](x) \coloneqq P_x \int_{\mathbb{S}^{d-1}} \mathrm{e}^{\beta\langle x,y\rangle} y \,\mathrm{d}\mu(y) \tag{3.2}$$

をベクトル場として与えた方程式

$$\partial_t \mu_t + \mathrm{div}_{\mathbb{S}^{d-1}}(\mu_t \chi_\beta[\mu_t]) = 0 \tag{3.3}$$

を考える。

## 3.1　勾配流構造

方程式 (3.3) は $\mu$ に関する非線形方程式ではあるが，以下に示す変分的な構造を有する。$\mathcal{P}(\mathbb{S}^{d-1})$ 上の汎関数 $E_\beta$ を

$$E_\beta(\mu) \coloneqq \frac{1}{2\beta}\iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \mathrm{e}^{\beta\langle x,y\rangle}\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)$$

で定める。

**命題 3.1** (Geshkovski et al. 2025). $\chi_\beta$ は式 (3.2) で与えられるとする。このとき，

$$\chi_\beta[\mu](x) = \nabla_{\mathbb{S}^{d-1}} \frac{\delta E_\beta}{\delta \mu}[\mu](x)$$

である。ここで，$\nabla_{\mathbb{S}^{d-1}}$ は（変数 $x \in \mathbb{S}^{d-1}$ に関する）勾配であり，$\frac{\delta}{\delta \mu}$ は（変数 $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ に関する）汎関数微分である。特に，$\mu_t$ を式 (3.3) の（超関数）解とすると，

$$\frac{\mathrm{d}}{\mathrm{d}t} E_\beta(\mu_t) = \int_{\mathbb{S}^{d-1}} \|\chi_\beta[\mu_t](x)\|^2 \, \mathrm{d}\mu_t(x) \geq 0$$

が成り立つ。

命題 3.1 から，平均場 Transformer (3.3) は $-E_\beta$ を目的関数として最適化しようとする力学系であることがわかる。このように，ある（汎）関数の，ある距離に関する勾配でベクトル場が与えられるような力学系を勾配流と呼ぶ。詳細は (Ambrosio et al. 2008) に譲るが，命題 3.1 は，ベクトル場 $\chi_\beta$ が，目的関数 $-E_\beta$ の，(2-)Wasserstein 距離 $W_2$ に関する勾配になっていることを表す。また，目的関数 $-E_\beta$ の最小化元について，次が知られている：

**命題 3.2** (Geshkovski et al. 2025, Proposition 3.4). $d \geq 3$ とし，$\beta > 0$ とする。$E_\beta$ の任意の局所最大点は大域最大点であり，ある $x_0 \in \mathbb{S}^{d-1}$ を用いて $\mu = \delta_{x_0}$ と書ける。ここに，$\delta_{x_0} \in \mathcal{P}(\mathbb{S}^{d-1})$ は Dirac のデルタ分布である。

命題 3.2 から，もし平均場 Transformer (3.3) が $-E_\beta$ を最適化できた，すなわち，層 $t$ を無限大に送った極限で $E_\beta$ を最大化する分布に収束するとすると，その分布は一点に集中していることを意味する。これは，Transformer が入力された初期分布 $\mu_0$ を，層を経るごとに適当な一点に集中させていく様子を表していると考えられる。この挙動は，表現学習の観点からはクラスタリングと捉えられる。

**注意 3.3** $Z_\beta$ による正規化について. 本稿では説明を省略するが，平均場 Transformer (2.3) も，仮定 (3.1) のもとであれば同じ目的関数の勾配流となることが示される Burger et al. (2025)。ただし，勾配を考える距離は，ある重み付きの Wasserstein 距離となる。

## 3.2 クラスタリングの定量的評価

命題 3.1 と 3.2 は，単一クラスターがエネルギー最大化状態であることを示す静的情報である。しかし，勾配流の解が実際に最大化元へ収束するかどうかは，別途証明すべき数学的問題である。S. Chen et al. (2025) は，この問題に対して定量的な収束評価を与えている。彼らはまず，入力分布 $\mu_0$ がある点 $u \in \mathbb{S}^{d-1}$ 周りに集中している場合を考えている。その集中度合い $\alpha \in [0, \pi/2)$ に対して，球冠集合を

$$S_\alpha^+(u) := \left\{ x \in \mathbb{S}^{d-1} \;\middle|\; \langle x, u \rangle \geq \cos \alpha \right\}$$

で定義する。

**定理 3.4** (S. Chen et al. 2025, Theorem 2.3)**.** $d \geq 2$, $\beta > 0$, $\alpha \in [0, \pi/2)$, $u \in \mathbb{S}^{d-1}$ とする。$\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ が $\operatorname{supp}\mu \subset S_\alpha^+(u)$ を満たすとする。さらに $10(1+\sqrt{\beta})\tan\alpha \leq 1$ を仮定する。このとき

$$E_\beta(\delta_u) - E_\beta(\mu) \leq 10\mathrm{e}^{-\beta} \int_{\mathbb{S}^{d-1}} \|\chi_\beta[\mu](x)\|^2 \,\mathrm{d}\mu(x)$$

が成り立つ。また，初期値 $\mu_0 = \mu$ から 式 (3.3) の解 $(\mu_t)_{t\geq 0}$ は，ある $x_\infty \in \mathbb{S}^{d-1}$ に対して $\langle x_\infty, u\rangle \geq \cos\alpha$ を満たし，さらに

$$W_2(\mu_t, \delta_{x_\infty}) \leq 20\mathrm{e}^{-\beta}\mathrm{e}^{-\frac{\mathrm{e}^\beta}{20}t}\left(\int_{\mathbb{S}^{d-1}} \|\chi_\beta[\mu](x)\|^2 \,\mathrm{d}\mu(x)\right)^{1/2}$$

が成り立つ。

また，初期分布が集中していない場合でも，$\beta$ が小さい場合には大域的な指数的収束が得られる。

**定理 3.5** (S. Chen et al. 2025, Theorem 2.5)**.** $d \geq 2$ とする。$(\mu_t)_{t\geq 0}$ を 式 (3.3) の解とする。初期値 $\mu_0$ は $R_0 := \|\int_{\mathbb{S}^{d-1}} x \,\mathrm{d}\mu_0(x)\| > 0$ を満たし，さらに一様分布に対する Radon–Nikodym 微分 $f_0 \in L^2(\mathbb{S}^{d-1})$ を持つと仮定する。このとき，任意の $t > 0$ で $\mu_t$ は密度 $f_t \in L^2(\mathbb{S}^{d-1})$ を持つ。さらに，$\mu_0$ に依存する定数 $\beta_0 > 0, C_0 > 0, T_0 > 0$ が存在し，$0 < \beta < \beta_0$ ならば，ある $x_\infty \in \mathbb{S}^{d-1}$ に対して

$$W_2(\mu_t, \delta_{x_\infty}) \leq C_0\mathrm{e}^{-t/100}, \qquad t > T_0$$

が成り立つ。

**注意 3.6 大きい $\beta$ での制限.** 定理 3.5 の $\beta$ が十分小さいという条件は単なる技術的仮定ではない。(S. Chen et al. 2025, Example 2.7) において，$d = 2$, $\beta = 100$ の円周上で，$R_0 > 0$ かつ $L^2$ 密度を持つ初期値から出発しても，解が一点に台を持つ分布ではなく，二点分布 $\frac{1}{3}\delta_{(1,0)} + \frac{2}{3}\delta_{(-1,0)}$ へ収束する例を構成している。

## 3.3 準安定的クラスタリング

3.2節では，$Q = K = V = I_d$ の平均場 Transformer 式 (3.3) に対して，絶対連続初期値から一点分布へ収束する十分条件を述べた。これは，層を十分深くすると，入力列 $X = (x_1, \ldots, x_N)$ は一つのクラスターに集まることを支持する結果である。一方で，有限だが大きい文脈長 $N$ から作られる経験測度では，最終的な一つのクラスターに至る前に，複数のクラスターが現れる中間時間スケールが存在する。本節では，このような状況を考察した Bruno et al. (2025) の結果を述べる。$d = 2$ とする。すなわち台空間を $\mathbb{S}^1 = \mathbb{R}/2\pi\mathbb{Z}$ と同一視する。初期分布 $\mu_0$ が経験分布 $\mu_0^N = \frac{1}{N}\sum \delta_{x_i}$ であるとすると，(3.3) の解 $\mu_t$ は $\mu_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i(t)}$ という表示を持つ。その各点 $x_i(t) \in \mathbb{S}^1$ を

$$x_i(t) = (\cos\theta_i(t), \sin\theta_i(t))$$

と書くと，各角度 $\theta_i(t)$ は，蔵本モデル (Kuramoto 1975) の亜種

$$\dot{\theta}_i(t) = -\frac{1}{N}\sum_{j=1}^{N} \mathrm{e}^{\beta\cos(\theta_i(t)-\theta_j(t))}\sin(\theta_i(t)-\theta_j(t)), \qquad i = 1,\dots,N,$$

に従う。$\mathbb{S}^1$ 上の偶関数

$$H_\beta(\theta) \coloneqq \frac{1}{\beta}\mathrm{e}^{\beta\cos\theta}, \qquad \theta \in \mathbb{S}^1$$

を導入すると

$$\partial_\theta(H_\beta * \mu)(\theta) = -\int_{\mathbb{S}^1} \mathrm{e}^{\beta\cos(\theta-\phi)}\sin(\theta-\phi)\,\mathrm{d}\mu(\phi) = \chi_\beta[\mu](\theta)$$

である。ここで，$H_\beta * \mu$ は畳み込みである。したがって，式 (3.3) は

$$\partial_t\mu_t + \partial_\theta(\mu_t\ \partial_\theta(H_\beta * \mu_t)) = 0$$

と書ける。$\omega$ を $\mathbb{S}^1$ 上の一様確率測度とする。$H_\beta$ の Fourier cosine 係数を $(\widehat{H}_{\beta,k})_{k=1}^{\infty}$ と書き，

$$\gamma_k \coloneqq k^2\widehat{H}_{\beta,k}, \qquad \gamma_{\max} \coloneqq \max_{k\geq 0}\gamma_k$$

と定める。

**定理 3.7** (Bruno et al. 2025, Theorems 4.2–4.3)**.** $\gamma_{\max} > 0$ がただ一つの $k_{\max} > 0$ で達成されると仮定する。また，$\rho_t^N \coloneqq \mu_t^N - \omega$ と置き，$\rho_0^N$ の $k$ 番目の Fourier 係数を $(\widehat{\rho}_0^N)_k$ と書く。

$$T_1 \coloneqq \frac{1}{\gamma_{\max}}\log\left(\frac{N^{-1/4}}{\|\rho_0^N\|_{H^{-1}}}\right)$$

とすると，$N\to\infty$ で確率収束の意味で

$$\mu_{T_1}^N = \omega + N^{-1/4}\frac{(\widehat{\rho}_0^N)_{k_{\max}}}{\|\rho_0^N\|_{H^{-1}}}\cos(k_{\max}\theta) + R_N, \qquad \|R_N\|_{H^{-1}} = o\left(N^{-1/4}\right)$$

が成り立つ。さらに，$\alpha_N \coloneqq N^{-1/4}\dfrac{\left|(\widehat{\rho}_0^N)_{k_{\max}}\right|}{\|\rho_0^N\|_{H^{-1}}}$ と置き，$f^\alpha$ を

$$\partial_t f^\alpha = -\partial_\theta(f^\alpha\ \partial_\theta(H_\beta * f^\alpha)), \qquad f_0^\alpha = \omega + \alpha\cos(k_{\max}\theta) \tag{3.4}$$

の解とする。$\delta > 0$ を十分小さく固定し，

$$T_2 \coloneqq \frac{1}{\gamma_{\max}}\log\left(\frac{\delta}{\alpha_N}\right)$$

と定める。このとき

$$W_1\left(\mu_{T_1+T_2}^N, f_{T_2}^{\alpha_N}\right) \longrightarrow 0$$

が確率収束の意味で成り立ち，さらに

$$W_1\left(\mu_{T_1+T_2}^N, \omega\right) > \delta$$

が確率 $1-o(1)$ で成り立つ。

**仮定 3.8** (Bruno et al. 2025, Assumption 3). 定理 3.7 と同じ仮定の下で，$f_* := \lim_{\alpha\to 0} f^\alpha(T_2)$ とする。時刻 $T_2$ で初期値 $f_*$ を課した 式 (3.4) の解を $f_*(t)$ と書く。また

$$\mu_{\mathrm{cluster}} := \frac{1}{k_{\max}} \sum_{j=0}^{k_{\max}-1} \delta_{2\pi j/k_{\max}}$$

と置く。このとき $W_1(f_*(t), \mu_{\mathrm{cluster}}) \longrightarrow 0, t \to \infty$ を仮定する。

**定理 3.9** (Bruno et al. 2025, Theorem 4.5). 仮定 3.8 の下で，任意の $\eta > 0$ に対して十分大きい $T_3 > 0$ を取れば

$$\mathbb{P}\Big(W_1\Big(\mu^N_{T_1+T_2+T_3}, \mu_{\mathrm{cluster}}\Big) > \eta\Big) \longrightarrow 0 \qquad (N \to \infty)$$

が成り立つ。

定理 3.7 と 3.9 は，一様分布からサンプルされた入力 $X$ 中に含まれる $N^{-1/2}$ サイズのノイズが $O(\log N)$ の時間スケールで増幅され，$k_{\max}$ 個のクラスターが選択される構造を作ることを示す。このように，安定な定常解に収束するまでに長い時間留まる状態があることを metastability（準安定性）という。

# 4　発展：学習による clustering 挙動の変化

ここまでの議論では，$Q, K, V$ のパラメータを単位行列に固定し，その下でトークン分布が層方向にどのように発展するかを考えてきた。一方，実際の Transformer では，Attention 機構だけでなく feed-forward network (FFN) も含まれ，さらにそれらの中にある重みパラメータは訓練によって選ばれる。我々は，このような訓練によって上で説明してきたクラスタリング現象がどのように変化するかを，同じく勾配流の枠組みで解析した (Isobe et al. 2026)。

今後の課題としては，この勾配流の枠組みを越えて，実際の Transformer で見られるような回転的な挙動 (Blayney et al. 2026) を理論解析する力学系理論のツールを開発することが挙げられる。

# References

L. Ambrosio, N. Gigli, and G. Savaré (2008). *Gradient flows in metric spaces and in the space of probability measures*. 2nd ed. Lectures in Mathematics ETH Zürich.

Y. Bengio, A. Courville, and P. Vincent (2013). “Representation Learning: A Review and New Perspectives”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1798–1828.

H. Blayney, Á. Arroyo, J. Obando-Ceron, P. S. Castro, A. Courville, M. M. Bronstein, and X. Dong (2026). *A Mechanistic Analysis of Looped Reasoning Language Models*. arXiv: 2604.11791 [cs.LG]. URL: https://arxiv.org/abs/2604.11791.

G. Bruno, F. Pasqualotto, and A. Agazzi (2025). "Emergence of meta-stable clustering in mean-field transformer models". *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=eBS3dQQ8GV.

M. Burger, S. Kabri, Y. Korolev, T. Roith, and L. Weigand (2025). "Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 383.2298, p. 20240233. eprint: https://royalsocietypublishing.org/rsta/article-pdf/doi/10.1098/rsta.2024.0233/2835670/rsta.2024.0233.pdf. URL: https://doi.org/10.1098/rsta.2024.0233.

V. Castin, P. Ablin, J. A. Carrillo, and G. Peyré (2025). *A Unified Perspective on the Dynamics of Deep Transformers*. arXiv: 2501.18322 [cs.LG]. URL: https://arxiv.org/abs/2501.18322.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud (2018). "Neural Ordinary Differential Equations". *Advances in Neural Information Processing Systems*. Vol. 31. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet (2025). *Quantitative Clustering in Mean-Field Transformer Models*. arXiv: 2504.14697 [cs.LG]. URL: https://arxiv.org/abs/2504.14697v1.

E. Cheng, D. Doimo, C. Kervadec, I. Macocco, L. Yu, A. Laio, and M. Baroni (2025). "Emergence of a High-Dimensional Abstraction Phase in Language Transformers". *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=0fD3iIBhlV.

B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet (2025). "A mathematical perspective on transformers". *Bull. Am. Math. Soc., New Ser.* 62.3, pp. 427–479.

K. He, X. Zhang, S. Ren, and J. Sun (2016). "Deep Residual Learning for Image Recognition". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

S. Ioffe and C. Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research, pp. 448–456. URL: https://proceedings.mlr.press/v37/ioffe15.html.

N. Isobe, D. Inoue, and M. Imaizumi (2026). *Training-Induced Escape from Token Clustering in a Mean-Field Formulation of Transformers*. arXiv: 2605.07772 [cs.LG]. URL: https://arxiv.org/abs/2605.07772.

Y. Kuramoto (1975). "Self-entrainment of a population of coupled non-linear oscillators". *International Symposium on Mathematical Problems in Theoretical Physics*, pp. 420–422.

nostalgebraist (2020). *Interpreting GPT: The Logit Lens*. URL: https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens (visited on 05/15/2026).

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever (2018). *Improving Language Understanding by Generative Pre-Training*. Technical report. OpenAI. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (visited on 05/15/2026).

P. Rigollet (2026). *The Mean-Field Dynamics of Transformers*. arXiv: 2512.01868 [cs.LG]. URL: https://arxiv.org/abs/2512.01868v4.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). "Attention is All you Need". *Advances in Neural Information Processing Systems*. Vol. 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

C. Wendler, V. Veselovsky, G. Monea, and R. West (2024). "Do Llamas Work in English? On the Latent Language of Multilingual Transformers". *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394. URL: https://aclanthology.org/2024.acl-long.820/.

B. Zhang and R. Sennrich (2019). "Root Mean Square Layer Normalization". *Advances in Neural Information Processing Systems*. Vol. 32. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf.