Consensus is all you get: the role of attention in transformers

Álvaro Rodríguez Abella [*1] **João Pedro Silvestre** [*1] **Paulo Tabuada** [*1]

Abstract

A key component of transformers is the attention mechanism orchestrating how each token influences the propagation of every other token along the layers of a transformer. In this paper we provide a rigorous, mathematical analysis of the asymptotic properties of attention in transformers. Although we present several results based on different assumptions, all of them point to the same conclusion, all tokens asymptotically converge to each other, a phenomenon that has been empirically reported in the literature. Our findings are carefully compared with existing theoretical results and illustrated by simulations and experimental studies using the GPT-2 and the GPT-Neo models.

1. Introduction

The incorporation of attention (Chorowski et al., 2015) in natural language processing was a significant breakthrough, particularly in the context of sequence-to-sequence models, enabling the creation of transformers (Vaswani et al., 2017) which revolutionized the field. Even initial transformer models such as GPT (Radford et al., 2018) or Bert (Devlin et al., 2018) showed drastic improvements over previous approaches such as the Long Short-Term Memory model (Sainath et al., 2015).

As practical applications of deep neural networks, such as image recognition (O'shea & Nash, 2015), natural language processing (Torfi et al., 2020), and autonomous driving (Grigorescu et al., 2020), continue to advance, our understanding of these networks is struggling to keep pace (Van Dijk et al., 2023). This underscores the critical importance of our study, which aims to delve deeper into transformers and their dynamics. Our understanding of transformers is currently limited by their inherent complexity, making it challenging to comprehensively explain their behavior (Peng et al., 2024). However, recent studies have shown the emergence of clusters of tokens empirically and theoretically (Dong et al., 2021; Feng et al., 2022; Geshkovski et al., 2023b; 2024). These findings suggest that without proper care, large transformers may collapse, a phenomenon where the tokens cluster, limiting the model's ability to produce different outputs.

Our work was motivated by the paper (Geshkovski et al., 2023a) where a mathematical model for attention was proposed, based on prior work on similar models (Lu et al., 2019; Dutta et al., 2021), and investigated. The authors share the vision outlined in (Geshkovski et al., 2023a), a better understanding of the role and importance of attention mechanisms can be achieved through the study of mathematical models. Our contribution lies in bringing ideas developed by the control community, where the study of asymptotic properties of dynamical and control systems is a central preoccupation, to bear on this problem. While deferring to the next section a more detailed comparison between our results and those available in the literature, we emphasize here that, in contrast with (Geshkovski et al., 2023a;b; 2024), we do not rely on stochastic and/or mean-field techniques and rather adopt a geometric perspective drawing from control theory, e.g., from consensus dynamics on manifolds (Sarlette & Sepulchre, 2009) such as spheres (Markdahl et al., 2017; Thunberg et al., 2018), and from Input-to-State Stability (Sontag, 1989; 2001; 2008).

Contributions of the paper

The main contribution of this work is to provide a number of results[1], for a differential equation model of attention[2] showing that all tokens converge to a single cluster thereby leading to a collapse of the model. We use the term *consensus equilibria* to refer to such clusters as is done in the consensus literature (Ren et al., 2005; Cao et al., 2013). These results hold under different assumptions on the parameters of the model —namely, the query (Q), key (K) and value matrices (U), as well as the number of heads (h)—

[*] Equal contribution [1] Department of Electrical and Computer Engineering, University of California, Los Angeles, USA. Correspondence to: João Pedro Silvestre <joaosilvestre@g.ucla.edu>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

[1] A detailed proof for each theorem stated in this paper can be found in (Rodríguez Abella et al., 2024).

[2] Since we focus on the attention mechanism, this model does not describe the effect of feed-forward layers.

Table 1. Summary of the results presented in this work for several particular cases of the continuous model (6), where $Q(t)$, $K(t)$, and $U(t)$ denote the query, key and value matrices, respectively.

	Full attention	Causal attention (auto-regressive)	
Section	§3	§4.1	§4.2
# of heads	$h \geq 1$	$h \geq 1$	$h = 1$
$P(t) = Q(t)^\top K(t)$	Time varying, uniformly continuous, bounded	Time varying, bounded	Time varying, bounded
$U(t)$	Identity	Identity	Time invariant, symmetric
Result	Theorem 3.2	Theorem 4.2	Theorem 4.5
Statement	Convergence to consensus	Asympt. stability of consensus	Asympt. stability of consensus
Domain of attraction	Some hemisphere	Conull (complement of zero measure)	Fixed hemisphere

that are summarized in Table 1.

In particular, Theorem 3.2 states that tokens converge to a consensus equilibrium whenever their starting positions lie in the interior of some hemisphere of the ellipsoid. This result holds for any number of heads and time varying matrix $P = Q^\top K$ provided that U is the identity and P is bounded and uniformly continuous as a function of the time. A similar result is reported in (Geshkovski et al., 2023a) under Lemma 4.2. However, its conclusions hold under the stronger assumptions that both U and $P = Q^\top K$ are the identity matrix and there is a single attention head. Theorem 3.2 makes no assumptions on the attention matrix other those induced by the assumptions on P. In contrast, Theorems 4.2 and 4.5 focus on the auto-regressive case, also known as causal attention, where the self-attention matrix is lower triangular. Theorem 4.2 states that when U is the identity, the first token is fixed and all the other tokens converge to the position of the first one for almost every initial position of the tokens. In fact, we have asymptotic stability of this consensus equilibrium. This holds for any number of heads and any time varying P matrix provided it is bounded. Similar conclusions are reported under Theorem 4.1 in (Karagodin et al., 2024) by imposing stronger assumptions: time invariance of $P = Q^\top K$ and existence of a single attention head. Theorem 4.5 extends these result to the case where U is a time invariant symmetric matrix and the multiplicity of its largest eigenvalue is one. In this case all the tokens will converge to a consensus equilibrium (moreover, that equilibrium is asymptotically stable) if they start in one of the two hemispheres defined by the eigenvector associated with the largest eigenvalue of U. We were only able to establish this result for the single-head case although we believe it holds in greater generality. To the best of the author's knowledge there is no result available in the literature for the case where U is not the identity matrix although this is conjectured, but not proved, in (Karagodin et al., 2024).

Our theoretical findings are validated by simulations of the mathematical model for attention. Moreover, experiments with the GPT-2 and the GPT-Neo models provide empirical evidence for convergence to consensus equilibria in more general situations than those captured by our theoretical results, thus providing additional confirmation for model collapse.

Notations

We use the letters n, ℓ, r, and s to denote elements of $\mathbb{N} = \{1, 2, \ldots\}$. The space of $r \times s$ real matrices is denoted by $\mathcal{M}_{r\times s}(\mathbb{R})$. In particular, $\mathbb{I}_r \in \mathcal{M}_{r\times r}(\mathbb{R})$ denotes the identity matrix. The Frobenius norm of a square matrix $A \in \mathcal{M}_{r\times r}(\mathbb{R})$ is denoted by $\|A\|$. The elements of $\mathbb{R}^{n+1}$ are denoted by x, and tuples of ℓ elements are denoted by $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_\ell) \in (\mathbb{R}^{n+1})^\ell$ (note the different font). The tangent space of a smooth manifold M at $p \in M$ and its elements are denoted by T_pM and $X_p \in T_pM$, respectively. Given another smooth manifold N and a smooth map $\phi : M \to N$, i.e., $\phi \in C^\infty(M, N)$, its tangent map is denoted by $T\phi : TM \to TN$.

2. Dynamics of transformers

2.1. Configuration space

Let $\ell, n \in \mathbb{N}$. A symmetric, positive-definite matrix $W \in \mathcal{M}_{(n+1)\times(n+1)}(\mathbb{R})$ defines an inner product on $\mathbb{R}^{n+1}$:

$$\langle X_\mathrm{x}, Y_\mathrm{x} \rangle_W = X_\mathrm{x}^\top \, W \, Y_\mathrm{x},$$

for each $X_\mathrm{x}, Y_\mathrm{x} \in T_\mathrm{x}\mathbb{R}^{n+1}$ and $\mathrm{x} \in \mathbb{R}^{n+1}$, where the superscript $\top$ denotes the transpose. The corresponding norm is denoted by $|X_\mathrm{x}|_W = (X_\mathrm{x}^\top \, W \, X_\mathrm{x})^{1/2}$. The points of $\mathbb{R}^{n+1}$ of unit norm define an n-dimensional ellipsoid, which is

$$\mathrm{y}_i(k+1) = \pi_W \left(\mathrm{y}_i(k) + \sum_{\eta=1}^{h} \sum_{j=1}^{\ell} W_\eta(k)\, V_\eta(k)\, D_\eta(k)_{ii}\, \exp\left(\mathrm{y}_j(k)^\top K_\eta(k)^\top Q_\eta(k)\, \mathrm{y}_i(k) \right) \mathrm{y}_j(k) \right). \tag{1}$$

denoted by:

$$\mathcal{E}_W^n = \{\mathrm{x} \in \mathbb{R}^{n+1} \mid \mathrm{x}^\top W \mathrm{x} = 1\}.$$

In this work, we consider a transformer consisting of ℓ tokens of dimension $n+1$ constrained to evolve on an ellipsoid. As we have ℓ tokens, the resulting state space is the Cartesian product of ℓ copies of the ellipsoid, i.e.:

$$(\mathcal{E}_W^n)^\ell = \underbrace{\mathcal{E}_W^n \times \ldots \times \mathcal{E}_W^n}_{\ell\text{-times}}.$$

Similarly, we consider the following projection:

$$\pi_W : \mathbb{R}_0^{n+1} \to \mathcal{E}_W^n, \quad \mathrm{x} \mapsto \pi_W(\mathrm{x}) = \mathrm{x}\, |\mathrm{x}|_W^{-1}, \tag{2}$$

where $\mathbb{R}_0^{n+1} = \mathbb{R}^{n+1} - \{0\}$, whose tangent map at each $\mathrm{x} \in \mathbb{R}_0^{n+1}$:

$$T_\mathrm{x}\pi_W : T_\mathrm{x}\mathbb{R}_0^{n+1} \to T_{\pi_W(\mathrm{x})}\mathcal{E}_W^n,$$

is given by:

$$T_\mathrm{x}\pi_W \cdot X_\mathrm{x} = |\mathrm{x}|_W^{-1} \left(\mathbb{I}_{n+1} - \mathrm{x}\,\mathrm{x}^\top W\, |\mathrm{x}|_W^{-2} \right) \cdot X_\mathrm{x}, \tag{3}$$

for each $X_\mathrm{x} \in T_\mathrm{x}\mathbb{R}_0^{n+1}$. In particular, for $\mathrm{y} \in \mathcal{E}_W^n$, we have $T_\mathrm{y}\pi_W \cdot X_\mathrm{y} = \left(\mathbb{I}_{n+1} - \mathrm{y}\,\mathrm{y}^\top W \right) \cdot X_\mathrm{y}$.

Remark 2.1 (Evolution on the sphere). There are a number of models in which the tokens evolve on the n-sphere, i.e., $\mathbb{S}^n = \mathcal{E}_{\mathbb{I}_{n+1}}^n$. For brevity, in that case we will drop the subscripts standing for the matrix $W = \mathbb{I}_{n+1}$. For instance, we will write $|\cdot| = |\cdot|_{\mathbb{I}_{n+1}}$, $\boldsymbol{\pi} = \boldsymbol{\pi}_{\mathbb{I}_{n+1}}$, etc.

2.2. Discrete-time attention model

In this section we present the mathematical model for a transformer. Similarly to (Geshkovski et al., 2023a), the model encompasses the self-attention mechanism, the skip connection, and the normalization layer, but excludes the feedforward layer.

Let $w \in \mathbb{N}$ be a design parameter. The weight matrices at the k-th layer of the transformer, $k \in \mathbb{N}$, are denoted by $Q(k) \in \mathcal{M}_{w\times(n+1)}(\mathbb{R})$, $K(k) \in \mathcal{M}_{w\times(n+1)}(\mathbb{R})$ and $V(k) \in \mathcal{M}_{w\times(n+1)}(\mathbb{R})$, and are typically known as the *Query*, *Key*, and *Value*[3] matrices, respectively. The input to the k-th layer is denoted by $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_\ell) \in \mathcal{M}_{(n+1)\times\ell}(\mathbb{R})$ and the output $\boldsymbol{z} \in \mathcal{M}_{w\times\ell}(\mathbb{R})$ of the self-attention mechanism is given by:

$$\boldsymbol{z}(k) = V(k)\mathbf{x}(k)D(k)\exp\left(\mathbf{x}(k)^\top K(k)^\top Q(k)\mathbf{x}(k)\right), \tag{4}$$

where $\exp(\cdot)$ denotes the entry-wise exponential (i.e., $[\exp(R)]_{ij} = e^{R_{ij}}$), and $D(k) \in \mathcal{M}_{\ell\times\ell}(\mathbb{R})$ is defined as:

$$D(k)_{ij} = \left(\sqrt{n+1} \sum_{l=1}^{\ell} \exp(\mathrm{x}_l(k)^\top K(k)^\top Q(k)\, \mathrm{x}_i(k)) \right)^{-1},$$

if $i = j$, and $D(k)_{ij} = 0$ otherwise.

Practical transformer applications often distribute the computations of the self-attention mechanism through several parallel *heads*, leading to what is commonly known as *multi-headed self-attention*. To make explicit the dependence on the head, we write (4) as:

$$\boldsymbol{z}_\eta(k) = V_\eta(k)\,\mathbf{x}(k)\,D_\eta(k)\,\exp\left(\mathbf{x}(k)^\top K_\eta(k)^\top Q_\eta(k)\,\mathbf{x}(k)\right),$$

for each $1 \leq \eta \leq h$.

The outputs from all attention heads are added after being multiplied by certain weight matrices $W_\eta \in \mathcal{M}_{(n+1)\times w}(\mathbb{R})$, $1 \leq \eta \leq \ell$. Then, the resulting sum is added to the input of the layer $\mathbf{x}(k)$, using what is often called a *skip connection*. Lastly, a normalization function is applied to ensure that the output is bounded. In this work, we consider functions that normalize each token of the transformer separately, which is known as *layer normalization* and was first proposed in (Ba et al., 2016). Hence, the normalization function $\mathbf{N} : \mathcal{M}_{(n+1)\times\ell}(\mathbb{R}) \to \mathcal{M}_{(n+1)\times\ell}(\mathbb{R})$ is of the form:

$$\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_\ell) \mapsto \mathbf{N}(\mathbf{x}) = (N(\mathrm{x}_1), \ldots, N(\mathrm{x}_\ell)),$$

for some $N : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$. Similarly to (Geshkovski et al., 2023a), in the following we consider the normalization function $N = \pi_W$ given in (2), which projects each token to the ellipsoid $\mathcal{E}_W^n$. In practice, this projection has been explicitly used in some models such as (Jiang et al., 2023). For clarity, we utilize the symbol $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_\ell)$ for the tokens evolving on the ellipsoid (after this explicit choice of normalization). The resulting discrete-time dynamical system is shown in (1) where $1 \leq i \leq \ell$ indexes each token.

Remark 2.2 (Standard layer normalization). The standard layer normalization utilized in most transformers is given by:

$$N(\mathrm{x}) = \frac{1}{\sigma(\mathrm{x})}(\mathrm{x} - \mu(\mathrm{x})\mathbf{1}) \star \gamma + \beta,$$

[3] In the introduction we used U to refer to the value matrix; this difference is resolved in this section.

$$\dot{\mathrm{y}}_i = T_{\mathrm{y}_i}\pi_W \cdot \left(\sum_{\eta=1}^{h} \sum_{j=1}^{\ell} W_\eta(t)\, V'_\eta(t)\, D_\eta(t)_{ii}\, \exp\left(\mathrm{y}_j^\top K_\eta(t)^\top Q_\eta(t)\, \mathrm{y}_i\right) \mathrm{y}_j \right). \tag{5}$$

for each $\mathrm{x} = (x^1, \ldots, x^{n+1}) \in \mathbb{R}^{n+1}$. In the previous expression, $\mathbf{1}$ denotes the vector $(1, \ldots, 1) \in \mathbb{R}^{n+1}$, $\star$ denotes the element-wise product of vectors, and $\gamma, \beta \in \mathbb{R}^{n+1}$ are the *learned scale* and *shift*, respectively. Similarly, $\mu(\mathrm{x})$ and $\sigma(\mathrm{x})$ denote the mean and standard deviation of x, respectively. Under this normalization, we have:

$$\begin{aligned} |N(\mathrm{x}) - \beta|^2 &= \frac{1}{\sigma(\mathrm{x})^2} \sum_{\mu=1}^{n+1} (x^\mu - \mu(\mathrm{x}))^2\, (\gamma^\mu)^2 \\ &= \frac{n+1}{\sum_{\mu=1}^{n+1} (x^\mu - \mu(\mathrm{x}))^2} \sum_{\mu=1}^{n+1} (x^\mu - \mu(\mathrm{x}))^2\, (\gamma^\mu)^2, \end{aligned}$$

where we used the notation $\gamma = (\gamma^1, \ldots, \gamma^{n+1})$. It is clear that, if $\gamma = \gamma_0 \mathbf{1}$ for some $\gamma_0 \neq 0$, then the tokens lie on the n-sphere of center β and radius $(n+1)\gamma_0^2$. Therefore, the results in this paper also apply to this type of normalization. We conjecture that, for arbitrary scale parameters, the tokens will also lie on a certain hypersurface of $\mathbb{R}^{n+1}$ diffeomorphic to the n-sphere, but a more careful analysis has to be carried out to understand the geometry of such hypersurface.

2.3. Continuous-time attention model

In this section we introduce additional notation that is only used to derive the continuous time model. Readers not interested in the model's derivation can skip the next two paragraphs and start reading the paragraph commencing with "To simplify notation".

Let $Y \in \mathfrak{X}((\mathcal{E}_W^n)^\ell)$ be a vector field and denote its flow by $Y^\tau : (\mathcal{E}_W^n)^\ell \to (\mathcal{E}_W^n)^\ell$. Note that Y is complete, i.e., its flow is defined for each $\tau \in \mathbb{R}$, since $(\mathcal{E}_W^n)^\ell$ is compact. Given a map $g : (\mathcal{E}_W^n)^\ell \times \mathbb{R} \to \mathbb{R}_0^+$, we use the notation $g(\mathbf{y}, \tau) = O_{\mathbf{y}}(\tau^2)$ to denote the existence of a constant $T \in \mathbb{R}^+$ and a function $\sigma : (\mathcal{E}_W^n)^\ell \to \mathbb{R}_0^+$ such that, for each $\tau \in [0, T]$ and $\mathbf{y} \in (\mathcal{E}_W^n)^\ell$, we have $g(\mathbf{y}, \tau) \leq \sigma(\mathbf{y})\tau^2$. A map $\phi : (\mathcal{E}_W^n)^\ell \times \mathbb{R} \to (\mathcal{E}_W^n)^\ell$ is a first order approximation to the flow Y^τ if $\mathbf{d}(Y^\tau(\mathbf{y}), \phi(\mathbf{y}, \tau)) = O_{\mathbf{y}}(\tau^2)$ where $\mathbf{d}$ denotes the distance on $(\mathcal{E}_W^n)^\ell$ induced by the Euclidean distance on $(\mathbb{R}_0^{n+1})^\ell$.

Using the concepts introduced in the previous paragraph, our objective is to construct a vector field Y such that the map defined by the right-hand side of (1) is the best first order approximation of Y^τ. To that end, we write $V_\eta(k)$ as $V_\eta(k) = \tau V'_\eta(k)$ for each $1 \leq \eta \leq h$, with $0 < \tau \ll 1$ being a small parameter. Hence, (1) may be rewritten as:

$$\mathrm{y}_i(k+1) = \pi_W(\mathrm{y}_i(k) + \tau\, f_k(\mathbf{y}(k)), \qquad 1 \leq i \leq \ell,$$

where $f_k : (\mathcal{E}_W^n)^\ell \to \mathbb{R}^{n+1}$ is defined as:

$$\begin{aligned} f_k(\mathbf{y}) = \sum_{\eta=1}^{h} \sum_{j=1}^{\ell} & W_\eta(k)\, V_\eta(k)'\, D_\eta(k)_{ii} \\ & \exp\left(\mathrm{y}_j^\top K_\eta(k)^\top Q_\eta(k)\, \mathrm{y}_i\right) \mathrm{y}_j, \end{aligned}$$

for each $\mathbf{y} \in (\mathcal{E}_W^n)^\ell$. For each $1 \leq i \leq \ell$, the best linear approximation in τ is given by:

$$\dot{\mathrm{y}}_i = \left.\frac{d}{d\tau}\right|_{\tau=0} \pi_W(\mathrm{y}_i + \tau\, f_k(\mathbf{y})) = T_{\mathrm{y}_i}\pi_W \cdot f_k(\mathrm{y}_i).$$

Therefore, the continuous-time model is given by (5), with $1 \leq i \leq \ell$, $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_\ell) \in (\mathcal{E}_W^n)^\ell$, and $t \geq 0$, as the differential equation whose solution provides the best first order approximation of (1).

To simplify notation we introduce the following (time-dependent) auxiliary matrices:

$$\begin{aligned} U_\eta(t) &= W_\eta(t)\, V'_\eta(t) \in \mathcal{M}_{(n+1)\times(n+1)}(\mathbb{R}), \\ P_\eta(t) &= Q_\eta(t)^\top K_\eta(t) \in \mathcal{M}_{(n+1)\times(n+1)}(\mathbb{R}), \end{aligned}$$

for each $1 \leq \eta \leq h$ and $t \geq 0$. We still refer to the matrix $U_\eta(t)$ as the value matrix since it plays a similar role. Similarly, we define the functions $\alpha_{ij}^\eta, Z_i^\eta : \mathbb{R}_0^+ \times (\mathcal{E}_W^n)^\ell \to \mathbb{R}$ by:

$$\begin{aligned} \alpha_{ij}^\eta(t, \mathbf{y}) &= \frac{1}{Z_i^\eta(t, \mathbf{y})} \exp(\mathrm{y}_i^\top P_\eta(t)\, \mathrm{y}_j), \\ Z_i^\eta(t, \mathbf{y}) &= D_\eta(t)_{ii}^{-1} = \sqrt{n+1} \sum_{j=1}^{\ell} \exp(\mathrm{y}_i^\top P_\eta(t)\, \mathrm{y}_j), \end{aligned}$$

respectively, for each $1 \leq i, j \leq \ell$, $1 \leq \eta \leq h$, $t \geq 0$ and $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_\ell) \in (\mathcal{E}_W^n)^\ell$. The matrix having α_{ij}^η as its i-th row and j-th column entry is usually called the attention matrix of head η.

With the notation just introduced, the dynamical system that describes the evolution of a transformer with h heads and ℓ tokens evolving on the ellipsoid $\mathcal{E}_W^n$ is given by:

$$\boxed{\dot{\mathrm{y}}_i = T_{\mathrm{y}_i}\pi_W \cdot \left(\sum_{\eta=1}^{h} \sum_{j=1}^{\ell} \alpha_{ij}^\eta(t, \mathbf{y})\, U_\eta(t)\, \mathrm{y}_j \right)} \tag{6}$$

for each $1 \leq i \leq \ell$, $t \geq 0$ and $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_\ell) \in (\mathcal{E}_W^n)^\ell$.

3. Full self-attention matrix

In this section we consider a particular case of the model (6) described by the following assumptions.

Assumption 3.1. For each head $1 \leq \eta \leq h$, we have:

1. $U_\eta(t) = \mathbb{I}_{n+1}$,
2. $P_\eta(t)$ is bounded, i.e., $\sup_{t\geq 0} \|P_\eta(t)\| < \infty$, and
3. $P_\eta(t)$ is uniformly continuous on $[0, \infty[$.

Recall that a set $\mathcal{C} \subset (\mathcal{E}_W^n)^\ell$ is said to be *attractive* for (6) with domain of attraction $\mathcal{H} \subset (\mathcal{E}_W^n)^\ell$ if the following conditions hold:

1. $\mathcal{C}$ is forward-invariant, i.e., each solution of (6) starting in $\mathcal{C}$ remains in $\mathcal{C}$ for all positive times.
2. $\lim_{t\to\infty} \operatorname{dist}(\mathbf{y}(t), \mathcal{C}) = 0$ for each solution $\mathbf{y} : \mathbb{R}_0^+ \to (\mathcal{E}_W^n)^\ell$ of (6) with $\mathbf{y}(0) \in \mathcal{H}$, where $\operatorname{dist}(\mathbf{y}(t), \mathcal{C}) = \inf_{z\in\mathcal{C}} \mathbf{d}(\mathbf{y}(t), \mathrm{z})$ and $\mathbf{d}$ denotes the distance on $(\mathcal{E}_W^n)^\ell$ induced by the Euclidean distance on $(\mathbb{R}_0^{n+1})^\ell$.

The next result claims attractivity of the consensus set provided that the initial position of the tokens is in some open hemisphere of the ellipsoid.

Theorem 3.2. Let $v \in \mathcal{E}_W^n$ and consider the open hemisphere:

$$\mathcal{H}^+(v) = \{\mathrm{y} \in \mathcal{E}_W^n \mid v^\top y > 0\}.$$

If Assumption 3.1 holds, then the consensus set $\mathcal{C}_\ell^+(v)$ in the product of hemispheres $\mathcal{H}^+(v)^\ell$, given by:

$$\mathcal{C}_\ell^+(v) = \{\mathbf{y} = (\mathrm{y}, \ldots, \mathrm{y}) \in (\mathcal{E}_W^n)^\ell \mid \mathrm{y} \in \mathcal{H}^+(v)\},$$

is attractive for (6) with domain of attraction $\mathcal{H}^+(v)^\ell$.

The idea of the proof is to first check that $\mathcal{H}^+(v)^\ell$ is forward-invariant and then define a Lyapunov function as follows:

$$V : \mathcal{H}^+(v)^\ell \to \mathbb{R}, \quad \mathbf{y} \mapsto V(\mathbf{y}) = \max_{1\leq i\leq \ell} V_i(\mathbf{y}), \tag{7}$$

where $V_i(\mathbf{y}) = 1 - v^\top \mathrm{y}_i$. In other words, V describes how far is each token y_i from being aligned with v by representing the largest misalignment, i.e., the worst case over all tokens. We show that V is strictly decreasing along solutions of (6) that start in $\mathcal{H}^+(v)^\ell$, except when the solution lies in $\mathcal{C}_\ell^+(v)$, i.e., when the tokens reach consensus. It follows from the Lyapunov stability theory that, under these conditions, $\mathcal{C}_\ell^+(v)$ is attractive. We only obtain attractivity instead of asymptotic stability due to the system being time-varying. It should be pointed out that V is defined by a maximum and, thus, it may not be differentiable. To overcome this, we compute the upper Dini derivative instead of the standard one (which is not defined at the discontinuities) using Danskin's theorem. This gives an upper bound on the change of V along the solution that allows for a Lyapunov-like theorem.

Remark 3.3 (Closest result available in the literature). Similar conclusions appear in (Geshkovski et al., 2023a) (see Lemma 4.2) under the stronger assumptions of a single attention head and that both U and $P = Q^\top K$ are the identity matrix.

Remark 3.4 (Higher dimensions). Let us restrict ourselves to the case where we have normalization to the sphere, i.e., $W = \mathbb{I}_{n+1}$. Wendel's theorem (cf. Eq. (1) of (Wendel, 1962)) gives the probability that ℓ tokens lie on the same hemisphere when distributed uniformly at random; namely:

$$\mathcal{P}_{\ell,n} = \frac{1}{2^{\ell-1}} \sum_{\mu=0}^{n-1} \binom{\ell-1}{\mu}.$$

In particular, $\mathcal{P}_{\ell,n} = 1$ whenever $n \geq \ell$. As a result, if the starting position of the tokens is chosen from a uniformly random distribution and $n \geq \ell$, then they will lie on the same hemisphere almost surely. The previous result thus deals with the most general situation for higher dimensions.

4. Auto-regressive self-attention matrix

This section addresses the auto-regressive (also known as causal) case, that is, the case where the dynamics of each token only depends on itself and the previous tokens. This corresponds to the model (6) with the so-called *auto-regressive self-attention matrix*, i.e.:

$$\alpha_{ij}^\eta(t, \mathbf{y}) = \begin{cases} \dfrac{1}{Z_i^\eta(t, \mathbf{y})} \exp(\mathrm{y}_i^\top P_\eta(t)\, \mathrm{y}_j), & i \geq j, \\ 0, & i < j, \end{cases}$$

$$Z_i^\eta(t, \mathbf{y}) = \sqrt{n+1} \sum_{j=1}^{i} \exp(\mathrm{y}_i^\top P_\eta(t)\, \mathrm{y}_j).$$

4.1. Identity value matrix

Let us consider the case where $W = \mathbb{I}_{n+1}$, i.e., the tokens evolve on the sphere.

Assumption 4.1. The model (6) is auto-regressive, $W = \mathbb{I}_{n+1}$ and, for each head $1 \leq \eta \leq h$, we have:

1. $U_\eta(t) = \mathbb{I}_{n+1}$, and
2. $P_\eta(t)$ is bounded, i.e., $\sup_{t\geq 0} \|P_\eta(t)\| < \infty$.

It is straightforward that, under the previous conditions, $\dot{\mathrm{y}}_1 = 0$ and, thus, the first token remains fixed: $\mathrm{y}_1(t) = \mathrm{y}_1^0$ for each $t \geq 0$. For each other token y_i, $2 \leq i \leq \ell$, the inner product between y_1^0 and y_i, i.e., the (cosine of the) angle between them, provides a projection of the dynamics onto the real line. In other words, using this angle we construct a scalar differential equation governing the evolution of the projection of the token on the real line. It them becomes simple to construct a Lyapunov function for the second token, ensuring convergence to y_1^0 for every initial condition except for $y_2(0) = -y_1^0$. For the remaining tokens, an input-to-state stability argument coupled with the triangular nature of the dynamics leads to the asymptotic stability of the consensus set for almost all initial conditions.

Theorem 4.2. If Assumption 4.1 holds, then the consensus set:

$$\mathcal{C}_\ell = \{\mathbf{y} = (y, \dots, y) \in (\mathbb{S}^n)^\ell\},$$

is asymptotically stable for the system (6) and the domain of attraction contains the following set:

$$\mathcal{D}_\ell^1 = \{(y_1, \dots, y_\ell) \in (\mathbb{S}^n)^\ell \mid y_j \neq -y_1,\ 2 \leq j \leq \ell\}.$$

Remark 4.3 (Closest result available in the literature). Similar conclusions are reported under Theorem 4.1 in (Karagodin et al., 2024) by imposing stronger assumptions, time invariance of $P = Q^\top K$ and existence of a single attention head, although the authors state that time-invariance is not explicitly used.

4.2. Symmetric value matrix

Now we extend the results of the previous section to more general value matrices. As above, the tokens evolve on the sphere, i.e., $W = \mathbb{I}_{n+1}$.

Assumption 4.4. The model (6) is auto-regressive, $W = \mathbb{I}_{n+1}$, and we have:

- There is only one head, i.e., $h = 1$,
- $U_1(t) = U$ with $U^\top = U$, and
- $P_1(t) = P(t)$ is bounded, i.e., $\sup_{t \geq 0} \|P(t)\| < \infty$.

We denote the spectrum of U by $\lambda(U)$. Note that $\lambda(U) \subset \mathbb{R}$ as U is symmetric. Given $\lambda \in \lambda(U)$, the corresponding eigenspace is denoted by $L_\lambda(U) \subset \mathbb{R}^{n+1}$.

Unlike the case $U = \mathbb{I}_{n+1}$ considered in the previous section, the first token is no longer fixed. However, it can be shown that it converges to a fixed position provided the multiplicity of the largest eigenvalue is one. This allows for establishing the asymptotic stability of two specific consensus points induced by the matrix U using the same technique as in Theorem 4.2. Namely, the dynamics of each other token y_i, $2 \leq i \leq \ell$, is projected to the real line using its inner product with the corresponding asymptotically stable equilibrium of y_1. By following an induction argument and treating the distance of the previous tokens to the equilibrium as an error (as it converges to zero within time), an input-to-state-stability-Lyapunov function for the projected dynamics of y_i can be found, yielding the result.

Theorem 4.5. Suppose that Assumption 4.4 holds and $\dim L_\lambda(U) = 1$, where $\lambda = \max \lambda(U)$. The elements of $L_\lambda(U) \cap \mathbb{S}^n$ are denoted by $\{-v, v\}$. If $\lambda > 0$, then $\mathbf{y}^* = (v, \dots, v)$ (resp. $\mathbf{y}^* = -(v, \dots, v)$) is an asymptotically stable equilibrium of (6) and its domain of attraction contains the set:

$$\mathcal{D}^\ell(v) = \{(y_1, \dots, y_\ell) \in (\mathbb{S}^n)^\ell \mid v^\top y_i > 0,\ 1 \leq i \leq \ell\},$$
$$\text{(resp. } \mathcal{D}^\ell(-v)\text{)}.$$

Remark 4.6 (Closest results available in the literature). The authors were not able to find results in the literature addressing the case where U is not the identity matrix although two conjectures are proposed, but not proved, in (Karagodin et al., 2024).

Remark 4.7 (Time-varying value matrix). In this remark we discuss a possible relaxation of Assumption 4.4 to a time-varying matrix $U(t)$. Once $U(t)$ becomes time-varying, so will $\lambda(t)$, $v(t)$, and the consensus equilibrium $\mathbf{y}^*(t) = (v(t), \dots, v(t))$. If $v(t)$ varies slowly, we expect convergence to a ball centered at $\mathbf{y}^*(t)$. The scalar differential equation, upon which the proof of Theorem 4.5 is built, becomes:

$$\dot{b} = \tilde{\alpha}(t)\,(\lambda(t) - y^\top U(t)\, y)\, b + \dot{v}^\top y, \quad t \geq 0. \tag{8}$$

Let us denote the extra term as $e = \dot{v}^\top y$. When $e = 0$, $b^* = 1$ is an asymptotically stable equilibrium with domain of attraction $]0, 1]$ (resp. $b^* = -1$ with domain $[-1, 0[$). Hence, an input-to-state stability argument would show convergence to a ball centered at $b^* = 1$ (resp. $b^* = -1$) whose radius is an increasing function of $\sup_{t \geq 0} |e(t)|$.

5. Simulations and empirical validation

In this section we illustrate the theoretical results and show that their conclusions appear to hold even when our assumptions are violated. We start by simulating the continuous transformer model and illustrating our theoretical results. In addition to simulations, we provide empirical evidence using the GPT-2 XL and the GPT-Neo 2.7B to show how token consensus seems to occur even if the assumptions in our theoretical results are not satisfied.

5.1. Numerical simulations

5.1.1. Illustration of Theorem 3.2.

We simulate the motion of 10 tokens, each of them randomly placed on the sphere $\mathbb{S}^2 \subset \mathbb{R}^3$, according to the dynamics (6) with $h = 2$. All matrices, except for $P_1(t)$ and $P_2(t)$,

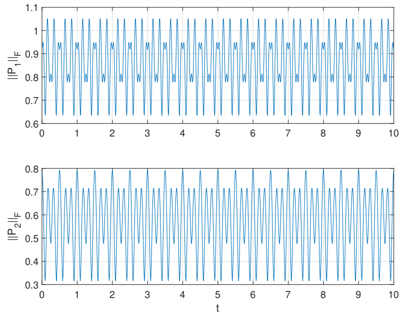

Figure 1. Frobenius norm of the matrices $P_1(t)$ and $P_2(t)$.

were randomly chosen, and each element was drawn from a uniform distribution in the interval $[-0.5, 0.5]$. The matrices $P_1(t)$ and $P_2(t)$ were computed as $P_1(t) = D_1(t)P_1'$, $P_2(t) = D_2(t)P_2'$ with P_1' and P_2' randomly generated:

$$P_1' = \begin{pmatrix} 0.08 & -0.19 & 0.20 \\ -0.23 & 0.31 & -0.23 \\ 0.18 & -0.17 & -0.16 \end{pmatrix},$$

$$P_2' = \begin{pmatrix} -0.31 & 0.03 & 0.11 \\ 0.06 & -0.06 & 0.13 \\ 0.14 & 0.11 & 0.10 \end{pmatrix}.$$

The matrices $D_1(t)$ and $D_2(t)$ were given by $D_1(t) = 2\,\mathrm{diag}\left(\cos(10\pi t), \ \sin(10\pi t), \ \cos(6\pi t)\right)$ and $D_2(t) = 2\,\mathrm{diag}\left(\cos(6\pi t), \ \sin(6\pi t), \ \cos(4\pi t)\right)$, where $\mathrm{diag} : \mathbb{R}^3 \to \mathcal{M}_{3\times 3}(\mathbb{R})$ denotes the function that maps a vector to the diagonal matrix with its components on the diagonal.

To better appreciate the time-varying nature of the matrices P_1 and P_2, in Figure 1 we shown their Frobenius norm.

In Figure 2 we show the motion of the tokens in blue with their initial position represented by a white circle and final position by a gray circle. We can appreciate that all the tokens start and remain in an hemisphere and that they converge to a consensus equilibrium.

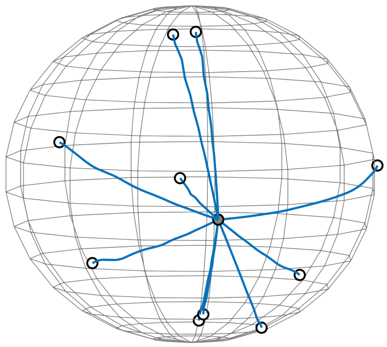

Figure 2. Convergence to a consensus equilibrium on the sphere $\mathbb{S}^2$. All the tokens start and remain in an hemisphere.

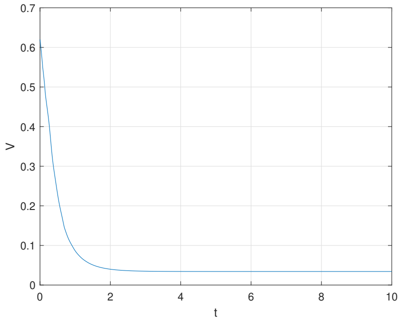

Figure 3. Evolution of the Lyapunov function (7) used in the proof of Theorem 3.2.

The proof of Theorem 3.2 is based on the Lyapunov function (7), whose time-evolution is displayed in Figure 3 for the case where $v = (1, 0, 0)$.

5.1.2. Illustration of Theorem 4.2

We now consider the auto-regressive model with 50 tokens on $\mathbb{S}^{499} \subset \mathbb{R}^{500}$. The number and dimension of the tokens were chosen to make them comparable to the GPT-2 model. We use two heads ($h = 2$) with the matrices $P_1 = D_1(t)P_1'$ and $P_2 = D_2(t)P_2'$ obtained by randomly generating P_1' and P_2', and taking $D_1(t)$ and $D_2(t)$ to be diagonal with entries $(D_\eta)_{jj} = |2\sin(wt+\phi)|$ for $\eta = 1, 2$, $j = 1, \ldots, 500$, w drawn from the uniform distribution on $]0, 1[$ and ϕ drawn from the uniform distribution on $]0, 2\pi[$. To measure the error between tokens we use the cosine similarity, $E : (\mathcal{E}_W^n)^\ell \to \mathbb{R}^+$, defined as:

$$E = 1 - \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{\mathrm{y}_1^\top \mathrm{y}_i}{|\mathrm{y}_1|\,|\mathrm{y}_i|}, \tag{9}$$

which becomes zero when all the tokens belong to the consensus set.

In Figure 4 we display the evolution of the function E along 100 trajectories of (6) for random initial conditions drawn from an element-wise uniform distribution on $]-0.5, 0.5[$, and then projected to the sphere. We can appreciate in Figure 4 how the function E converges to zero along all the trajectories.

5.1.3. Illustration of Theorem 4.5

In the final case we use the auto-regressive model with 10 tokens on $\mathbb{S}^2$ with randomly assigned initial positions. As for the previous cases, we choose $P(t) = D(t)P'$, with

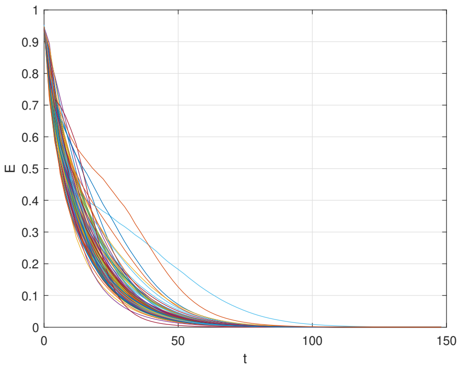

Figure 4. Illustration of Theorem 4.2; evolution of the function E defined in (9) along 100 solutions of (6) with random initial conditions drawn from an element-wise uniform distribution on $]-0.5, 0.5[$ and then projected to the sphere.

randomly generated P' and U given by:

$$P' = \begin{pmatrix} 0.36 & 0.42 & 0.13 \\ 0.10 & -0.07 & -0.20 \\ 0.15 & -0.21 & 0.12 \end{pmatrix},\ U = \begin{pmatrix} -0.26 & 0.50 & 0.56 \\ 0.50 & -0.72 & -0.50 \\ 0.56 & -0.50 & -0.02 \end{pmatrix},$$

and $D(t) = 2\,\mathrm{diag}\,\big(\cos(10\pi t),\ \sin(10\pi t),\ \cos(6\pi t)\big)$.

In Figure 5 we can observe convergence of the tokens to a consensus equilibrium point whereas in Figure 6 we have the time evolution of $V_1 = 1-\mathrm{y}_1^\top v$ and $V_2 = 1-\mathrm{y}_2^\top v$ where $v \in \mathbb{R}^3$ is the eigenvector of U corresponding to its largest eigenvalue. Note that V_2 is not a Lyapunov function, and therefore it may increase, although the proof of Theorem 4.5, establishes that it will eventually converge to zero.

5.2. GPT-2 and GPT-Neo Experiments

In this section we report on experiments conducted on the GPT-2 XL model and the GPT-Neo 2.7B model suggesting that our theoretical findings hold under more general assumptions. Since our results are asymptotic, we need to increase the depth of both models. We do so by running the same set of tokens through the model multiple times. In other words, we extract the tokens at the end of the model, after the final normalization, and feed them to the model for another pass thereby simulating a model of increased length, the code used for our experiments is available at (Silvestre, 2025).

For all our experiments, we used the same set of 100 random prompts, each generated by uniformly sampling 200 tokens from the GPT-2 tokenizer's vocabulary. In each experiment, we plot the average of E across all the prompts. As an example, the first 10 tokens of the first sampled prompt are:

```
divest anxYou coasts Oz
Vi Happy appreciate tcp.
```

In the first experiments, we removed the feedfoward layers of each model, to make them closer to the structure we assume in our theoretical work. The experiments were then repeated without removing the feedforward layers, showing that in both cases convergence to consensus occurs.

The experiments were conducted on the standard configuration of the GPT-2 XL model and the GPT-Neo 2.7B model, using the pre-trained weights provided by the Hugging Face library (Wolf et al., 2020). The multiple passes through the model results in matrices P and U that are time-varying but periodic with period corresponding to the depth each model: 48 layers for the GPT-2 XL and 32 for the GPT-Neo 2.7B. To measure how far the tokens are from each other we used the function (9) whose evaluation after each layer is depicted in Figures 7 and 8.

We can observe that in both models the average of the function E over all the prompts converge asymptotically to 0, thus implying the tokens converge to a consensus equilibrium. We recall that our theoretical results predict this

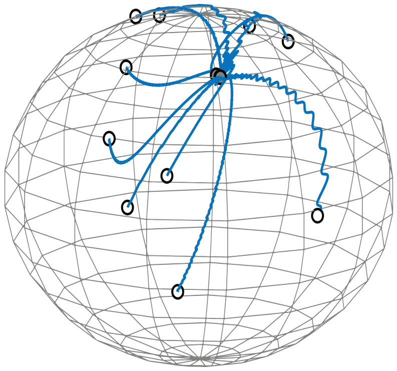

Figure 5. Convergence to a consensus equilibrium on the sphere $\mathbb{S}^2$. All the tokens start and remain in the hemisphere defined by v.

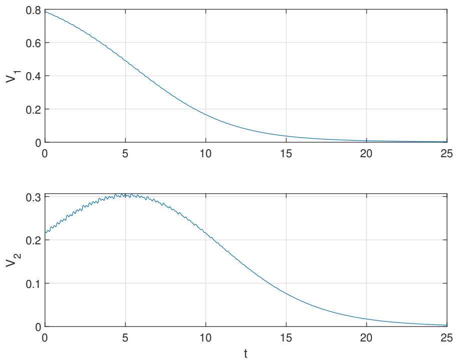

Figure 6. Evolution of the functions V_1 and V_2 used in the proof of Theorem 4.5.

observation **only** when feedforward layers are absent, i.e., the case depicted in Figure 7. However, as can be seen in Figure 8, even when the feedfoward layers are present, convergence still occurs. The rate of convergence appears to be dependent on the weights of the feedfoward layers as their presence increases the convergence rate in the GPT-2 model, but decreases it in the GPT-Neo model. In both cases, these findings suggest that feedfoward layers may not be sufficient to preclude consensus.

Although the previous experimental results suggest that consensus occurs even in the presence of feedfoward layers, it does not address the question of consensus being a structural property of the the transformer architecture or of the choice of weight matrices. To address this question we repeated the experiments by using random matrices in GPT-2 and GPT-Neo. Since this results in time-varying matrices, we further repeated the experiments by randomly selecting new weight matrices before each model pass. Moreover, we conducted these experiments with the full model and also by removing the feedforward layers (including the associated normalization function and skip connection) to better understand the impact of these on token consensus. The results are reported in Figures 9 and 10, where we can see that convergence towards consensus still occurs across all experiments. Furthermore, Figures 9 and 10 suggest that the feedfoward layer may decrease the convergence rate.

Our experiments suggest that the convergence phenomenon is a product of the structure of the transformers and not of the choice of weights. We observe convergence with trained, random periodic, and random aperiodic matrices P and U. In terms of rate of convergence, the choice of weight matrices appears to have an impact with faster convergence being observed when pretrained matrices were used.

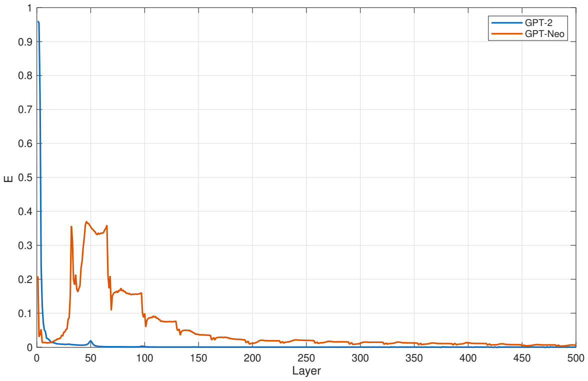

Figure 7. Comparison between the GPT-2 XL and GPT-Neo 2.7B architectures with feedforward layers removed; evaluation of the average of (9) across all the random prompts.

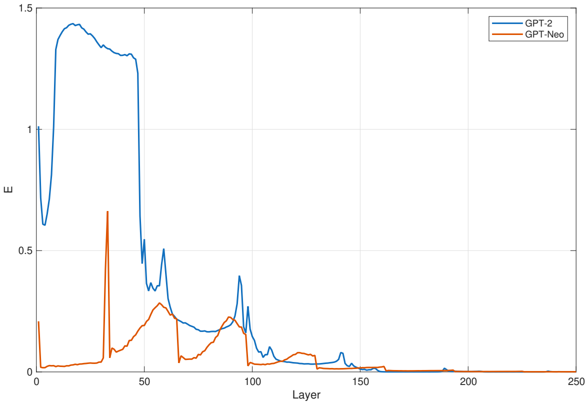

Figure 8. Comparison between the GPT-2 XL and GPT-Neo 2.7B architectures with the full model; evaluation of the average of (9) across all the random prompts.

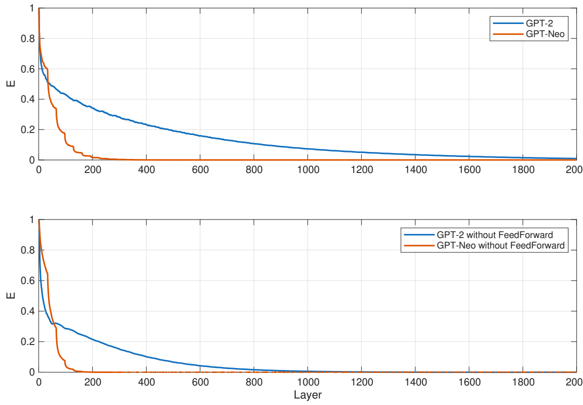

Figure 9. Comparison between the GPT-2 XL and GPT-Neo 2.7B architectures with fixed and randomly chosen weight matrices. Each model was evaluated with and without feedfoward layers using the average of (9) across all the random prompts.

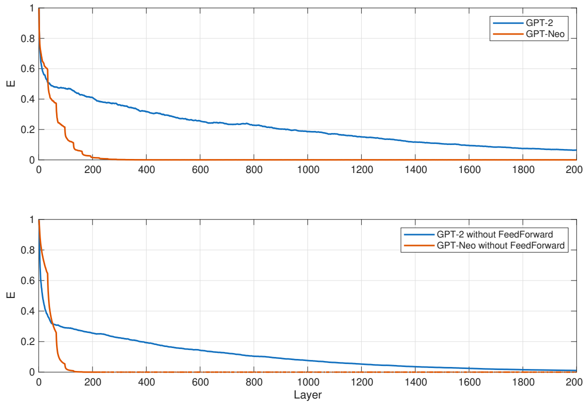

Figure 10. Comparison between the GPT-2 XL and GPT-Neo 2.7B architectures with random weight matrices chosen before each model pass. Each model was evaluated with and without feedfoward layers using the average of (9) across all the random prompts.

Acknowledgments

This work was partially supported by the NSF award 2211146. The work of João Pedro Silvestre was partially supported by the PhD fellowship 2023.01843.BD from the Fundação para a Ciência e a Tecnologia (FCT), Portugal.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Cao, Y., Yu, W., Ren, W., and Chen, G. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics*, 9(1):427–438, 2013. doi: 10.1109/TII.2012.2219061.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28, 2015.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dong, Y., Cordonnier, J.-B., and Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pp. 2793–2803. PMLR, 2021.

Dutta, S., Gautam, T., Chakrabarti, S., and Chakraborty, T. Redesigning the transformer architecture with insights from multi-particle dynamical systems. *arXiv preprint arXiv:2109.15142*, 2021.

Feng, R., Zheng, K., Huang, Y., Zhao, D., Jordan, M., and Zha, Z.-J. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35: 33054–33065, 2022.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. A mathematical perspective on transformers. *arXiv preprint arXiv:2312.10794*, 2023a.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. The emergence of clusters in self-attention dynamics. *arXiv preprint arXiv:2305.05465*, 2023b.

Geshkovski, B., Koubbi, H., Polyanskiy, Y., and Rigollet, P. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.

Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. A survey of deep learning techniques for autonomous driving. *Journal of field robotics*, 37(3):362–386, 2020.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Karagodin, N., Polyanskiy, Y., and Rigollet, P. Clustering in causal attention masking. *arXiv preprint arXiv:2411.04990*, 2024.

Lu, Y., Li, Z., He, D., Sun, Z., Dong, B., Qin, T., Wang, L., and Liu, T.-Y. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

Markdahl, J., Thunberg, J., and Gonçalves, J. Almost global consensus on the n-sphere. *IEEE Transactions on Automatic Control*, 63(6):1664–1675, 2017.

O'shea, K. and Nash, R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

Peng, B., Narayanan, S., and Papadimitriou, C. On limitations of the transformer architecture. *arXiv preprint arXiv:2402.08164*, 2024.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

Ren, W., Beard, R., and Atkins, E. A survey of consensus problems in multi-agent coordination. In *Proceedings of the 2005, American Control Conference, 2005.*, pp. 1859–1864 vol. 3, 2005. doi: 10.1109/ACC.2005.1470239.

Rodríguez Abella, Á., Silvestre, J. P., and Tabuada, P. The asymptotic behavior of attention in transformers. *arXiv preprint arXiv:2412.02682*, 2024.

Sainath, T. N., Vinyals, O., Senior, A., and Sak, H. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4580–4584. Ieee, 2015.

Sarlette, A. and Sepulchre, R. Consensus optimization on manifolds. *SIAM journal on Control and Optimization*, 48(1):56–76, 2009.

Silvestre, J. P. GPT-Consensus. `https://github.com/cyphylab/GPT-Consensus`, 2025.

Sontag, E. Smooth stabilization implies coprime factorization. *IEEE Transactions on Automatic Control*, 34(4): 435–443, 1989. doi: 10.1109/9.28018.

Sontag, E. D. The ISS philosophy as a unifying framework for stability-like behavior. In Isidori, A., Lamnabhi-Lagarrigue, F., and Respondek, W. (eds.), *Nonlinear control in the year 2000 volume 2*, pp. 443–467, London, 2001. Springer London. ISBN 978-1-84628-569-1.

Sontag, E. D. *Input to State Stability: Basic Concepts and Results*, pp. 163–220. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-77653-6. doi: 10.1007/978-3-540-77653-6_3. URL https://doi.org/10.1007/978-3-540-77653-6_3.

Thunberg, J., Markdahl, J., Bernard, F., and Goncalves, J. A lifting method for analyzing distributed synchronization on the unit sphere. *Automatica*, 96:253–258, 2018.

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., and Fox, E. A. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*, 2020.

Van Dijk, B., Kouwenhoven, T., Spruit, M. R., and van Duijn, M. J. Large language models: The need for nuance in current debates and a pragmatic perspective on understanding. *arXiv preprint arXiv:2310.19671*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wendel, J. G. A problem in geometric probability. *Mathematica Scandinavica*, 11(1):109–111, 1962. doi: https://doi.org/10.7146/math.scand.a-10655.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.