Picking an LLM's Brain: Thought and Fixation in Hidden Representations

Murdock Aubry Haoming Meng Anton Sugolov Vardan Papyan[1]

University of Toronto

Table 1. Summary of properties observed during each phase.

PHASE	LINEAR	EQUISPACED	ALIGNED	CONFIDENT
THOUGHT	LESS	LESS	MORE	LESS
FIXATION	MORE	MORE	LESS	MORE

Abstract

Large Language Models (LLMs) have made significant strides in natural language processing, and a precise understanding of the internal mechanisms driving their success is essential. We regard LLMs as discrete, coupled, nonlinear, dynamical systems in high dimensions. This perspective motivates tracing the trajectories of individual tokens as they pass through transformer blocks, and linearizing, along these trajectories, the system through their Jacobian matrices. These investigations uncover two distinct operational stages:

Thought. An exploratory phase, occurring in shallower layers where the model considers various possibilities for the next token. This stage is characterized by coordinated processing across layers, gradual expansion away from the origin through less linear trajectories, and low variation in confidence in the next token prediction.

Fixation. A more focused state, occurring in the deeper layers, akin to a fixation on a specific outcome or solution. This phase is distinguished by linear token trajectories at an increased velocity, reduced coordination in processing between layers, and highly variable prediction certainty.

Collectively, these findings reinforce the viewpoint of LLMs as dynamical systems and reveal a remarkable level of regularity that has previously been overlooked. These results lay the groundwork towards further transparency, explainability, and improvements in LLMs.

1. Introduction

Large language models (LLMs), as exemplified by BERT and GPT-series (Devlin et al., 2019; Brown et al., 2020), have revolutionized the field of natural language processing through their adoption of the transformer architecture (Vaswani et al., 2017). Despite their widespread success, the internal mechanisms that underpin their performance are not fully understood.

Previous works viewed certain types of deep networks as implementing discrete, nonlinear dynamical systems, operating in high dimensions (Greff et al., 2016; Papyan et al., 2017; Ebski et al., 2018; Chen et al., 2018; Bai et al., 2019; Rothauge et al., 2019; Li & Papyan, 2023; Gai & Zhang, 2021; Haber & Ruthotto, 2017; Ee, 2017). The term *discrete* reflects the network's finite depth; *nonlinear* refers to the model's nonlinear components; and *dynamical* is due to the residual connections spanning various layers.

In this work, we view LLMs as being *coupled* dynamical systems, due to the interdependent token trajectories enabled by self-attention. Adopting this perspective motivates us to trace the dynamics of individual tokens as they traverse through the numerous transformer blocks, and to linearize the system through Jacobian matrices along their trajectory. This investigation reveals two distinct operational phases, Thought and Fixation, characterized by properties summarized in Table 1.

1.1. Thought Phase

Prevalent in the shallower transformer blocks, this phase is characterized by:

Intuitively. An exploratory state where the LLM weighs different possibilities for the next token without definitive commitment. Representations appear to move erratically and indecisively, mirroring the process of *thinking*.

Algebraically. Coordinated processing across depth, as indicated by aligned top left and right singular vectors of several consecutive Residual Jacobians.

Geometrically. Less linear token trajectories with a gradual distancing from the origin.

Probabilistically. Comparable levels of confidence across

various prompts in predicting the next token.

1.2. Fixation Phase

Occurring in deeper transformer blocks, this phase is distinguished by:

Intuitively. Transitioning to a more focused, persistent mode, akin to a fixation on a particular outcome or solution. Representations move in a determined and focused manner, indicative of a strong conviction and decisive decision-making process.

Algebraically. Decreased alignment in Residual Jacobians' top singular vectors, signifying less coordinated processing at these depths.

Geometrically. Direct, linear token trajectories, accompanied by uniformly spaced embeddings.

Probabilistically. High variability and increased confidence in predicting the subsequent token.

1.3. Residual Alignment

Our investigation draws inspiration from a recent study by Li & Papyan (2023) on Residual Networks (ResNets) (He et al., 2016), which uncovered a phenomenon they termed Residual Alignment (RA). This phenomenon is marked by several distinct characteristics which include: *linear* trajectories in layer-wise progression, *equispaced* positioning of hidden representations, and *aligned* top left and right singular vectors in the linearizations of residual blocks across depths.

Our research takes builds on this phenomenon by contrasting these findings in ResNets with the patterns seen in LLMs. Through this approach, we aim to enrich the overall understanding of deep learning models, particularly in their function as discrete, nonlinear dynamical systems operating within high-dimensional spaces.

2. Background on Large Language Models

In the input layer, $l = 0$, textual prompts undergo tokenization and are combined with positional encodings to create an initial high-dimensional embedding, denoted by $x_i^0 \in \mathbb{R}^{d_{\text{model}}}$ for the i^{th} token. When these embeddings are stacked together, they form a matrix:

$$X^0 = (x_1^0, x_2^0, \ldots, x_n^0) \in \mathbb{R}^{d_{\text{model}} \times n}.$$

The embeddings then pass through L transformer blocks:

$$X^0 \xrightarrow{f^1_{\text{block}}} X^1 \xrightarrow{f^2_{\text{block}}} \cdots X^{L-1} \xrightarrow{f^L_{\text{block}}} X^L.$$

Here, $X^l = f^l_{\text{block}}(X^{l-1})$ denotes the embeddings after the l^{th} block, consisting of causal multi-headed attention (MHA), a feed-forward network (FFN), and normalization layers (LN) with residual connections:

$$\begin{aligned} h^{l+1}(X^l) &= \text{MHA}(\text{LN}(X^l)) \\ g^{l+1}(X^l) &= \text{LN}(X^l + h^{l+1}(X^l)) \\ f^{l+1}_{\text{block}}(X^l) &= X^l + h^{l+1}(X^l) + \text{FFN}(g^{l+1}(X^l)), \end{aligned}$$

where the MHA, LN, FFN are implicitly indexed by layer. In the final representation, an additional layer normalization is applied:

$$\begin{aligned} f^L_{\text{block}}(X^{L-1}) = \text{LN}(&X^{L-1} + h^L(X^{L-1}) \\ &+ \text{FFN}(g^L(X^{L-1}))). \end{aligned}$$

The output X^L from the final block f^L is passed into a bias-free linear layer $M \in \mathbb{R}^{d_{\text{vocab}} \times d_{\text{model}}}$, with d_{vocab} denoting the size of the token vocabulary and d_{model} is the dimension of the token embeddings. This layer M computes final-layer logits for each token embedding, $\ell_i^L = Mx_i^L$. The prediction for the next token is then determined by selecting the maximal logit value: $\arg\max_{v \in \text{tokens}} \ell^L_{v,n}$.

3. Methods

3.1. Suite of Large Language Models

Our empirical study focuses on three LLMs: Llama-2 (Touvron et al., 2023), Falcon (Almazrouei et al., 2023), and GPT-2 (Radford et al., 2019). These models, provided through HuggingFace (Wolf et al., 2020), vary in terms of parameter budgets, number of layers, and hidden dimensions. A summary of the models under consideration is presented in Table 2 below.

Table 2. Summary of models used for the experiments in this paper.

MODEL	PARAMETERS	LAYERS (L)	DIM. (d_{MODEL})
LLAMA-2	13 B	40	5120
	7 B	32	4096
FALCON	7 B	32	4544
GPT-2	1.5 B	48	1600
	774 M	36	1280
	355 M	24	1024
	117 M	12	768

3.2. Prompt Data

We evaluate these LLMs using prompts of varying length, ambiguity, and context, sourced from the SQuAD v2.0 QA dataset (Rajpurkar et al., 2018). The prompts are structured in a consistent format: `(context) + (question) + "The answer is:"`. For detailed information, see Appendix B.1. Post-prompting, we assess the models using several metrics, detailed in the following subsections.

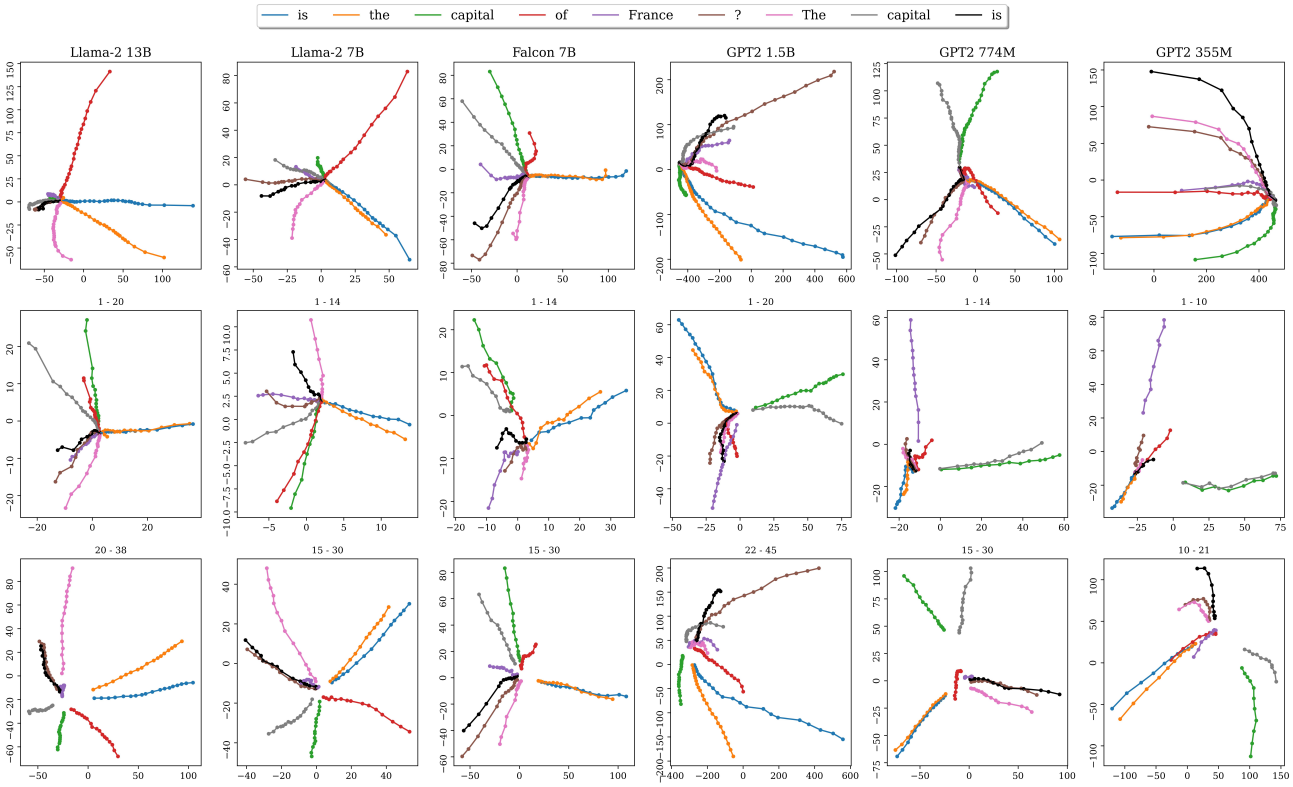


Figure 1. **Hidden Trajectories in LLMs.** Principal components of the trajectories of the hidden representations through various LLMs (columns, decreasing in model size, see Table 2) in the prompt: `What is the capital of France? The capital is`. Top row: all layers. Middle Row: layers in shallower transformer blocks (layers specified above plot). Bottom Row: layers in deeper transformer blocks (layers specified above plot). Trajectories of each input token (last token 'is' is plotted in black) are plotted in latent space, visualized with a 2-dimension principal component projection. Representations proceed in distinct outward directions, especially in the second half of transformer blocks (lower row) during which the norm of representations increases, with possible abrupt change in the last layer (outer points in upper row). A clear direction of movement is visible in each token trajectory.

3.3. Visualization of Trajectories

Each initial embedding x_i^0 forms the trajectory $x_i^0, x_i^1, \ldots, x_i^L$ as it passes through L transformer blocks. The dynamics in high-dimensional space are visualized through a 2-dimensional principal component (PC) projection, PC_L, fitted to the last layer embeddings $X^L = (x_1^L, x_2^L, \ldots x_n^L)$. The projected embeddings, $\text{PC}_L(x_i^0)$, $\text{PC}_L(x_i^1)$, $\ldots$, $\text{PC}_L(x_i^L)$, are plotted for each of the $i = 1, \ldots, n$ trajectories.

3.4. Linearity of Trajectories

Linearity in intermediate embeddings is quantified with the *line-shape score* (LSS), defined by Gai & Zhang (2021) as

$$\text{LSS}_i^{0,\ldots,L} = \frac{L}{\left|\left|\tilde{x}_i^L - \tilde{x}_i^0\right|\right|_2}, \qquad (1)$$

where $\tilde{x}_i^0 = x_i^0$ and $\tilde{x}_i^l$ is defined recursively as

$$\tilde{x}_i^l = \tilde{x}_i^{l-1} + \frac{x_i^l - x_i^{l-1}}{\left|\left|x_i^l - x_i^{l-1}\right|\right|_2} \quad \text{for } l = 1, \ldots, L.$$

Note that $\text{LSS} \geq 1$, with $\text{LSS} = 1$ if intermediate representations $x_i^0, \ldots, x_i^L$ form a co-linear trajectory.

3.5. Equidistance of Embeddings

Equispacing of consecutive hidden representations of the i-th token is quantified via

$$U_i = \sqrt{\frac{1}{L}\sum_{\ell=0}^{L-1} \frac{\left(\|x_i^l - x_i^{l+1}\|_2 - \overline{x}_i\right)^2}{\overline{x}_i^2}} \quad i = 1, \ldots, n, \qquad (2)$$

where $\overline{x}_i = \text{Ave}_i \|x_i^l - x_i^{l+1}\|_2$. This is the standard deviation over the mean of the perturbation norms, $\|x_i^l - x_i^{l+1}\|_2$.

3.6. Alignment of Residual Jacobians

To further investigate the hidden representations of transformer architectures, we examine the properties of the transformer blocks and the relationships between them. This is done by analyzing the linearizations of the blocks given by their *Residual Jacobian* matrices

$$J_l^i = \frac{\partial}{\partial x_i^{l-1}} \left(\left(h^l(X^{l-1}) + \text{FFN}_l(g^l(X^{l-1})) \right)_i \right),$$

for $l = 1, \ldots, L$, $i = 1, \ldots, n$. Note that this is the Jacobian matrix for each block without the contribution from the skip connection from the input, analogous to the quantities measured by Li & Papyan (2023).

The singular value decompositions of these J_l^i are computed, i.e., $J_l = U_l S_l V_l^\top$ (with superscript i, indicating the token, omitted for clarity), where $U_l \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ and $V_l \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ are the matrices of left and right singular vectors respectively, and $S_l \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ is the singular value matrix.

The matrices $A_{i,j,K} \coloneqq U_{j,K}^\top J_i V_{j,K}$ are plotted over all pairs of depths $i, j \in \{1, \ldots, L\}$, where $U_{j,K}$ and $V_{j,K}$ are the sub-matrices with columns that are the top-K left and right singular vectors of J_j, respectively.

To quantify the alignment of singular vectors of the Residual Jacobians at depths i and j, we measure the ratio of the average absolute off-diagonal entry to the average absolute diagonal entry. More precisely, if the entries of $A_{i,j,K}$ are given by $(a_{k_1 k_2})_{k_1,k_2=1}^K$, we define

$$r(A_{i,j,K}) = \frac{\frac{1}{K(K-1)} \sum_{k_1 \neq k_2} |a_{k_1 k_2}|}{\frac{1}{K} \sum_{k_1 = k_2} |a_{k_1 k_2}|} = \frac{\sum_{k_1 \neq k_2} |a_{k_1 k_2}|}{(K-1) \sum_{k_1 = k_2} |a_{k_1 k_2}|}.$$

3.7. Uncertainty of Predictions

Typically, the output of the last transformer block, x_n^L, is passed to the linear classifier, M. A softmax function is then applied to the logits, Mx_n^L, to obtain the probabilities for the next token in the vocabulary, $P_n^L = \text{Softmax}(\ell_n^L) = \text{Softmax}(Mx_n^L)$.

One could alternatively pass the output of any intermediate transformer block, x_n^l, to the linear classifier, M. Applying a softmax to the result, Mx_n^l, also yields the probabilities for the next token, $P_n^l = \text{Softmax}(\ell_n^l) = \text{Softmax}(Mx_n^l)$, however, as determined up to that specific, earlier layer, rather than the final one.

Given the probabilities P_n^l, uncertainty in the next-token prediction can be quantified through its entropy, defined as:

$$S(P_n^l) = -\sum_{v \in \text{vocab.}} P_n^l(v) \log(P_n^l(v)).$$

Maximal uncertainty (entropy) $S(P) = \log(d_{\text{vocab}})$ is attained for uniform probabilities P over d_{vocab} elements. Minimal uncertainty $S(P) = 0$ is achieved when a single vocabulary element v is assigned probability $P(v) = 1$.

4. Results

4.1. Expansion of Embeddings

Expansive dynamics are observed throughout all layers of the model. The initial token embeddings X^0 are small in norm and begin to grow linearly and radially outward (Figures 1, 2). The representations consistently exhibit growth through the Thought phase, with the average layer-wise norm appearing largely prompt-independent (Figure 2). The beginning of the Fixation phase is characterized by an increase in token velocity and layer-wise norm becoming highly prompt-dependent. Changes in scale can also be observed from low dimensional projections of the representations (Figure 1; rows 2,3). In Llama-2, the mean norm of trajectories appears linear through each phase, while for GPT2 the mean norm grows faster than linear (Figure 2).

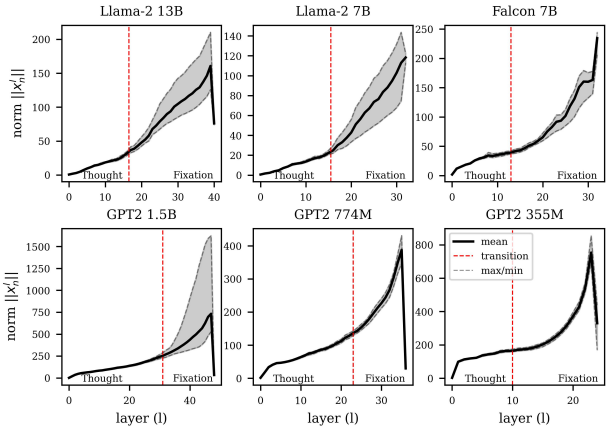


Figure 2. **Norms of Embeddings Increase With Depth.** Norm of the (last token) embedding versus depth for various LLMs (Table 2), averaged across 100 prompts. Error bands capture the minimal and maximal norm of a token trajectory across prompts. The dotted red line indicates the approximate transition between Thought and Fixation, characterized by increased embedding velocity and greater variability across prompts.

4.2. Linearity of Trajectories

Qualitatively, the linearity of the trajectories is evident in their low-dimensional projections (Figure 1).

Quantitatively, for each LLM, the trajectories attain a maximal LSS, indicative of minimal linearity, approximately at $1/4$ of the transformer depth, in the Thought phase (Figure

3). The transition from Thought to Fixation is distinguished by an inflection point in LSS, occurring around the same layer as the increase in norm velocity. During the Fixation phase, there is a notable decrease in LSS for each model, signifying enhanced linearity (Figure 3).

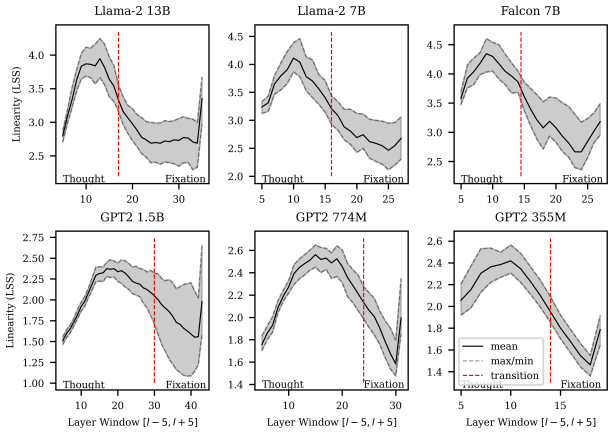


Figure 3. **Linearity of Intermediate Representations.** LSS versus depth for various LLMs (Table 2), averaged across 100 prompts. LSS is computed over a local window of 11 layers centered at varying depths l, i.e., $[l-5, l+5]$. The error band captures the minimal and maximal values across prompts. Included are approximate locations of the transition between Thought, characterized by minimal linearity (maximal LSS), and Fixation, characterized by maximal linearity (minimal LSS).

4.3. Equidistance of Embeddings

Qualitatively, the equidistance of the intermediate embeddings is clearly visible in their low-dimensional projections (Figure 1). Indeed, points are, for the most part, uniformly spaced from each other across all LLMs and depths.

Quantitatively, the variation in distances between intermediate embeddings peaks during the Thought phase (Figure 4), where the ratio of the standard deviation to the mean in the norms of perturbations reaches as high as 0.3 for the largest of the LLMs. The transition from Thought to Fixation phase is characterized by a significant reduction in this ratio, indicating that token perturbations become more uniform in norm. This change corresponds with the layer where norm velocity increases and where an inflection point in linearity measures is observed.

Throughout the Fixation phase, the distances between embeddings become notably consistent, maintaining a standard deviation to mean ratio of about 0.1 (Figure 4).

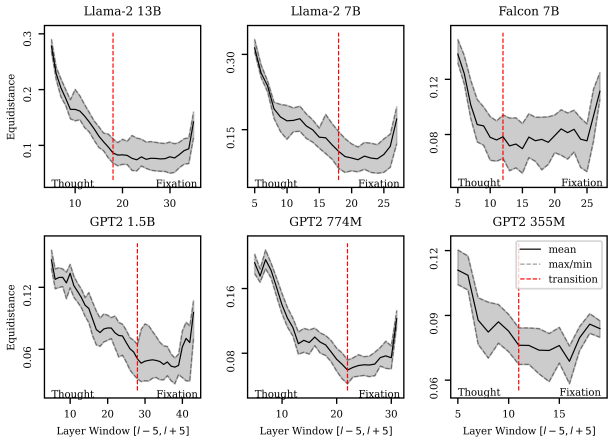


Figure 4. **Equidistance of Intermediate Representations.** Mean equidistance versus depth for various LLMs (Table 2), averaged across 100 prompts. Equidistance is computed over a local window of 11 layers centered at varying depths l, i.e, $[l-5, l+5]$. The error band captures the minimal and maximal values across prompts. Included are approximate locations of the transition between Thought, characterized by less equidistant embeddings, and Fixation, characterized by more equidistant embeddings.

4.4. Uncertainty of Predictions

Prediction certainty generally increases as tokens pass through transformer blocks (Figure 5). The Thought stage shows almost identical confidence among prompts, reflected in the low variation in entropy measurements. In each LLM, a maximal entropy of $\log(d_{\text{vocab}})$ is nearly achieved in the first layer, followed by a general increase in confidence. The transition between Thought and Fixation phases is characterized by an increase in variation between prompts. Greater confidence emerges in the Fixation stage, and varies depending on the information included in an individual input prompt. In some prompts, a prediction is completely fixated and total certainty is assigned in the deeper layers of GPT2 and Falcon (Figure 5). In other prompts, confidence fluctuates until a next token emerges, and a prediction is decided with low certainty.

4.5. Alignment of Residual Jacobians

We observe alignment of the top left and right singular vectors of the Jacobians J_l across depth (Figure 7), evident in the distinct diagonal lines present in the matrix subplots. This phenomenon is consistently observed across LLMs, similar to the observations made for ResNets (Li & Papyan, 2023).

Notably, the depths demonstrating the strongest alignment vary among the models, as depicted in Figure 6. Both Llama-2 models (7B and 13B) exhibit their strongest align-

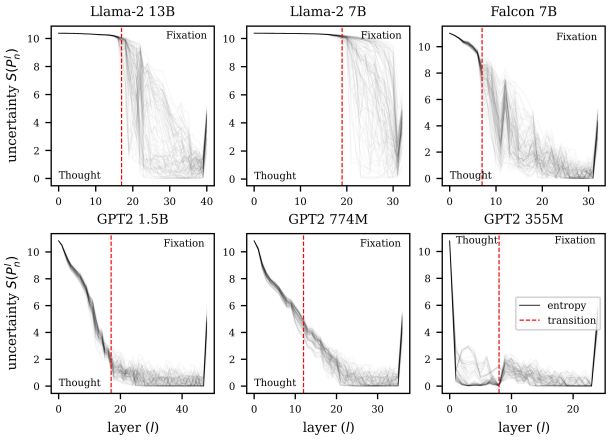


Figure 5. **Entropy of Final Token Representations.** Uncertainty in prediction probabilities P_n^l at each layer in various LLMs (Table 2), for 100 prompts (Section 3.2). Entropy after each decoder layer is plotted for each prompt. The approximate transition between the Thought phase (shallower layers) and Fixation phase (deeper layers) is labelled.

ment in the Thought phase, with some alignment reappearing in the final few layers. Falcon 7B has less Jacobian alignment in the initial few blocks, but demonstrates alignment deeper in the network. Meanwhile, the GPT-2 models have the strongest Jacobian alignment overall, extending to the deeper layers of the network. Across these models, strong alignment is consistently observed during the Thought phase, with models such as Llama-2 exhibiting notably less alignment during the Fixation phase.

4.6. Relation Between Metrics

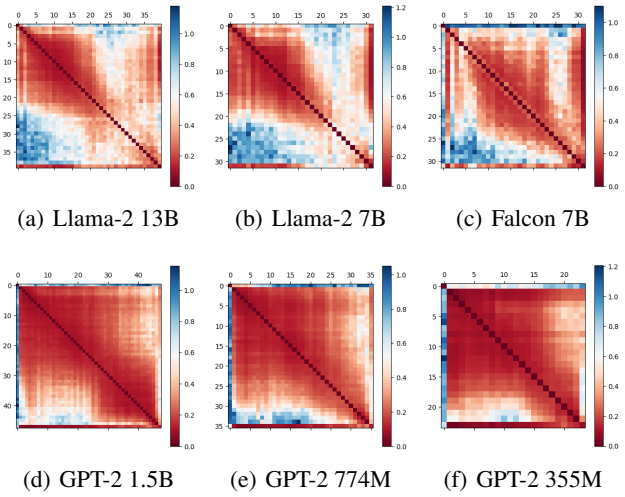
Our observations show expansion in norm (Section 4.1) linear trajectories (Section 4.2), and a general increase in model confidence in late embeddings (Section 4.4). These properties are interconnected, as detailed in the subsequent result.

Proposition 1. *Let $v, b \in \mathbb{R}^m$ where v has a single largest component, $v_1 > v_2, \ldots, v_m$. For elements on the line $\{\lambda v + b \mid \lambda \in \mathbb{R}\}$, probabilities Softmax$(\lambda v + b)$ decay in entropy for large λ.*

$$\lim_{\lambda \to \infty} S(softmax(\lambda v + b)) = 0$$

Proof. See Appendix A.2. □

Throughout our results, the embeddings $x_n^0, x_n^1, \ldots, x_n^L$ trace a line. Together with the linearity of the classifier, M, this implies that the logits are also traversing a line. Proposition 1 above suggests that such linear movement of logits induces a decrease in entropy, which aligns with our uncertainty measurements.



(a) Llama-2 13B (b) Llama-2 7B (c) Falcon 7B

(d) GPT-2 1.5B (e) GPT-2 774M (f) GPT-2 355M

Figure 6. **Strength of Residual Jacobian Alignment Across Different Depths**. The ratio of the average absolute value of the off-diagonal entries to the average absolute value of the diagonal entries of $A_{i,j,30} = U_{j,30}^\top J_i V_{j,30}$ are visualized across pairs of depth. In each subplot, each entry (i, j) has the value $r(A_{i,j,30})$ plotted, averaged across question prompts from SQuAD v2.0 Wikipedia questions about dogs. A lower value of $r(A_{i,j,30})$ (darker red in the plot) indicates stronger alignment between Residual Jacobians at depths i and j.

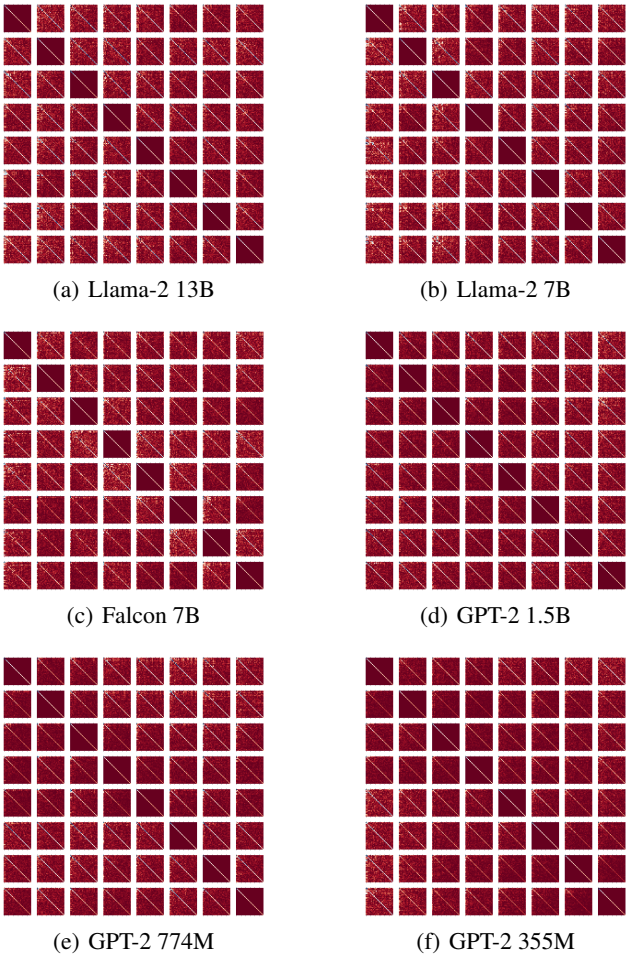
5. Discussion

5.1. Comparison to ResNets

Linearity and Equidistance. Linearity and equidistance in hidden representations, as observed in ResNets (Li & Papyan, 2023), are also present in LLMs. The LSS values (Figure 3) across all tested LLMs (Table 2) are between 2.0 and 3.0 through the Fixation stage. These values are similar to those observed for ResNets (Gai & Zhang (2021), page 18).

Residual Jacobian Alignment. The phenomenon of Residual Jacobian alignment is generally present in LLMs (Figure 7), with some variation among models in terms of the layers where it occurs. Stronger alignment is usually observed in the earlier to middle layers, contrasting with the previously observed alignment in ResNets, where alignment may be weaker in the initial layers but tends to strengthen later in the network (Li & Papyan, 2023). An interesting observation is that a significant portion of Jacobian alignment occurs before the model makes a decision on the next token, hinting at a potential relationship between alignment and the model's predictive capabilities.

The relationship between the strength of Jacobian alignment and linearity remains unclear in transformers; the Thought phase generally shows greater alignment despite having lower linearity, while the opposite is true in the Fixation phase. Further investigation is required to determine the exact relationship between Residual Jacobian alignment, linearity, and its emergence in transformers.



(a) Llama-2 13B (b) Llama-2 7B

(c) Falcon 7B (d) GPT-2 1.5B

(e) GPT-2 774M (f) GPT-2 355M

Figure 7. **Residual Jacobian Alignment of Top Singular Vectors**. The figure illustrates the alignment of Residual Jacobians across transformer blocks 9 to 16 (part of the Thought phase). For each subplot, in the square located at entry (i, j), the absolute values of the entries of matrices $A_{i,j,30} = U_{j,30}^{\top} J_i V_{j,30}$ (using the top 30 singular vectors) are visualized using the prompt `'What is the capital of France? The capital is'` with its final token. The diagonal line in the plots indicates alignment of Residual Jacobians, where the top singular vectors of J_j diagonalize J_i, with diagonal entries that are sufficiently large compared to the off-diagonal entries. Visualizations over all depth pairs are included in Appendix C.

5.2. Comparison Between Large Language Models

Llama-2. vs. Falcon. Thought and Fixation are clearly distinguished in linearity of trajectories, uncertainty of predictions, and alignment of Jacobians in Llama-2. The phase transition is not clear in Falcon, however the two phases are evident in shallower and deeper layers. Despite Llama-2 7B and Falcon 7B being comparable in size, measurements differ greatly.

Llama-2 vs. GPT-2. Llama-2 and the larger GPT-2 models show a distinct transition between Thought and Fixation. GPT-2 shows greater linearity and equispacing than Llama-2. Llama-2 features an almost maximal uncertainty during the Thought phase, while prompt-independent decrease is generally present in GPT-2. Residual Jacobians align throughout several layers in GPT-2, otherwise occuring only in Thought in Llama-2.

GPT-2 vs. Falcon. The transition between Thought and Fixation is clear in GPT-2, in contrast to Falcon. Linearity is significantly greater in GPT-2. Residual Jacobian alignment is evident in several layers in both LLMs.

5.3. Final Layer

There is a clear irregularity in the final representation as reflected in every metric (Figures 1, 2, 3, 4, 5). This may be attributed to a layer normalization applied in the final layer before a prediction is made. Among many prompts, the correct next-token prediction is fixated before the very last layer, while the normalization introduces greater uncertainty. The significance of the final normalization and its effect on training and generalization requires further investigation.

5.4. Effect of Model Size

In the LLMs considered in our work, the dynamical linearity of the representations X^l and overall uniformity of trajectories decrease with number of parameters (Figure C). Moreover, a consistent and significant drop in LSS and uniformity is observed in the Fixation phase (Figures 3, 4), the magnitude of which appears independent of model size. While still demonstrating Thought and Fixation phases, there is less distinction in smaller GPT2 models. To reinforce our understanding of the relationship between model size and the geometry of the trajectories, more LLMs with a varying number of parameters must be considered, and is left for future work.

5.5. Regularity and Training

The Llama-2 and Falcon models share similar architectures and sizes, and both were trained with identical weight decay and learning rate schedulers. From available data, their training differs only in the maximal learning rate: 3.0×10^{-4}

in Llama-2 7B while 1.85×10^{-4} in Falcon 7B (Touvron et al., 2023; Almazrouei et al., 2023). Yet, generalization capabilities are significantly greater in Llama-2, particularly in reading comprehension performance: Llama-2 7B scores 61.3, while Falcon 7B scores 36.0. This discrepancy in performance is further echoed by our findings that these models differ in the properties of their hidden representations, suggesting a potential relationship between generalization and regularity (Section 5.2).

Understanding the relationship between training methods and regularity in representations holds potential implications for the enhancement of LLMs. The development of techniques to amplify specific types of regularity could pave the way for more principled approaches to improving model performance. The exploration of the connection between hidden representation regularity, generalization, and LLM training, as prompted by our work, presents an intriguing research direction with important practical implications.

6. Related Work

Residual Networks. ResNets have been viewed as an ensemble of shallow networks (Veit et al., 2016), with studies delving into the scaling behaviour of their trained weights (Cohen et al., 2021). The linearization of residual blocks by their Residual Jacobians was first explored by Rothauge et al. (2019), who examined Residual Jacobians and their spectra in the context of stability analysis, and later by Li & Papyan (2023) who discovered Residual Alignment. We continue this line of work by further investigating Residual Jacobians in transformer architectures.

Information Processing in LLMs. Previously, Katz & Belinkov (2023); Bietti et al. (2023) have described memory and semantic flow in intermediate states of transformers while van Aken et al. (2019) have examined BERT hidden trajectories. Investigation of LLMs for reasoning tasks and interpretability remains an important research focus (Huang & Chang, 2023). Our work provides insight into information processing and prediction in LLMs through identifying two phases in the hidden representations, characterized by their geometric, algebraic, and probabilistic properties.

Hidden Representation Dynamics. Previous studies (Gai & Zhang, 2021; Haber & Ruthotto, 2017; Ee, 2017), have focused on interpreting deep neural networks from a dynamical systems perspective. In particular, the layer-wise transformations between transformer blocks are interpreted as discrete approximations to continuous curves through representation space. Throughout training, numerical experiments have shown that both plain networks and ResNets are aptly approximated by a geodesic curve in Wasserstein space (Gai & Zhang, 2021). Other interpretations involve partitioning activation space into basins of attraction, with fixed points representing discrete thoughts (Nam et al., 2023). For further related works, see (Geshkovski et al., 2023b;a; Tarzanagh et al., 2023; Valeriani et al., 2023).

Structure in Representations. Emergence of regularity in last-layer representations during training was noticed in Neural Collapse (Papyan et al., 2020). Motivated by this, Parker et al. (2023) examined a structure of representations analogous to Neural Collapse in intermediate hidden states. Our work studies the representations of all layers in LLMs, and discovers regularity in geometric, algebraic, and probabilistic properties in LLMs.

Information Geometry. A space of probability measures may be endowed with geometric structure, allowing for the analysis of information manifolds. In recent years, optimal transport has been investigated on spaces equipped with different geometric properties (Khan & Zhang, 2022; Rankin & Wong, 2023). Gai & Zhang (2021) have proposed ResNets as transporting an input distribution with the optimal transport map and induce equidistancing of representations in trajectories. Our work provides insight into the transformation of distributions in LLMs, with further relations to trajectories on the probability simplex in Appendix A.3.

7. Conclusion

A detailed description of the complexities of transformer blocks in the context of language modeling is an active area of research, and our primary goal was to enhance the understanding of the mechanics underlying transformer architectures. In this paper, we provided a quantitative description of the hidden representations and embedding trajectories within LLMs. We introduce Thought and Fixation phases which characterize and qualitatively describe the stages that the hidden representations endure when passing through the transformer blocks. The characteristics of each of these phases remains consistent among the models considered in this paper and occur across a wide variety of textual prompts, demonstrating vast applicability.

Our results have identified the dependence of prediction certainty, embedding trajectories, and Jacobian alignment on model size and training methods, and contain strong connections with the recently adopted dynamical system interpretation of deep learning. These findings open avenues for future research for a deeper understanding the connections between regularity of hidden representations and model specifications, LLM architecture, and generalizability.

8. 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

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., and Penedo, G. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

Bai, S., Kolter, J. Z., and Koltun, V. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. Birth of a transformer: A memory viewpoint, 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. 2020.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Cohen, A.-S., Cont, R., Rossier, A., and Xu, R. Scaling properties of deep residual networks, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Ebski, S. J., Arpit, D., Ballas, N., Verma, V., Che, T., and Bengio, Y. Residual connections encourage iterative inference. In *International Conference on Learning Representations*, 2018.

Ee, W. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5:1–11, 02 2017. doi: 10.1007/s40304-017-0103-z.

Gai, K. and Zhang, S. A mathematical principle of deep learning: Learn the geodesic curve in the wasserstein space, 2021.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. The emergence of clusters in self-attention dynamics. 2023a.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. A mathematical perspective on transformers, 2023b.

Greff, K., Srivastava, R. K., and Schmidhuber, J. Highway and residual networks learn unrolled iterative estimation. *arXiv preprint arXiv:1612.07771*, 2016.

Haber, E. and Ruthotto, L. Stable architectures for deep neural networks. *CoRR*, abs/1705.03341, 2017. URL http://arxiv.org/abs/1705.03341.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Huang, J. and Chang, K. C.-C. Towards reasoning in large language models: A survey, 2023.

Katz, S. and Belinkov, Y. Visit: Visualizing and interpreting the semantic information flow of transformers, 2023.

Khan, G. and Zhang, J. When optimal transport meets information geometry. *Information Geometry*, 5(1): 47–78, June 2022. ISSN 2511-249X. doi: 10.1007/s41884-022-00066-w. URL http://dx.doi.org/10.1007/s41884-022-00066-w.

Li, J. and Papyan, V. Residual alignment: Uncovering the mechanisms of residual networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Nam, A., Elmoznino, E., Malkin, N., Sun, C., Bengio, Y., and Lajoie, G. Discrete, compositional, and symbolic representations through attractor dynamics, 2023.

Papyan, V., Romano, Y., and Elad, M. Convolutional neural networks analyzed via convolutional sparse coding. *The Journal of Machine Learning Research*, 18(1):2887–2938, 2017.

Papyan, V., Han, X. Y., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117.

Parker, L., Onal, E., Stengel, A., and Intrater, J. Neural collapse in the intermediate hidden layers of classification neural networks, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad, 2018.

Rankin, C. and Wong, T.-K. L. Bregman-wasserstein divergence: geometry and applications, 2023.

Rothauge, K., Yao, Z., Hu, Z., and Mahoney, M. W. Residual networks as nonlinear systems: Stability analysis using linearization, 2019.

Tarzanagh, D. A., Li, Y., Zhang, X., and Oymak, S. Max-margin token selection in attention mechanism, 2023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., Ansuini, A., and Cazzaniga, A. The geometry of hidden representations of large transformer models, 2023.

van Aken, B., Winter, B., Löser, A., and Gers, F. A. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pp. 1823–1832, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3358028. URL https://doi.org/10.1145/3357384.3358028.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Veit, A., Wilber, M., and Belongie, S. Residual networks behave like ensembles of relatively shallow networks, 2016.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Huggingface's transformers: State-of-the-art natural language processing, 2020.

A. Mathematical Results

A.1. Additional metrics

For a given prompt, a certain target next-token prediction may be represented with an indicator probability density $P_{\text{target}}(v)$ ranging over all tokens v in the vocabulary.

$$P_{\text{target}}(v) = \begin{cases} 1, & v \text{ is target token} \\ 0, & \text{otherwise} \end{cases}$$

The resulting distance between intermediate densities P_n^t from P_{target} is quantified through the normalized total variation

$$\delta(P_n^l, P_{\text{target}}) = \frac{1}{2} \sum_{v \in \text{vocab.}} \left| P_n^l(v) - P_{\text{target}}(v) \right| \in [0, 1]$$

A.2. Proofs

Proposition 2. *Let $v, b \in \mathbb{R}^m$ where v has a single largest component, $v_1 > v_2, \ldots, v_m$. For elements on the line $\{\lambda v + b \mid \lambda \in \mathbb{R}\}$, probabilities Softmax$(\lambda v + b)$ decay in entropy for large λ.*

$$\lim_{\lambda \to \infty} S(\textit{softmax}(\lambda v + b)) = 0$$

Proof. Let $P_{\lambda v+b} = \text{Softmax}(\lambda v + b)$. Selecting the first component, namely $(P_{\lambda v+b})_1$.

$$\lim_{\lambda \to \infty} \frac{e^{\lambda v_1 + b_1}}{\sum_k e^{\lambda v_k + b_k}} = \lim_{\lambda \to \infty} \frac{1}{1 + \sum_{k>1} e^{\lambda(v_k - v_1) + b_k - b_1}} = 1$$

Since $v_k - v_1 < 0$ by assumption. A similar argument shows that the v_1 exponential term dominates in the component $(P_{\lambda v+b})_j$ for $1 < j \leq m$.

$$\begin{aligned} &\lim_{\lambda \to \infty} \frac{e^{\lambda v_j + b_j}}{\sum_k e^{\lambda v_k + b_k}} \\ &= \lim_{\lambda \to \infty} \frac{1}{e^{\lambda(v_1 - v_j) + b_1 - b_j} + \sum_{k \neq j} e^{\lambda(v_k - v_j) + b_k - b_j}} \\ &= 0 \end{aligned}$$

Given arbitrary $\delta > 0$ and choosing λ large such that $(P_{\lambda v+b})_1 > 1 - \delta$, all other components must satisfy $(P_{\lambda v+b})_j < \delta$. Computing the entropy:

$$\begin{aligned} S(P_{\lambda v}) &= -\sum_{j=1}^{m} (P_{\lambda v+b})_j \log(P_{\lambda v+b})_j \\ &\leq (1 - \delta) \log(1 - \delta) + n\delta \log(\delta) \end{aligned}$$

As $\delta \to 0$, we have $S(P_{\lambda v+b}) \to 0$. In particular,

$$\lim_{\lambda \to \infty} S(P_{\lambda v+b}) = 0$$

□

A.3. Geometry and Probability Density

Geometric properties and entropy decay coincide during the Thought and Fixation phases, related by result 1. A relationship between these properties is motivated by the correspondence of last-token hidden representations x_n^l to prediction probabilities P_n^l through $P_n^l = \text{softmax}(Mx_n^l)$. Proposition 1 and linearity of the transformer head classifier M therefore suggest a decay in classifier entropy $S(\text{softmax}(Mx_n^l))$ as $l \to L$, as discussed in Section 4.4.

This perspective suggests a general geometric description of hidden representations X^l and prediction. The space of possible d_{vocab}-dimensional probability vectors (the probability simplex) corresponding to all possible densities over d_{vocab} elements (i.e. the vocabulary of tokens) is

$$C = \left\{ x \in \mathbb{R}^{d_{\text{vocab}}} \;\middle|\; \sum_{i=1}^{d_{\text{vocab}}} x_i = 1,\, x_i \geq 0 \right\}$$

Suppose $x_l \in \mathbb{R}^H$ is a representation and $f : \mathbb{R}^H \to \mathbb{R}^V$ is a function on representation space. $\text{softmax} : \mathbb{R}^d_{\text{vocab}} \to C$ represents each element of $\mathbb{R}^V$ as an element of the probability simplex. The composition $\text{softmax} \circ f$ assigns to each hidden representation $x_t \in \mathbb{R}^H$ a probability vector $\text{softmax}(f(x_l)) \in C$ corresponding to a particular probability measure on vocabulary tokens. The trajectory $(x_l)_{l=1}^L$ therefore corresponds to a trajectory on the space of categorical probability densities over d_{vocab} tokens. In this work, we suggest that trajectories $(x_l)_{l=1}^L$ on the probability simplex tend towards densities with maximal probability assigned to the correct next-token vocabulary component.

B. Additional Figure Detail

B.1. Prompt Details

Several illustrative prompts are tested across all models, particularly non-ambiguous short prompts. (e.g "What is the capital of France? The capital is", "United States of") or questions that cannot be answered without additional context (e.g "What is my favourite colour? The answer is", "Which month is my birthday in? The answer is"). We utilize 100 questions with prepended context from the 'Dog' category are illustrated (Figures 2, 3, 5).

B.2. Intermediate Probabilities

The Thought phase begins at the initial layer, with almost maximal entropy $S(P_n^l)$, corresponding to the entropy of the uniform distribution over vocabulary items $S = \log(|d_{\text{vocab}}|)$. For each LLM, entropy decay in the initial Thought stage occurs with similar behaviour across prompts (Figures 8, 5); entropy is constant in Llama-2 7B, 13B models until layer 20, GPT-2 1.5B, 762M feature an decreasing decay, while Falcon 7B features several overlapping jumps. A gradual decay pattern is not clearly observed in GPT-2 345M, however, entropy behaviour agrees across prompts in early layers before becoming prompt-dependent in later transformer blocks (Figure 5). A constant maximal entropy in Llama-2 7B, 13B corresponds to uniform and low probability being assigned by P_n^l to the next-token prediction during the thought stage. In Falcon 7B and GPT-2 variants, larger next-token prediction probabilities are assigned in earlier layers (Figures 9, 5), with large layer-wise jumps in predicted token and probability until the Fixation phase.

A highly prompt-dependent Fixation stage follows the thought phase: question prompts with sufficient context and clarity show almost zero entropy, while ambiguous prompts fluctuate in entropy (Figure 8). During the Fixation stage in a non-ambiguous prompt, the density P_n^l assigns high probability (commonly ≈ 1) to a single vocabulary token, settling on the correct next token prediction (Figure 10). In an ambiguous prompt, a higher probability may be assigned to several tokens, typically including the correct next token (Figure 11). The entropy fluctuates in ambiguous prompts due to significant changes in layer-wise next-token prediction probability, since P_n^l does not settle on a single candidate token. The Fixation phase in Llama-2 13B, 7B is abrupt; certainty in a candidate token quickly grows during the transition from the thought phase (Figure 10, 11), immediately settling on the correct token. In Falcon 7B and GPT-2 variants, the transition from the thought phase is less distinct (Figure 9, 5), and features several changes before settling on the correct next-token prediction. The end of the Fixation stage is marked by a sudden increase in entropy in the last several layers across all measured GPT models. In the last layer, several new tokens are assigned non-zero probability (Figures 8, 10, 11, 5). Jumps in the last layer are also highly prompt-dependent, with additional context reducing entropy (Figure 8).

C. Additional Figures

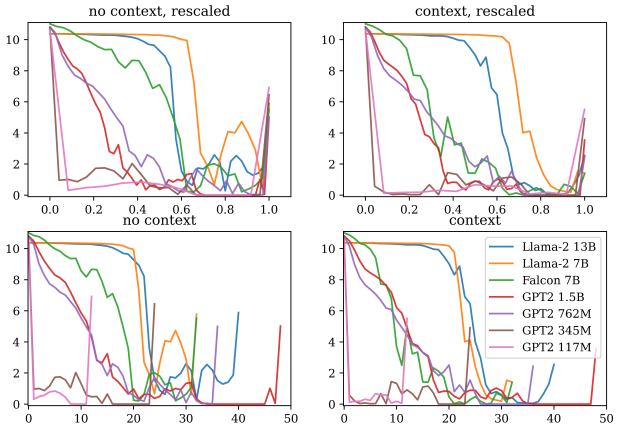


Figure 8. Intermediate entropy $S(P_n^l)$ among hidden layers for variants of Llama-2, Falcon, and GPT2 for the prompt 'What is the capital of this country? The capital is' with optional context. Context is introduced by prepending the first Wikipedia paragraph about the United Kingdom (3.2). Upper plots are rescaled for comparison of models with varying depths. With context, models make the correct next-token prediction ('London') while those without context do not. In the no context prompt, large oscillations are observed in the deeper transformer layers.

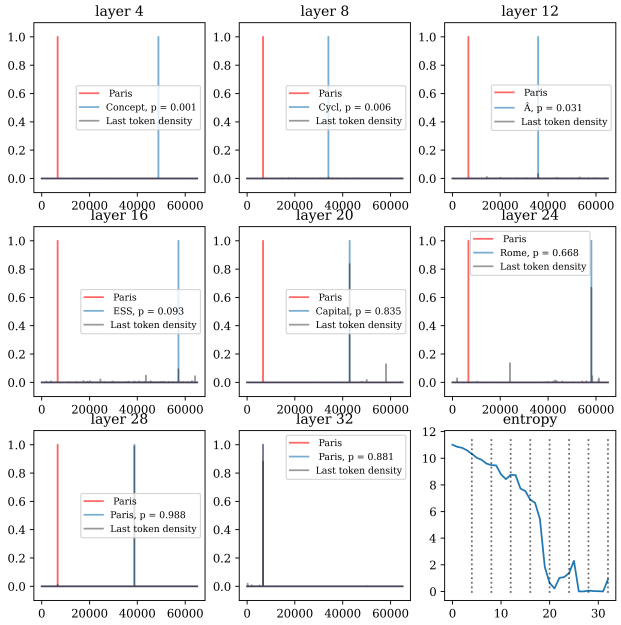


Figure 9. Falcon 7B, unambiguous prompt ('What is the capital of France? The capital is'), intermediate transformer head predictions and entropy at layers 4, 8, 12, 16, 20, 24, 28, 32. At the first layer, entropy is almost maximal ($S(P_n^0) = 11.067 \approx \log(64,000)$, 64k Falcon vocabulary size). Decay occurs for layers 0-20 and probability $> 10^{-3}$ is assigned to various incorrect tokens. In layer 20 the model correctly predicts 'Paris' with $p = 0.835$ and settles on the prediction throughout the remaining layers, despite briefly jumping to 'Rome' with high probability in layer 24. The final prediction is 'Paris' with $p = 0.861$.

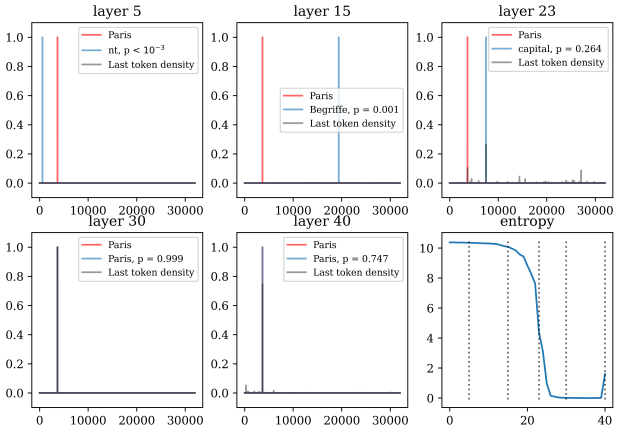


Figure 10. Llama-2 13B, unambiguous prompt ('What is the capital of France? The capital is'), intermediate transformer head predictions and entropy at layers 5, 15, 23, 30, and 40. At initial layers (0-18), predictions are highly uncertain (entropy is almost maximal: $S(P_n^0) = 10.373 \approx \log(32,000)$). A transition occurs at layer 23; a larger probability is assigned to 'Paris' and other tokens (e.g. 'capital'). From layer 25, the Fixation stage occurs: the 'Paris' token is assigned $p \approx 1$ until the final layer, in which it is assigned $p = 0.747$.

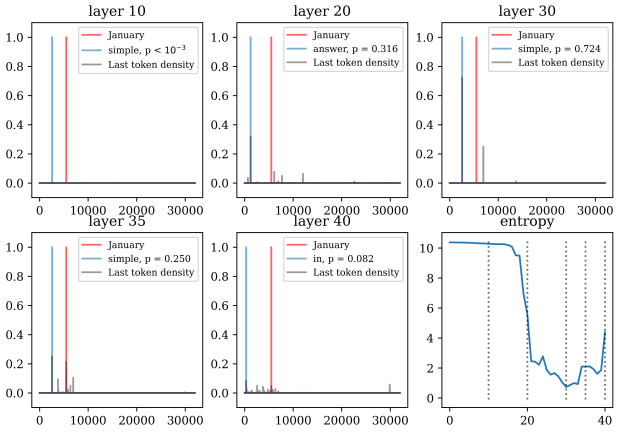


Figure 11. Llama-2 13B, ambiguous prompt ('What month was I born in? The answer is'), intermediate transformer head predictions and entropy at layers 10, 20, 30, 35, and 40. At initial layers (0-18), predictions are highly uncertain (entropy is almost maximal, $S(P_n^0) = 10.373 \approx \log(32,000)$). The entropy transition near layer 20 is followed by oscillations in entropy ($S(P_n^l) \neq 0$, in contrast to Figure 10). Non-zero probability is assigned to many tokens (e.g. 'simple' and 'January'). The final prediction is 'in' ($p = 0.082$), with similar probability to many other candidates (eg. 'January').

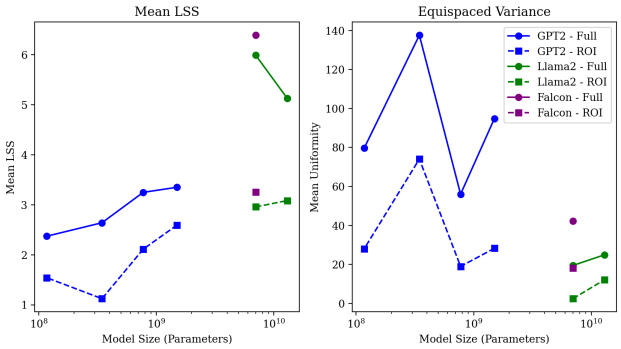


Figure 12. Number of parameters of Llama-2 13B, 7B, Falcon 7B, and GPT-2 1.5B, 762M, 345M versus mean layer-wise LSS (left) and mean variance (right) over (1) every layer in the transformer block and (2) only of the layers considered in the Fixation phase. In both cases (1) and (2), the GPT2 mean uniformity generally increases across the various model sizes, as is the case with Llama-2 and variants. Constant across all models is a sharp increase in dynamical uniformity in case (2) in comparison with case (1).

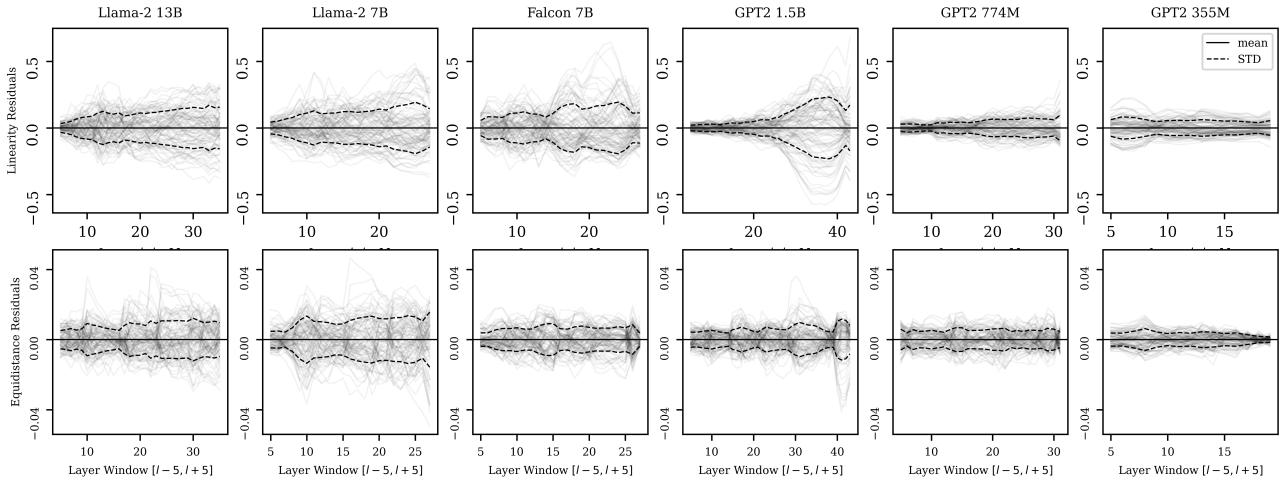


Figure 13. **LSS and Equidistance Residuals.** Residuals from the mean curve versus depth for various LLMs (Table 2), averaged across 100 prompts. Both LSS and Equidistance are computed over a local window of 11 layers centered at varying depths l, i.e., $[l-5, l+5]$. The dotted lines captures the first standard deviation of the data set at each layer. Included are approximate locations of the transition between Thought, characterized by minimal linearity (maximal LSS), and Fixation, characterized by maximal linearity (minimal LSS).

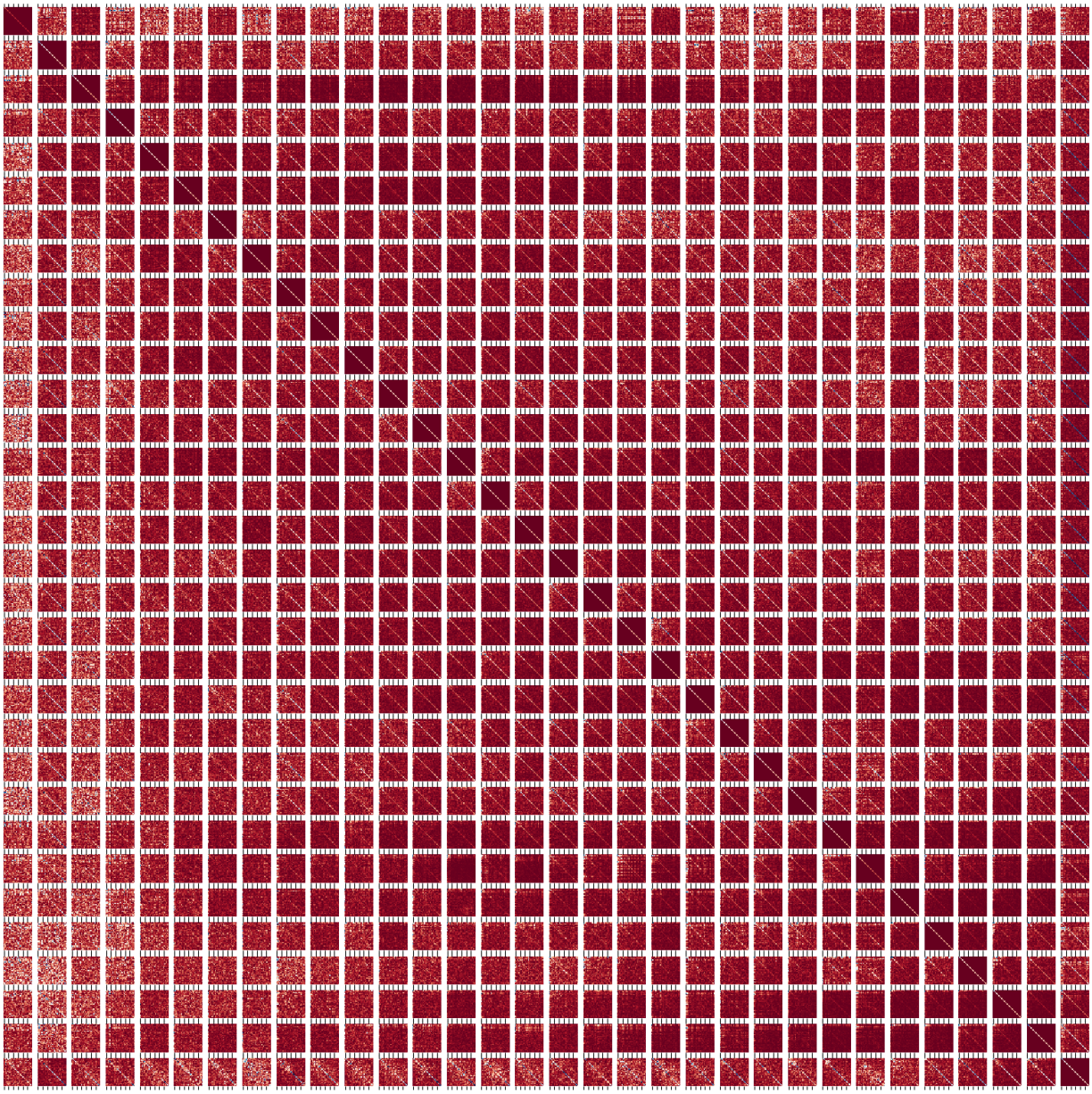


Figure 14. Residual Jacobian alignment on Falcon-7b for prompt 'What is the capital of France? The capital is' with its final token.

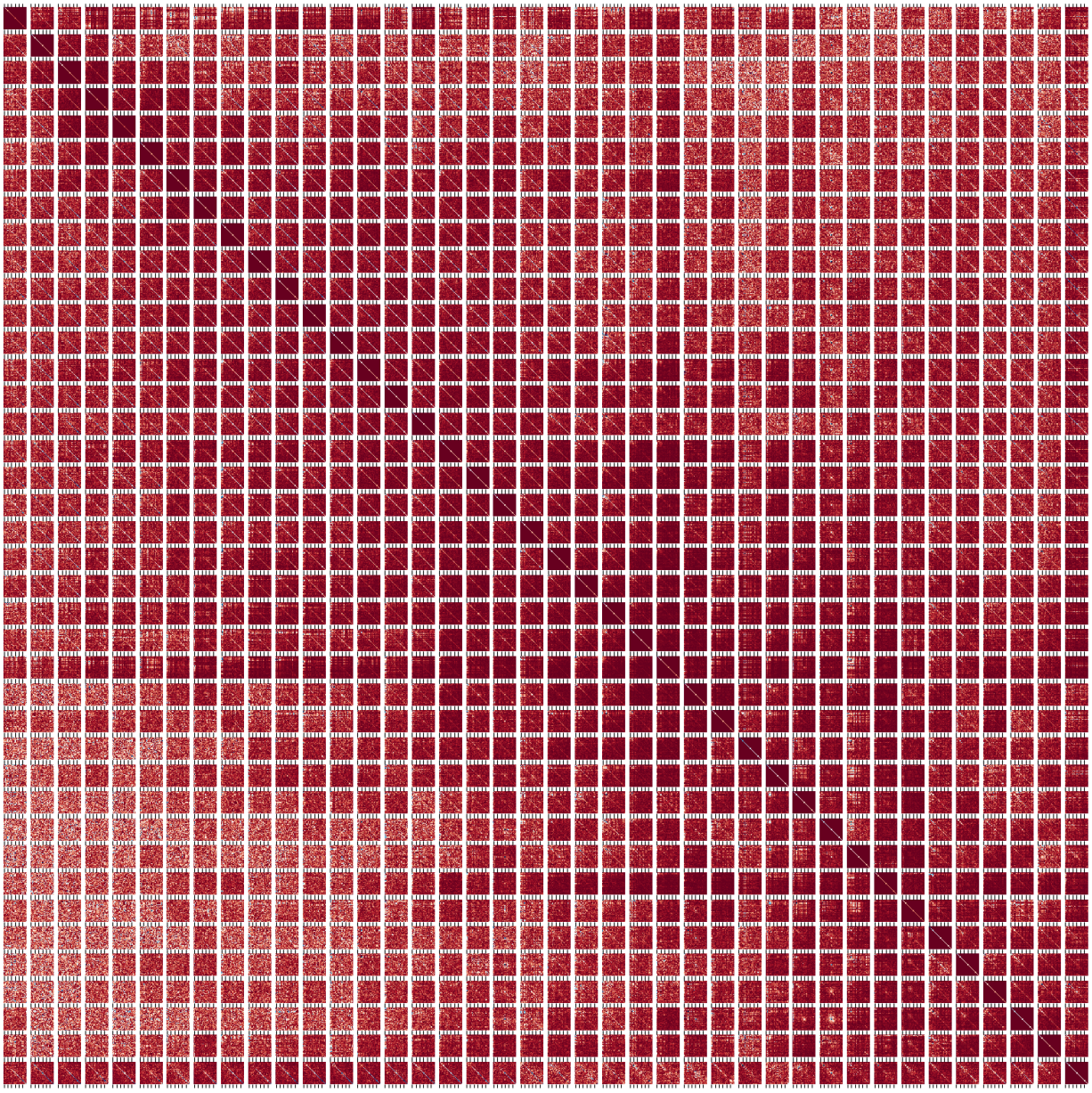


Figure 15. Residual Jacobian alignment on Llama2-13b for prompt 'What is the capital of France? The capital is' with its final token.

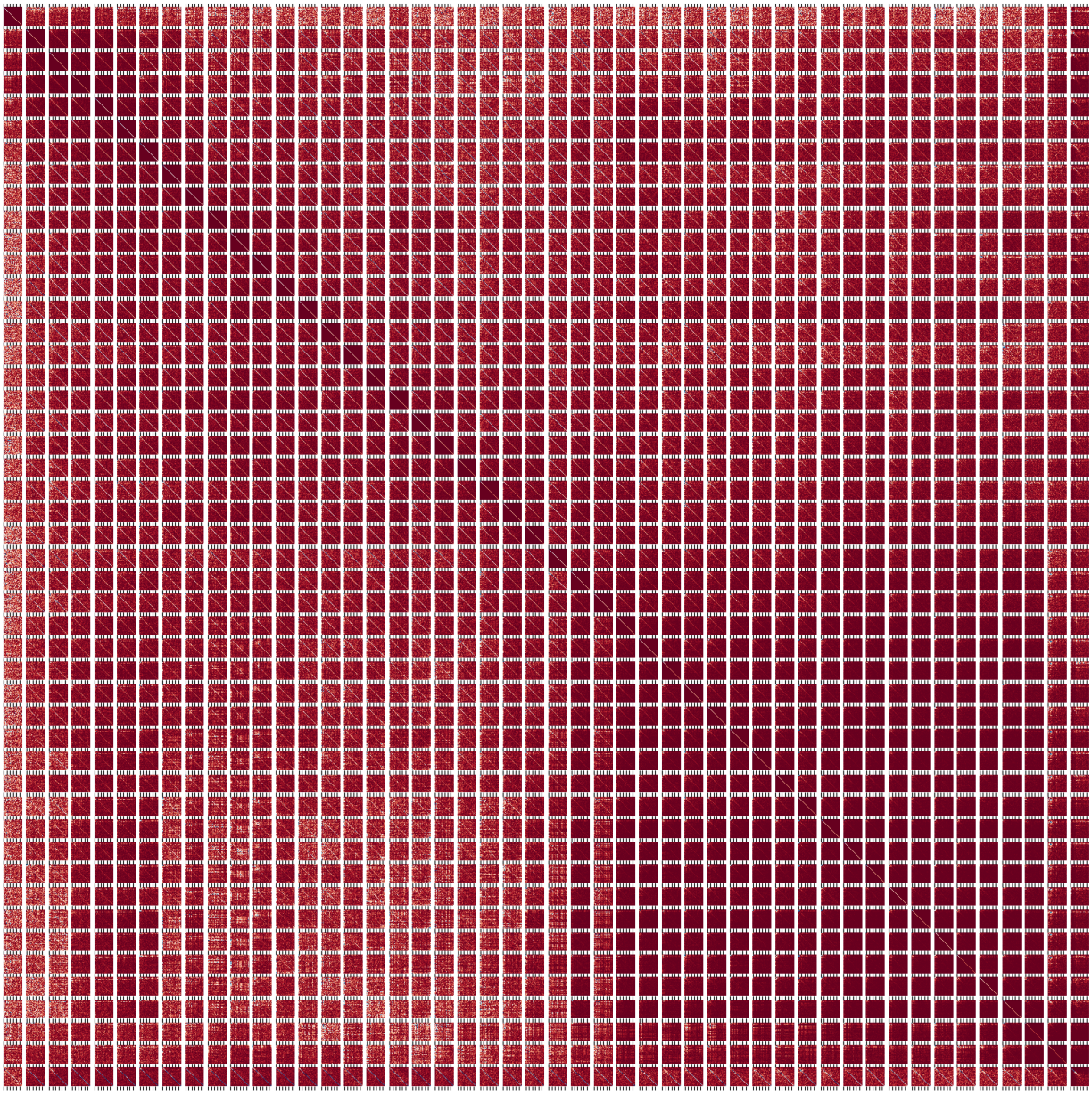


Figure 16. Residual Jacobian alignment on GPT2-XL for prompt 'What is the capital of France? The capital is' with its final token.

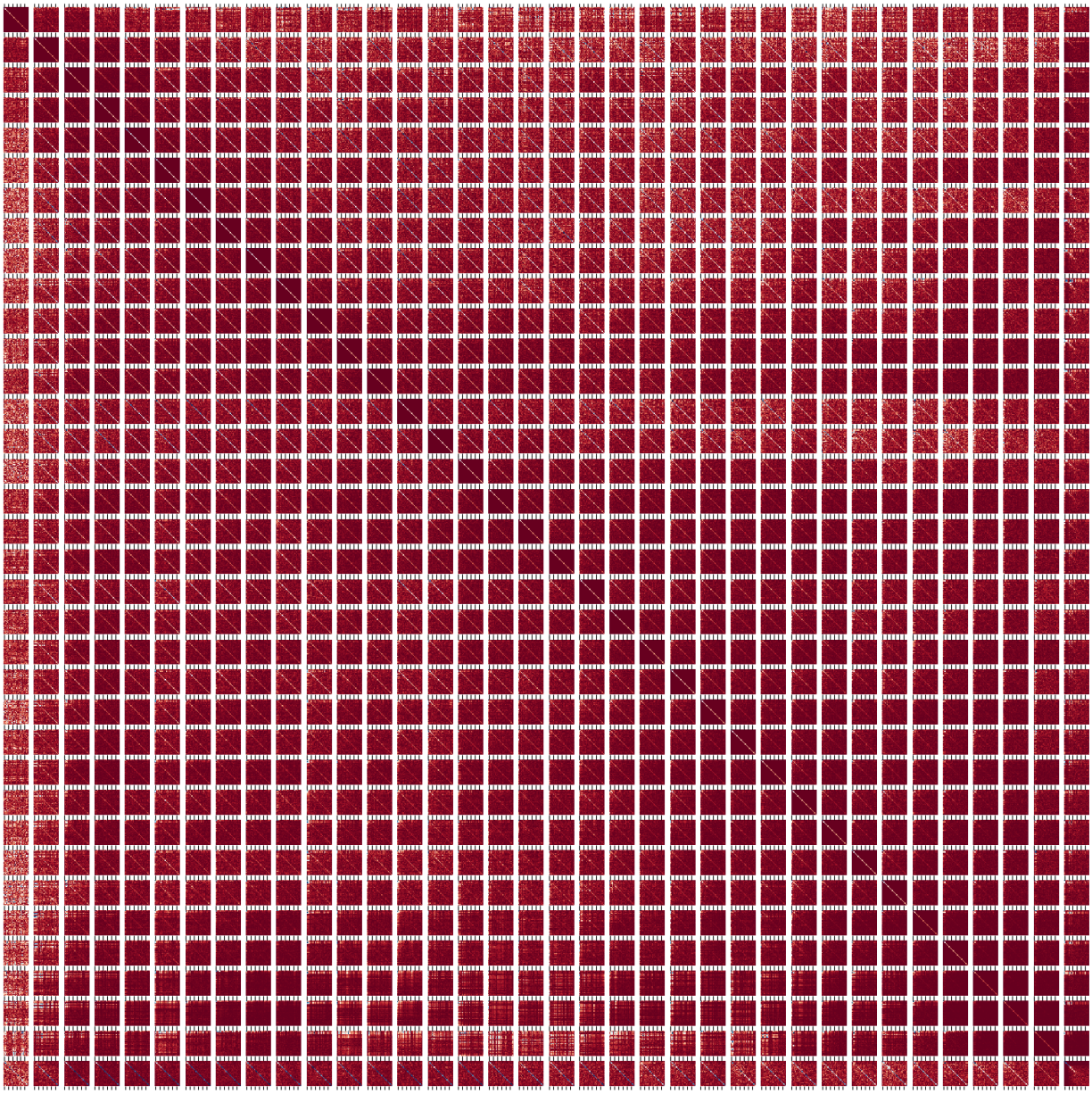


Figure 17. Residual Jacobian alignment on GPT2-L for prompt 'What is the capital of France? The capital is' with its final token.

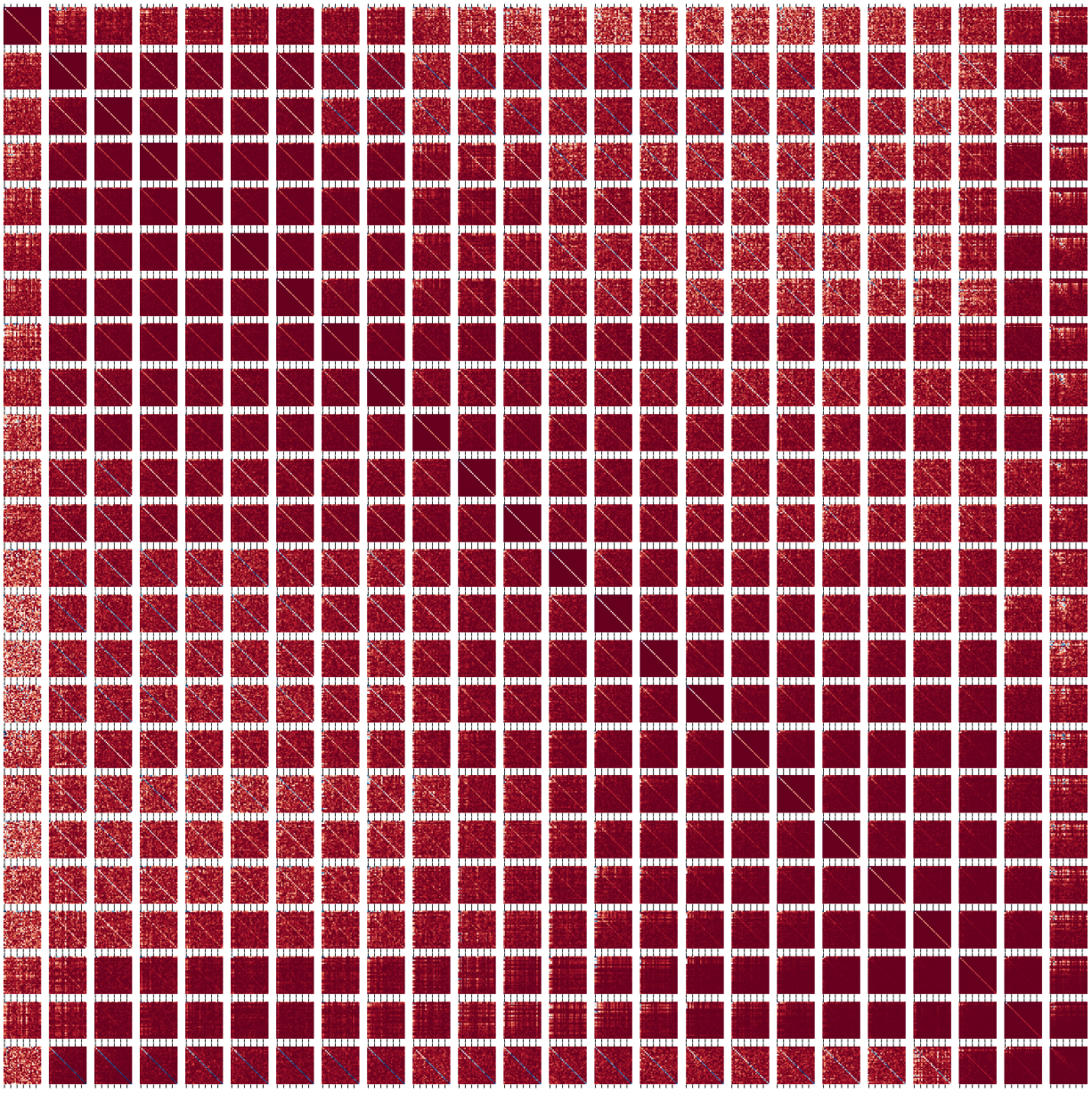


Figure 18. Residual Jacobian alignment on GPT2-M for prompt 'What is the capital of France? The capital is' with its final token.